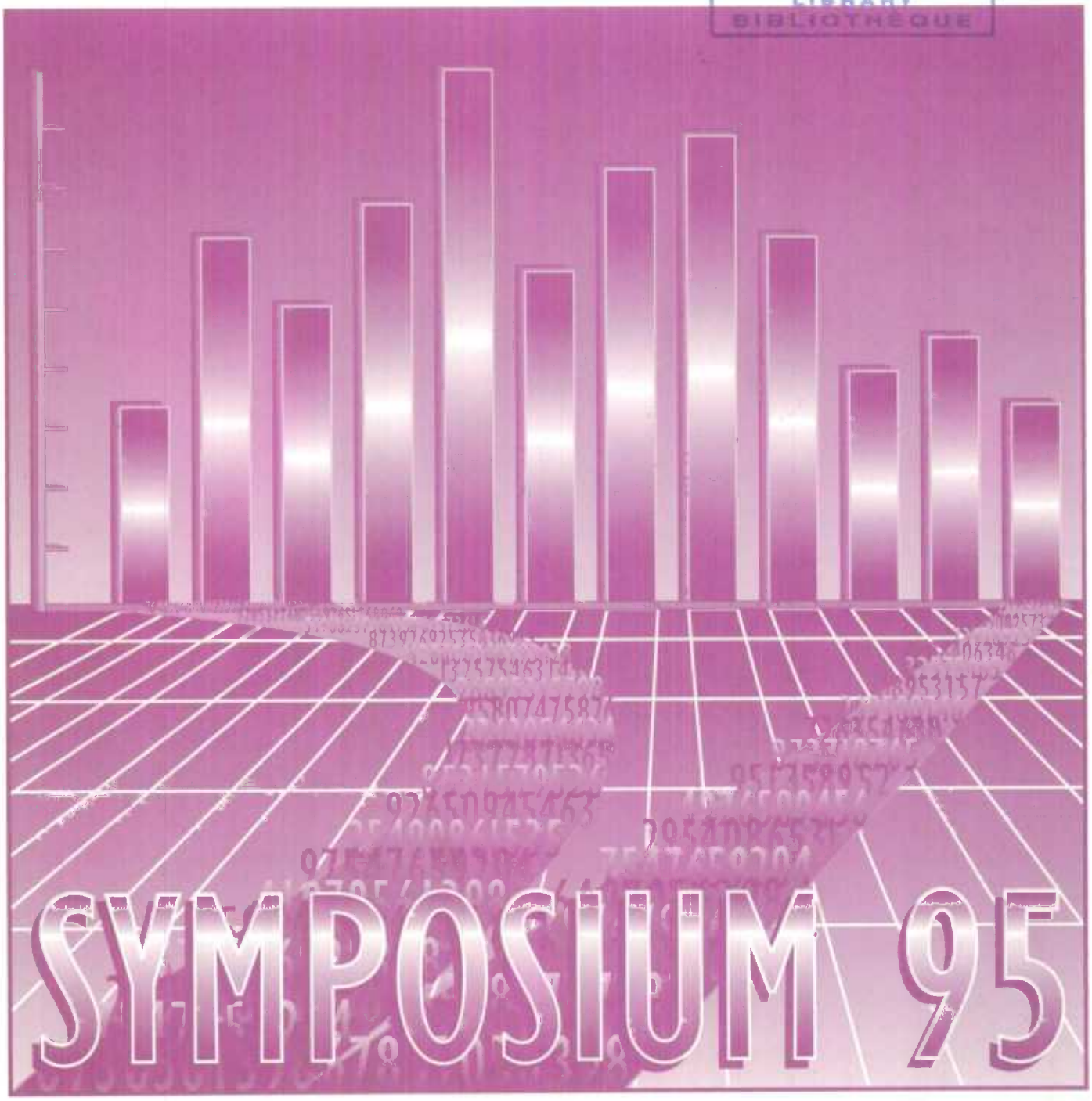
N° 11-522-XPF au catalogue

SYMPOSIUM 95

Des données à l'information
Méthodes et systèmes

RECUEIL

STATISTICS CANADA STATISTIQUE CANADA
OCT 22 1996
LIBRARY BIBLIOTHÈQUE



Statistique Canada Statistics Canada

Canada

Des données sous plusieurs formes

Statistique Canada diffuse les données sous formes diverses. Outre les publications, des totalisations habituelles et spéciales sont offertes. Les données sont disponibles sur Internet, disque compact, disquette, imprimé d'ordinateur, microfiche et microfilm, et bande magnétique. Des cartes et d'autres documents de référence géographiques sont disponibles pour certaines sortes de données. L'accès direct à des données agrégées est possible par le truchement de CANSIM, la base de données ordinolingue et le système d'extraction de Statistique Canada.

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet de la présente publication ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes-entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : (613) 951-8615) ou à l'un des centres de consultation régionaux de Statistique Canada :

Halifax	(902) 426-5331	Regina	(306) 780-5405
Montréal	(514) 283-5725	Edmonton	(403) 495-3027
Ottawa	(613) 951-8116	Calgary	(403) 292-6717
Toronto	(416) 973-6586	Vancouver	(604) 666-3691
Winnipeg	(204) 983-4020		

Vous pouvez également visiter notre site sur le Web : http//www.statcan.ca

Un service d'appel interurbain sans frais est offert **à tous les utilisateurs qui habitent à l'extérieur des zones de communication locale** des centres de consultation régionaux.

Service national de renseignements	**1 800 263-1136**
Service national d'appareils de télécommunications pour les malentendants	**1 800 363-7629**
Numéro pour commander seulement (Canada et États-Unis)	**1 800 267-6677**

Comment commander les publications

On peut se procurer les publications de Statistique Canada auprès des agents autorisés et des autres librairies locales, par l'entremise des centres de consultation régionaux de Statistique Canada, ou en écrivant à :

Statistique Canada
Division des opérations et de l'intégration
Gestion de la circulation
120, avenue Parkdale
Ottawa (Ontario)
K1A 0T6

Téléphone : (613) 951-7277
Télécopieur : (613) 951-1584
Toronto (carte de crédit seulement) : (416) 973-8018
Internet : order@statcan.ca

Normes de service au public

Afin de maintenir la qualité du service au public, Statistique Canada observe des normes établies en matière de produits et de services statistiques, de diffusion d'information statistique, de services à recouvrement des coûts et de services aux répondants. Pour obtenir une copie de ces normes de service, veuillez communiquer avec le centre de consultation régional de Statistique Canada le plus près de chez vous.

Statistique Canada
Direction de la méthodologie

SYMPOSIUM 95

Des données à l'information
Méthodes et systèmes

RECUEIL

STATISTICS CANADA STATISTIQUE CANADA
OCT 22 1996
LIBRARY
BIBLIOTHÈQUE

Publication autorisée par le ministre responsable de Statistique Canada

© Ministre de l'Industrie, 1996

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistrement sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Octobre 1996

Prix : Canada : 53 $
États-Unis : 53 $ US
Autres pays : 53 $ US

N° 11-522-XPF au catalogue

ISBN 0-660-95355-2

Ottawa

This publication is available in English upon request (Catalogue no. 11-522-XPE).

Note de reconnaissance

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

Symposium '95, Des Données à l'Information
(1995 : Ottawa, Ont.)
Symposium '95, Des Données à l'Information :
méthodes et systèmes : recueil.

Publ. aussi en anglais sous le titre: Symposium '95,
From Data to Information : methods and systems :
proceedings.
ISBN 0-660-95355-2
CS11-522-XPF

1. Services statistiques -- Congrès. 2. Statistiques
-- Congrès. I. Statistique Canada. Direction de la
méthodologie. II. Titre.

H21 S9514 1996 001.4'22
C96-988018-9

AVANT-PROPOS

Le Symposium de 1995 était le douzième de la série des symposiums internationaux sur les questions de méthodologie parrainée par Statistique Canada. Chaque année, le symposium porte sur un thème particulier. Celui de 1995 soulignait les étapes d'analyse et de diffusion des processus du développement de l'information.

Le symposium a réuni près de 300 personnes pendant trois jours au Centre de conférences Simon-Goldberg à Ottawa. On y a entendu des présentations de statisticiens du milieu universitaire et d'agences gouvernementales, des spécialistes du traitement et de la gestion de l'information, des vendeurs et des utilisateurs de données. En tout, 29 communications ont été présentées par des conférenciers et panélistes invités. Mise à part la traduction et le travail de mise en page, ce recueil contient les présentations telles que soumises par les auteurs. Les allocutions de deux panélistes ont été reproduites à partir d'enregistrements et ont subi quelques légères modifications pour en affiner la présentation en format textuel.

Les organisateurs du symposium de 1995 tiennent à exprimer leurs remerciements aux nombreuses personnes qui ont contribué à la réalisation de cet ouvrage ainsi qu'à toutes celles qui les ont aidé lors de la tenue du Symposium en novembre. En particulier nous tenons à remercier Josée Morel, Sophie Arsenault, Christine Larabie et Nick Budko pour leur assistance lors de la préparation du matériel et des préparatifs reliés au Symposium de 1995.

Il convient, naturellement, de remercier les conférenciers et panélistes pour avoir pris le temps de mettre leurs idées par écrit. La publication de ce recueil a aussi nécessité le concours de nombreuses autres personnes. Le traitement des manuscrits a été exécuté habilement par Christine Larabie aidée de Judy Clarke, Sandy Diloreto et Suzanne Fleury-Bertrand. La correction des épreuves a été effectuée par de nombreux méthodologistes, dont Jean-Luc Bernier, Alana Boltwood, René Boyer, Guylaine Dubreuil, Sylvie Gauthier, John Higginson, Tony LaBillois, Éric Langlet, Éric Lesage, Mary March, Josée Morel, Carole Morin, Sylvain Perron, Craig Seko, Michelle Simard, Jack Singleton et Larry Swain. Christine Larabie a vu à la coordination de la production de ce recueil.

Le treizième symposium annuel de Statistique Canada, tenu à Ottawa du 13 au 15 novembre 1996, sera précédé d'un atelier d'une journée. Leurs thèmes seront les erreurs non dues à l'échantillonnage.

Comité organisateur du Symposium de 1995

John Berigan — Jean Dumais
Georgia Roberts — Jean-Louis Tambay

Le lecteur peut reproduire sans autorisation des extraits de cette publication à des fins d'utilisation personnelle à condition d'indiquer la source en entier. Toutefois, la reproduction de cette publication en tout ou en partie à des fins commerciales ou de redistribution nécessite l'obtention au préalable d'une autorisation écrite de Statistique Canada.

LA SÉRIE DES SYMPOSIUMS DE STATISTIQUE CANADA

1984 - L'analyse des données d'enquête
1985 - Les statistiques sur les petites régions
1986 - Les données manquantes dans les enquêtes
1987 - Les utilisations statistiques des données administratives
1988 - Les répercussions de la technologie de pointe sur les enquêtes
1989 - L'analyse des données dans le temps
1990 - Mesure et amélioration de la qualité des données
1991 - Questions spatiales liées aux statistiques
1992 - Conception et analyse des enquêtes longitudinales
1993 - International Conference on Establishment Surveys *(en anglais seulement)*
1994 - Restructuration pour les organismes de statistique
1995 - Des données à l'information - méthodes et systèmes

LA SÉRIE DES SYMPOSIUMS INTERNATIONAUX DE STATISTIQUE CANADA
RENSEIGNEMENTS CONCERNANT LA COMMANDE DES RECUEILS

Utilisez le bon de commande sur cette page pour commander des copies additionnelles du recueil du Symposium 95: «Des données à l'information - méthodes et systèmes». Vous pouvez aussi commander les recueils des derniers symposiums. Une fois complété, retournez le formulaire à:

RECUEIL DU SYMPOSIUM 95
STATISTIQUE CANADA
DIVISION DES MÉTHODES D'ENQUÊTES - ENTREPRISES
IMMEUBLE R.H. COATS, 11e ÉTAGE
PARC TUNNEY
OTTAWA (ONTARIO)
K1A 0T6
CANADA

Veuillez inclure le paiement avec votre commande (chèque ou mandat poste, en dollars canadiens ou l'équivalent, à l'ordre du «Receveur général du Canada». Veuillez indiquer sur votre chèque ou mandat poste: Recueil du Symposium 95).

RECUEIL DU SYMPOSIUM: NUMÉROS DISPONIBLES

1987 -	Les utilisations statistiques des données administratives - FRANÇAIS	____ @ $10
1987 -	Statistical Uses of Administrative Data - ANGLAIS	____ @ $10
1987 -	ENSEMBLE DE 1 ANGLAIS ET 1 FRANÇAIS	____ @ $12 L'ENSEMBLE
1988 -	Les répercussions de la technologie de pointe sur les enquêtes - BILINGUE	____ @ $10
1989 -	L'analyse des données dans le temps - BILINGUE	____ @ $15
1990 -	Mesure et amélioration de la qualité des données - FRANÇAIS	____ @ $18
1990 -	Measurement and improvement of Data Quality - ANGLAIS	____ @ $18
1991 -	Questions spatiales liées aux statistiques - FRANÇAIS	____ @ $20
1991 -	Spatial Issues in Statistics - ANGLAIS	____ @ $20
1992 -	Conception et analyse des enquêtes longitudinales - FRANÇAIS	____ @ $22
1992 -	Design and Analysis of Longitudinal Surveys - ANGLAIS	____ @ $22
1993 -	International Conference on Establishment Surveys - ANGLAIS (disponsible en anglais seulement, recueil publié aux É.-U.)	____ @ $58
1994 -	Restructuration pour les organismes de statistique - FRANÇAIS	____ @ $53
1994 -	Re-engineering for Statistical Agencies - ANGLAIS	____ @ $53
1995 -	Des données à l'information - méthodes et systèmes - FRANÇAIS	____ @ $53
1995 -	From Data to Information - Methodes and Systems - ANGLAIS	____ @ $53
	S.V.P. AJOUTEZ LA TAXE SUR LES PRODUITS ET SERVICES (7%) (Résidents du Canada seulement)	$ ____
	MONTANT TOTAL DE LA COMMANDE	$ ____

S.V.P. INCLURE VOTRE ADRESSE COMPLÈTE AVEC VOTRE COMMANDE !

NOM _____

ADRESSE _____

_____

VILLE _____ PROV/ÉTAT _____ PAYS _____

CODE POSTAL _____ TÉLÉPHONE _____ TÉLÉCOPIEUR _____

Pour plus de renseignements, communiquer avec John Kovar, téléphone: (613) 951-8615, télécopieur (613) 951-1462, internet kovar@statcan.ca

DES DONNÉES À L'INFORMATION - MÉTHODES ET SYSTÈMES

TABLE DES MATIÈRES[1]

---

[1] Dans le cas de co-auteurs, le nom de l'orateur est imprimé en caractères **gras**.

ALLOCUTION D'OUVERTURE

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

ALLOCUTION D'OUVERTURE

G.J. Brackstone[1]

Au nom de Statistique Canada, je vous souhaite la bienvenue à ce douzième colloque d'une série dont les origines remontent à 1984. Pour ceux qui viennent d'autres villes, bienvenue à Ottawa et, pour ceux qui viennent d'autres pays, bienvenue au Canada.

Depuis son inauguration en 1984, cette série de colloques a touché une foule de sujets se rapportant de près ou de loin aux enquêtes. Nous avons par exemple abordé la question des statistiques régionales, l'utilisation des données administratives en statistique, les répercussions de la technologie sur les enquêtes, la conception et l'analyse des enquêtes longitudinales et les enquêtes sur les établissements. L'an dernier, nous nous sommes attaqués à un sujet plus vaste, qui dépassait la méthodologie et l'analyse des enquêtes, en parlant de la restructuration au sein des organismes de statistique. Nous voici aujourd'hui avec un autre thème très général, qui n'est pas sans rappeler celui du premier colloque, l'analyse des données d'enquête, mais qui s'étend en outre aux difficultés que pose la transformation des statistiques en information utiles aux intéressés.

Cette année, l'accent sera mis sur ce qu'on pourrait appeler, sans aucune intention dépréciative, la «dernière étape» du processus d'enquête - c'est-à-dire les activités qui ont lieu une fois que les opérations intensives de collecte et de traitement sont achevées et que nous avons en main les données dont nous voulons assurer l'utilisation efficace.

Ces dernières années, dans de nombreux organismes de statistique, on a accordé une importance toujours plus considérable à ces activités, qui sont axées sur les résultats. À mon avis, deux grands facteurs motivent cette attention accrue. Premièrement, on s'est rendu compte que les organismes de statistique devaient renforcer leur orientation-clientèle, afin de conserver leur pertinence et, par voie de conséquence, leurs assises et leur financement. En second lieu, il fallait réagir aux pressions financières : une fois recueillies, les données devaient être utilisées de façon optimale et les coûts entraînés par leur diffusion intégralement recouvrés.

Depuis les années 60, la conception et la finalité des enquêtes ont bien changé. Au départ, à chaque enquête ou à chaque instrument de collecte devaient correspondre une ou des publications. On a commencé par admettre que le résultat visé par une enquête n'était pas au premier chef une publication, mais plutôt une base de données pouvant donner lieu à une publication, et dont on pourrait également tirer des tableaux auxiliaires et des analyses sur requête. En d'autres termes, le produit d'une enquête était primordialement une base de données, assortie de mécanismes de récupération de ces données. Mais ce point de vue cède maintenant le pas à une autre perspective. Il faut désormais envisager l'enquête comme facteur contributif d'une base d'information collective ou intégrée, réunissant des données de nombreuses sources différentes, données pouvant être récupérées par des mécanismes communs et intégrés - c'est-à-dire une base de données générale, comme fondement d'un service d'information exploitant tous les ensembles de données disponibles, tant individuellement que globalement.

Cette évolution reflète la compréhension du fait que les résultats d'une enquête ne se résument pas à un ensemble isolé de tableaux; ils constituent un ajout à une base de données qui peut servir à toutes sortes de fins, tant prévues qu'imprévisibles.

En prenant pour thème du colloque «Des données à l'information», nous nous efforçons d'attirer l'attention sur les problèmes et les difficultés qui surgissent lorsqu'on s'efforce de transformer des données, donc des chiffres, en éléments d'information pour faire progresser le savoir. L'enjeu consiste à trouver des méthodes et des solutions pour aider les organismes de statistique à faire en sorte que leurs précieuses bases de données engendrent une information non seulement utile, mais dont on se servira.

---

[1] G.J. Brackstone, Statisticien en chef adjoint, Secteur de l'informatique et de la méthodologie, Statistique Canada, 26-J, Édifice R.H. Coats, Ottawa, (Ontario) K1A 0T6.

Pour cela, nous ferons des incursions dans de nombreux domaines qui, je crois, sont bien représentés dans le programme. Nous examinerons la gestion de l'information et le stockage des données - la façon d'organiser les données en notre possession et d'en faciliter l'accès, ainsi que de fournir des éclaircissements à leur sujet; nous nous pencherons sur l'intégration des données et la manière de faciliter l'usage collectif des renseignements recueillis dans le cadre de diverses enquêtes; il sera question de l'accès des analystes aux données et des mesures qui les aideront à entreprendre les travaux qu'ils désirent effectuer sans compromettre la nature confidentielle des renseignements personnels. La diffusion de l'information sur les données au grand public et le rôle des médias seront aussi examinés; nous étudierons la qualité des données et les méthodes d'analyse qui tiennent compte à la fois de la provenance et de la qualité des données; enfin, nous nous intéresserons à l'incidence de la technologie sur la diffusion et verrons comment le partenariat peut faciliter la tâche des organismes de statistique dans le secteur de l'information.

Mais pourquoi soulever cette question aujourd'hui? À ma connaissance, il existe au moins quatre raisons de le faire.

a) On attache maintenant à la compréhension des processus et des facteurs qui les modèlent plus de poids qu'à la seule description de la situation en résultant. En ce qui nous concerne, il nous faut comprendre les facteurs liés, par exemple, au phénomène de la paupérisation et aux moyens d'en sortir - et non simplement connaître le nombre de personnes en cause. Il nous faut comprendre pourquoi certaines entreprises réussissent alors que d'autres échouent, et non simplement en connaître les nombres respectifs.

b) J'ai déjà fait mention des pressions financières. Les compressions budgétaires et le coût élevé de la collecte de données nous dictent la meilleure utilisation possible des fonds de données existants. Cela signifie non seulement l'exploitation optimale de chaque ensemble de données, mais aussi l'intégration ou l'appariement des différentes sources de données.

c) La technologie rend l'innovation possible. Nous pouvons traiter des données plus nombreuses plus facilement et plus rapidement. Nous pouvons diffuser les données dans un rayon plus étendu, plus rapidement. Nous disposons aujourd'hui de méthodes d'analyse fortement informatisées, qui auraient été inconcevables il y a dix ans. Par ailleurs, la technologie a un prix. Nous sommes confrontés à des problèmes plus ardus de gestion et de contrôle des données, ainsi que d'accès à ces données : il faut faciliter cet accès lorsqu'il est approprié et l'empêcher dans le cas contraire. Il se pose aussi des problèmes relatifs à la protection de la vie privée lors du couplage des données et de leur utilisation secondaire.

d) Enfin, des bases de données développées, les bases de données longitudinales en particulier, permettent la réalisation d'analyses plus poussées. La possibilité d'extraire et d'exploiter une information plus riche croît au fur et à mesure de la constitution de bases de données plus vastes et plus complètes, que ce soit au moyen d'enquêtes ou de sources administratives.

Voilà donc quelques-unes des raisons qui nous ont incités à aborder ce sujet. Statistique Canada se heurte à maints problèmes complexes à cet égard. Il est rassurant de voir que tant de personnes sont venues, parfois de très loin, pour partager leur expérience et leur expertise, et nous aider à surmonter nos difficultés. Je suis fort heureux de voir autant d'éminents conférenciers annoncés au programme du colloque, dans les domaines se rapportant au thème retenu.

Il vous intéressera sûrement d'apprendre que notre colloque réunit de 250 à 300 personnes, de tous horizons: organismes de statistique, autres ministères fédéraux du Canada, gouvernements provinciaux, milieux universitaires et secteur privé.

Puisqu'il nous reste tant de choses à voir, je ne retarderai pas les travaux plus longtemps. Permettez-moi encore une fois de vous remercier d'avoir bien voulu participer au colloque. Nous espérons que les trois prochains jours s'avérerons fructueux. Je puis vous assurer d'emblée que Statistique Canada retirera beaucoup des exposés et de l'échange d'idées, mais j'ose espérer que chacun rentrera chez lui avec des idées neuves, des solutions peut-être ou, au moins, des informations qui l'aideront dans ses activités et par la même occasion, qui faciliteront la tâche de l'organisme qu'il représente.

DISCOURS PRINCIPAL

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

LE RÔLE DES STATISTIQUES DANS L'ÉLABORATION DES POLITIQUES SOCIALES

P. Hicks[1]

La présente conférence a pour thème : «Des données à l'information». En effet, l'information n'est pas une fin en soi. Lorsqu'on l'utilise, elle se transforme en connaissances. Je souhaite profiter de l'occasion qui m'est donnée, en ce début de conférence, pour me pencher sur la question plus vaste du cheminement «des données à l'information et aux connaissances». Je m'attacherai en particulier à examiner comment les connaissances statistiques sont utilisées dans l'élaboration des politiques sociales.

Cet examen nous amènera à conclure que les pressions inhérentes aux finances et à la conduite des affaires publiques nous incitent à adopter des politiques sociales qui s'appuient davantage sur l'expérience. Or, nous ne disposons pas encore de toutes les données statistiques qu'ils nous faudrait pour cela.

Toutefois, le Canada est aujourd'hui en excellente position pour devenir un chef de file mondial dans l'établissement des bases de données propices à la mise en place de nouvelles politiques sociales axées sur les résultats. Il en découlera une réduction des coûts, une augmentation remarquable de l'efficacité et une situation où l'exercice du pouvoir et l'établissement de normes nationales seront davantage fondés sur l'expérience concrète.

DÉFINITIONS ET HYPOTHÈSES

Permettez-moi d'abord d'établir certaines définitions.

Par politiques, j'entends l'ensemble des activités proposées par ceux qui conseillent les gouvernements au sujet des orientations de leurs programmes.

Données objectives et données subjectives. Les informations sur lesquelles s'appuient les politiques peuvent être objectives ou subjectives. Les données objectives comprennent celles issues des enquêtes nationales bien connues de Statistique Canada. Elles englobent également les informations du type de celles qu'on peut trouver dans les bonnes évaluations de programmes: ce qu'il advient aux groupes de participants avant, pendant et après la mise en oeuvre d'un programme, et comment la situation de ces groupes se compare à celle des groupes témoins.

À l'opposé, les données subjectives, sont celles qui portent sur les valeurs et les attitudes. Il peut s'agir par exemple d'informations provenant des enquêtes publiques et des sondages. À partir de quel moment le public et les experts considèrent-ils qu'une tendance sociale - par exemple, l'évolution de la violence conjugale - devient un «problème politique» justifiant une intervention gouvernementale prioritaire? Que nous enseigne l'opinion publique et les avis des experts sur la façon la plus appropriée de traiter les jeunes contrevenants?

À défaut de solides données objectives, nous sommes contraints de nous en remettre à l'intuition, aux anecdotes et à l'idéologie. À défaut de solides données subjectives, les politiques risquent de devenir peu à peu coupées de la réalité.

Dans le présent exposé, je limiterai mes commentaires à la situation actuelle en ce qui a trait aux données statistiques objectives. Toutefois, je trouvais important d'attirer votre attention sur les données subjectives, puisqu'elles revêtent la même importance que les données objectives dans le processus d'élaboration des politiques. Par ailleurs, l'ampleur des changements requis dans ce domaine est tout aussi grande, mais je laisserai à d'autres le privilège d'aborder cette question.

Le meilleur des mondes statistiques. Avant d'entreprendre l'évaluation de l'état actuel des statistiques sociales, permettez-moi d'abord de poser une prémisse importante : si on dispose de bonnes statistiques, on n'hésitera pas à les utiliser dans l'élaboration des politiques.

---

[1] Peter Hicks, Conseiller politique principal, 56 rue Sparks, salle 600, Ottawa, (Ontario), K1P 5A9.

Beaucoup d'entre vous ici présents pourront rester sceptiques. Il y a d'innombrables anecdotes au sujet de la surabondance des informations et des masses de données qui s'accumulent sans jamais être utilisées. On soupçonne souvent les statistiques de servir à justifier des décisions déjà prises au lieu de faire partie intégrante de l'élaboration des politiques.

Pourtant, mon expérience personnelle m'incite à croire que la situation est toute autre. Il existe un réel besoin de statistiques pertinentes. Si les statistiques ne sont pas utilisées, c'est qu'elles ne doivent pas être utiles. Les gens ont toujours le souci du travail bien fait. Ils accueillent volontiers les données qui leur permettront d'y arriver. Personne à mon avis ne préconise des politiques d'emploi inefficaces ni des politiques de soins de santé qui rendent les gens plus malades.

Il peut bien sûr souvent arriver que pour défendre leurs intérêts personnels, certains maintiennent des politiques inefficaces - par exemple, certains professionnels dont les aptitudes sont devenues superflues par suite de réformes. Toutefois, même dans de tels cas, les pressions du public et un certain sens du professionnalisme auront tôt fait de briser cette résistance, si on peut faire la preuve qu'il existe des méthodes plus efficaces.

Pour être utilisées, les statistiques doivent être connues. Entre leur collecte et leur utilisation pour l'élaboration des politiques, les données statistiques suivent un cheminement complexe. Les réunions du Cabinet ne sont pas consacrées à l'étude des plus récents tableaux de Statistique Canada. Les informations statistiques aident plutôt les conseillers en matière de politiques à se faire une opinion et à évaluer les options qui pourront être finalement présentées aux ministres.

Ce processus souffre de certaines lacunes. Les services chargés de l'élaboration des politiques, à Ottawa et dans les capitales provinciales, sont souvent portés à concentrer leur attention sur les problèmes de l'heure. Beaucoup sont mal équipés pour traiter d'une façon quantitative des problèmes qui se posent à plus long terme. Les cellules de réflexion indépendantes ont souvent, elles aussi, un horizon temporel rapproché. Les groupes d'intérêts sont souvent trop fragmentés pour adopter une perspective fondée sur l'expérience. Finalement, rares sont les universitaires qui ont le temps ou les ressources suffisantes pour se pencher sur l'interface entre les statistiques et les politiques sociales.

Néanmoins, je persiste à croire qu'il serait relativement facile et rapide d'établir de tels liens si on pouvait compter sur les statistiques voulues, c'est-à-dire, si nous étions dans une situation idéale de ce point de vue.

Dans le reste de mon exposé, je m'attacherai à répondre à trois questions. Premièrement, à quoi ressemblerait un ensemble idéal de statistiques sociales? Deuxièmement, que nous manque-t-il pour réaliser de telles conditions? Troisièmement, comment peut-t-on améliorer les choses?

LES STATISTIQUES SOCIALES IDÉALES

Dans une situation idéale, les statistiques sociales seraient complètes et intégrées, tant du point de vue de leur portée que de celui de leur polyvalence.

La dimension horizontale. Par portée, j'entends la qualité de données statistiques qui peuvent se prêter à des comparaisons entre les différents domaines des politiques sociales. Les anciennes frontières s'érodent rapidement. Il y a une génération à peine, le soutien du revenu et les services sociaux étaient considérés comme deux aspects passablement distincts des politiques sociales. Le monde scolaire, le monde du travail et celui des retraités étaient traités séparément. Les liens évidents entre la pauvreté, la santé, la nature du travail, l'intégration sociale, l'apprentissage, la criminalité et le bien-être n'étaient pas pris en compte dans les politiques sociales.

Cette situation a aujourd'hui commencé à changer. On admet maintenant volontiers qu'il existe des rapports entre tous ces domaines, à tout le moins au niveau de la rhétorique et des discours - programmes. Cette prise de conscience commence également à se faire sentir au niveau opérationnel. Par exemple, la création récente du ministère fédéral du Développement des ressources humaines a entraîné la fusion de programmes sociaux qui étaient auparavant jugés passablement distincts.

Idéalement, les statistiques devraient pouvoir favoriser cette nouvelle politique sociale plus complète. Elles devraient pouvoir nous renseigner sur ce qui se passe dans les foyers, dans les écoles et dans les cabinets des médecins. Elles devraient traiter des questions intergénérationnelles et des questions du cycle vital, de la nature changeante du travail et de l'apprentissage, et des incidences combinées des programmes sur le bien-être et sur le développement des individus et de la société.

La dimension verticale. Par polyvalence, j'entends la qualité d'une base de données intégrée à pouvoir servir à plusieurs types d'utilisateurs. La même base de données devrait fournir des indicateurs sociaux utiles pour cerner les nouveaux problèmes et promouvoir les interventions publiques. Elle devrait appuyer les évaluateurs de programmes qui se demandent ce qui a fonctionné par le passé, et les conseillers en matière de

politiques qui se demandent ce qui pourrait fonctionner à l'avenir. Elle devrait servir aux membres du personnel des points de service qui se demandent quelles sont les interventions qui risquent de donner les meilleurs résultats en diverses circonstances. Elle devrait finalement venir en aide aux particuliers qui souhaitent connaître les types de cours de formation, de soins de santé ou de régimes d'exercices physiques qui leur seront les plus salutaires.

Les données idéales devraient donc avoir les qualités requises pour mesurer l'efficacité des programmes - leurs coûts et les résultats obtenus - et pour prévoir quelles sont les interventions qui présentent les meilleures possibilités de succès.

La base de données idéale autoriserait l'intégration des données axées sur les programmes aux informations portant sur le fonctionnement des principales institutions de notre société : la famille, le marché du travail, la collectivité, les écoles, les établissements de soins de santé et de services sociaux et les organisations culturelles. Elle permettrait de mesurer des tendances sur les plans des activités individuelles et du bien-être des individus.

LA SITUATION ACTUELLE

La deuxième question consiste à savoir à quel point nous sommes éloignés de cette situation idéale.

En toute justice, il convient d'abord de mentionner les nombreuses améliorations récentes et d'insister sur le fait que de nombreuses autres améliorations sont en bonne voie d'être apportées. Il faut rappeler qu'en matière de statistiques, le Canada est en bien meilleure position que d'autres pays. Il faut finalement souligner l'excellence des données existantes utilisées aux fins de la commercialisation, et la puissance des nouvelles enquêtes longitudinales utilisées aux fins de la recherche, notamment la nouvelle Enquête sur la dynamique du travail et du revenu (EDTR) et les enquêtes sur la santé et sur les enfants.

Toutefois, **vue sous l'angle de l'élaboration des politiques sociales, la situation n'est pas aussi reluisante**. Les responsables de l'élaboration des politiques souhaitent avant tout pouvoir compter sur des statistiques cohérentes, intégrées, qui permettent de bien cerner les problèmes et de proposer, à terme, des solutions utiles. Or, les statistiques de base dans ce domaine ont peu changé au cours des dernières décennies.

Digression sur l'importance des indicateurs de tendances. Arrêtons-nous un instant pour nous pencher sur l'importance de bons indicateurs de tendances pour l'élaboration des politiques.

> Les chercheurs universitaires s'intéressent aux causes des problèmes et aux études approfondies, souvent longitudinales, qui peuvent leur fournir ces informations. Les spécialistes des marchés s'intéressent pour leur part à des données qui se prêteront à des classifications croisées instructives, par régions géographiques et par groupes socio-économiques.
>
> Les conseillers en matière de politiques trouvent également ce genre d'informations utiles. Toutefois, il leur importe avant tout de pouvoir compter sur des séries chronologiques longues et cohérentes. Les données recueillies doivent couvrir de nombreuses années, de manière que les dimensions cycliques des problèmes ainsi que les incidences des divers programmes gouvernementaux puissent être bien comprises.
>
> D'où vient l'importance de l'étude des tendances? Notre société est pourvue d'institutions sociales hautement développées qui se penchent déjà sur presque tous les problèmes sociaux imaginables. Les gouvernements doivent aujourd'hui faire des choix : faire un peu plus ou un peu moins dans certains domaines; faire les choses de façon légèrement différente ou répartir les responsabilités autrement entre les divers ordres de gouvernement et le secteur privé. Pour prendre toutes ces décisions, il convient de savoir à tout moment si les problèmes s'aggravent ou s'amenuisent avec le temps, ou si les programmes gouvernementaux deviennent plus ou moins efficaces ou plus ou moins abordables.

Dans les domaines qui présentent le plus d'intérêts en matière de politiques sociales aujourd'hui, nous sommes pratiquement dépourvus de bonnes données sur les tendances, même si les enquêtes sociales générales et certaines enquêtes connexes sur la main-d'oeuvre active peuvent nous fournir certaines indications.

Ce que nous avons déjà mérite d'être préservé. Nous pouvons déjà compter sur de bonnes séries chronologiques à l'appui des politiques qui étaient importantes au cours de la période d'après-guerre, jusque vers la fin des années 1960, lorsque nos politiques sociales se résumaient à assurer un solide filet de sécurité sociale, à lutter contre les inégalités des revenus dues au fonctionnement du marché et à assurer une

accessibilité plus large à l'école, aux soins de santé et au travail rémunéré.

Autrement dit, nous disposons d'informations raisonnablement bonnes sur les tendances des revenus des Canadiens, sur la nature des groupes engagés dans divers types d'institutions et sur les coûts qui leurs sont attribuables. Tous ces renseignements nous sont toujours utiles, même si les orientations principales de nos politiques ont aujourd'hui changé.

En d'autre mots, je ne crois pas qu'il faille aujourd'hui abandonner les outils statistiques que nous avons déjà. Ils sont toujours utiles pour traiter de plusieurs questions, et les séries chronologiques très longues présentent toujours d'énormes avantages, même si le cadre conceptuel qui les sous-tend est aujourd'hui périmé. Tout ce que je cherche à dire, c'est que l'utilité des séries chronologiques existantes n'est pas aussi évidente, lorsqu'on les envisage sous l'angle des questions politiques actuelles.

REGARD VERS LE FUTUR

Je voudrais terminer mon exposé en me penchant sur ce qui devrait être fait. J'ai plusieurs observations à formuler, et elles sont toutes marquées par l'optimisme.

En bref, nous sommes près de pouvoir nous doter d'un cadre qui autorisera l'intégration requise. Nous avons déjà la technologie qui nous permettra de traiter les masses considérables de données requises, sans enfreindre les règles de la confidentialité. Par ailleurs, il existe une foule de motifs justifiant des changements, tant du point de vue des finances que de celui de la conduite des affaires publiques.

Établissement d'un cadre. J'ai déjà décrit un ensemble idéal de statistiques sociales qui est très complet. Nous disposons dès à présent de bon nombre des éléments de cet ensemble, mais il nous manque toujours le cadre qui nous permettrait de les utiliser d'une manière intégrée. Fort heureusement, nous sommes très proches de la mise en place d'un nouveau cadre conceptuel qui nous permettra d'étayer à la fois les politiques sociales et les statistiques sociales.

En ce qui concerne l'aspect politique, nous sommes soit trop timides, soit insuffisamment courageux pour aborder explicitement la question du cadre. Toutefois, un cadre apte à guider nos politiques semble malgré tout émerger. L'OCDE a déjà établi ses nouvelles orientations en matière de politiques sociales. On s'appuie de plus en plus sur des notions semblables dans toutes les autres disciplines sociales et dans tous les autres pays développés. Ces orientations gravitent autour de notions telles que le développement humain, les investissements dans le capital humain, l'éducation permanente et d'autres approches globales; elles portent une attention beaucoup plus grande à la famille et aux questions intergénérationnelles, et abordent la question du travail en termes d'habiletés plutôt qu'en termes d'occupations.

Je crois que le jour n'est plus loin où des idées comme celles-là déboucheront sur la mise en place d'un cadre statistique qui fera pour les politiques sociales ce que le système des comptes nationaux a réalisé pour les politiques économiques. Ce cadre sera même plus puissant que le système des comptes nationaux puisqu'il sera fondé sur la façon dont les gens s'occupent, une notion beaucoup plus complète et fondamentale que la façon dont les gens dépensent.

Le nouveau cadre statistique prendra en compte la façon dont les gens répartissent leur temps entre l'école, le travail et les loisirs, le degré d'interactions sociales de ces activités, les contraintes exercées sur les gains matériels et intellectuels et la satisfaction qu'ils en tirent. Le nouveau cadre nous permettra de surveiller l'évolution du temps passé par les Canadiens à l'école, à la maison, au travail, à la retraite, à soigner les autres, à participer à divers types de programmes gouvernementaux et aux activités de diverses institutions.

Technologie. Nous possédons dès maintenant la technologie requise pour le traitement des masses considérables de données nécessaires à la mise en oeuvre de ce cadre. Nombre de ces exposés présentés au cours de cette conférence traiteront de technologie, et je ne m'étendrai donc pas sur cette question.

Le résultat est passablement simple à décrire : un ensemble de modèles de micro-simulation interdépendants, alimentés par les enquêtes existantes et nouvelles ainsi que par les registres de données administratives. D'énormes quantités d'informations seront emmagasinées au sujet d'individus fictifs et de leurs rapports avec les institutions. La confidentialité ne posera pas de problèmes puisque les « individus » en question ne seront pas de vrais Canadiens. Les données seront plutôt fondées sur des personnes fictives qui, lorsque considérées dans leur ensemble (ou en groupes), présenteront les mêmes caractéristiques que les vrais Canadiens (ou que des sous-populations de Canadiens).

À ce point-ci de mon exposé, plusieurs sentiront probablement le doute les envahir. Des bases de données énormes? L'intégration des politiques sociales sur plusieurs secteurs? Des données sur ce qui fonctionne vraiment? Des personnes fictives? Aucune menace à la confidentialité? Une augmentation radicale de

l'efficacité? Tout cela peut bien vous paraître trop beau pour être vrai.

Rappelez-vous cependant que nous appréhendons aujourd'hui le monde par le biais de statistiques qui ont été conçues il y a longtemps. Le système des comptes nationaux, le recensement et l'enquête sur la population active ont été conçus avant l'ère des ordinateurs. La puissance de ces instruments a certes été multipliée par l'informatique, mais leur structure sous-jacente est fondée sur la technologie et sur le paradigme de la machine à calculer. Il n'est donc pas exagéré de s'attendre à des améliorations énormes lorsque les plans d'échantillonnage seront élaborés sur des bases entièrement nouvelles, fondées sur des technologies informatiques extrêmement puissantes.

Il existe un plan. Je voudrais vous assurer que tout ce que je vous ai décrit ne relève pas de l'utopie. Il existe bel et bien un plan pour réaliser tout cela.

Un plan qui permettra aux programmes de s'appuyer sur les données de l'expérience. En ce qui concerne les données orientées sur les programmes, le ministère fédéral du Développement des ressources humaines procède actuellement à l'essai d'un modèle qui permettra aux responsables des politiques d'évaluer simultanément la rentabilité de divers types d'interventions et qui permettra à des particuliers d'évaluer leurs chances de succès dans différents genres de formation ou d'autres interventions en matière d'emploi. Ce système s'articule autour d'une banque de données qui permettra d'évaluer les décisions actuelles à l'aune des succès de toutes les interventions passées auprès de gens présentant des caractéristiques semblables, dans des circonstances comparables.

Ce type de technologie révolutionnera l'élaboration des programmes en matière de services sociaux et leur permettra de s'appuyer sur des données concrètes. Il permettra un transfert de la prise des décisions aux membres du personnel des points de services, qui seront de ce fait en mesure d'apprendre et d'ajuster continuellement leurs interventions en fonction des résultats obtenus. Tôt ou tard, la nouvelle technologie mettra l'information et le pouvoir qu'elle représente entre les mains des citoyens eux-mêmes. Les économies réalisées par les gouvernements, à mesure que cette technologie se répandra dans les divers secteurs des sciences sociales et des sciences de la santé, finiront par se compter en milliards de dollars. Le contrat social passé entre les citoyens et l'État sera révisé. Les programmes sociaux fonctionneront vraiment.

Révolution de notre connaissance de la société. Par ailleurs, du côté des statistiques portant sur l'ensemble de la population, une ébauche de projet a déjà été élaborée par des responsables de Statistique Canada et du ministère du Développement des ressources humaines. Ce projet décrit le nouveau cadre, les nouveaux modèles de micro-simulation et les nouvelles méthodes de collecte des données qui seront requises. Sa mise en oeuvre pourrait également révolutionner les statistiques sociales et les politiques sociales, c'est-à-dire, notre analyse de la société et des moyens d'en améliorer le fonctionnement.

Ce projet de statistiques n'a pas encore été financé ni validé par un plus large ensemble de groupes d'intérêts et de chercheurs. Il est toujours difficile de trouver des financements, en particulier en cette période de restrictions. Toutefois, j'envisage avec un optimisme prudent les futurs développements dans ce domaine. D'énormes pressions s'exercent déjà qui favoriseront l'évolution des politiques sociales vers le recours à des données plus complètes, même s'il faut pour cela songer à un financement initial. Ces pressions proviennent autant des impératifs financiers que de ceux liés à la conduite des affaires publiques.

Pressions financières. L'évolution de la situation financière poussera à accorder une plus grande attention à l'efficacité, à favoriser les interventions qui ont une réelle utilité pour les gens, les sociétés et le trésor des gouvernements. Ce changement ne s'est pas encore réalisé, mais lorsque les diverses vagues initiales de coupures aveugles des programmes sociaux seront passées, l'attention se tournera presque certainement sur ceux des éléments des programmes restants qui donnent des résultats et ceux qui sont inutiles.

Pressions liées à la conduite des affaires publiques. La conduite des affaires publiques serait largement simplifiée si on adoptait une approche plus scientifique des politiques sociales.

Aujourd'hui, par exemple, lorsque nous songeons aux aspects importants des politiques sociales, nous pensons typiquement aux normes nationales ou aux contraintes qui s'exercent dans des domaines tels que l'accès, le financement ou les processus. Nous nous préoccupons de la transférabilité des programmes, de l'imposition de frais aux utilisateurs ou des règles qui limitent à certains organismes du secteur public ou à des organisations certifiées le privilège de fournir certains services.

Nous vivons ici dans un monde d'énormes transferts fiscaux, de principes abstraits et d'ententes constitutionnelles complexes.

Demain, nous nous inquiéterons davantage des effets des normes sur les résultats des programmes. Nous porterons une attention particulière à la comparabilité des habiletés acquises d'une région à

l'autre du pays, et non simplement à l'application de normes pédagogiques par les divers programmes de formation. L'accent portera davantage sur la comparabilité des résultats des programmes de santé et moins sur des questions d'accessibilité ou sur l'utilité médicale de tel ou tel service.

Nous vivrons dans un monde où les coûts principaux des normes et des principes nationaux seront limités aux coûts relativement mineurs des enquêtes et des recherches, et non à ceux découlant des ententes fiscales.

Nous vivrons dans un monde où les normes pourront être considérées non pas comme des valeurs sociétales abstraites, mais plutôt comme des balises plus terre-à-terre et plus concrètes faisant partie intégrante des indicateurs statistiques.

Nous vivrons dans un monde de prises de décisions décentralisées, où les nombreux partenaires seront en mesure de se prononcer sur des questions concrètes, qu'ils fassent partie des usagers, des praticiens, des universitaires, des fonctionnaires de tous les niveaux ou même des représentants des organisations internationales. À mesure que s'effondrent les barrières des politiques sociales, et à mesure que ces dernières s'imbriquent dans les politiques fiscales et les politiques économiques, beaucoup d'intervenants devront par nécessité intervenir et dépendre de plus en plus les uns des autres.

Soyons clairs, je ne prétends pas que les normes et les principes relatifs aux intrants ou aux processus ne seront plus nécessaires. Celles et ceux qui traitent de la mobilité et de la transférabilité des avantages continueront à avoir une importance toute particulière.

Je ne prétends pas non plus que les transferts fiscaux sont sans importance. Ils sont nécessaires pour assurer la comparabilité dans l'aptitude des provinces à financer de bons programmes sociaux.

Je ne prétends certainement pas non plus nier l'importance des solutions constitutionnelles - le partage des rôles et des responsabilités et l'élimination des recoupements et des lacunes - en particulier en cette période post-référendaire.

Ce que je prétends, c'est que notre défi fondamental, en matière de politiques sociales, devrait être de trouver des moyens qui permettront aux nombreux intervenants de travailler ensemble d'une manière productive. De bonnes statistiques fourniront un langage commun qui rendra cette collaboration possible.

Un langage statistique commun facilitera de beaucoup le traitement des questions de politiques sociales, y compris l'élaboration de normes et de buts communs. Il permettra la réalisation de progrès sensibles répondant aux impératifs de la décentralisation, axés sur l'atteinte de résultats concrets et fondés sur des données objectives.

Les discussions menées aux niveaux plus élevés, qu'elles surviennent dans le contexte d'accords constitutionnels, de chartes sociales ou des principes qui sous-tendent le nouveau programme fédéral de santé et de transfert sociaux, pourraient s'avérer particulièrement productives si elles donnent le coup d'envoi à la mise en pratique de cette approche plus scientifique.

LE CANADA PEUT ÊTRE UN CHEF DE FILE

Le Canada se trouve dans une position idéale pour jouer un rôle d'avant-garde en ces matières. Statistique Canada, en collaboration avec ses nombreux partenaires, se trouve en excellente position pour prendre la direction des travaux d'élaboration de ce nouveau langage commun. Il est déjà un chef de file mondial dans ce domaine.

Avec un solide esprit de leadership et beaucoup de persistance, nous pourrons jeter les bases qui permettront à nos successeurs de bâtir une société meilleure au cours du prochain millénaire.

Nous pouvons faire une énorme différence.

SESSION 1

Intégration des données

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

INTÉGRATION DES DONNÉES : LE POINT DE VUE DE CEUX QU'ON RELÈGUE À L'ARRIÈRE DE L'AUTOBUS

G. Priest[1]

RÉSUMÉ

Les bureaux de la statistique mettent traditionnellement l'accent sur la méthodologie. Autrement dit, la mise au point des véhicules de collecte de données s'articule autour de méthodes précises. Chaque véhicule est, en général, adapté aux exigences d'une clientèle spécialisée, sans égard aux besoins d'autres organismes. Par conséquent, au lieu de constituer une entreprise, le bureau de la statistique finit souvent par devenir un groupement, voire un groupement fragmenté, de producteurs de données relativement autonomes. Or, l'élaboration indépendante des méthodes, des systèmes, des concepts, des définitions, des classifications, des produits et des services engendre des pratiques inefficaces, des redondances et des discordances, et contrarie une partie de la clientèle.

Jusqu'ici, le client content d'une source unique d'information a été relativement bien servi. Par contre, celui qui recherchait des renseignements complets sur une question précise, un segment de population ou un lieu géographique particulier ne l'était pas autant. Aujourd'hui, toutefois, les progrès de la technologie de l'information précipitent un changement de paradigme.

En effet, une nouvelle génération de clients est en train de s'initier à l'Internet et de former de nouvelles attentes, surtout en ce qui concerne la recherche d'information. Ces clients, qui expriment des besoins uniques et distincts, veulent pouvoir survoler des métadonnées de façon thématique, déterminer les sources, faire des choix et même télécharger des données en direct, en temps réel, de façon transparente et à peu de frais, voire même gratuitement.

Le défi et l'enjeu pour les bureaux de la statistique consistent à s'adapter au changement de paradigme en répondant aux besoins de cette nouvelle clientèle. Or, la solution au problème réside dans l'intégration. Il s'agit, d'une part, de créer des liens entre les sources et, d'autre part, d'éliminer ou de réduire les discordances. L'intégration est, en outre, essentielle au passage des données brutes à l'information, car elle facilite le regroupement de toutes les données en entrée pertinentes et disponibles. La prise de décisions éclairées en dépend.

MOTS CLÉS: Silos; méta-information; discordances; données de sortie de source unique; thématique; intégration.

1. INTRODUCTION

1.1 Intégration

Au début des années 70, j'ai participé, dans le cadre de la conférence des statisticiens européens, à une séance en vue de produire un ensemble harmonisé de données internationales sur le logement, les ménages et les familles. Au dîner organisé à la fin de la première journée de travaux, un délégué canadien représentant la Société canadienne d'hypothèques et de logement s'est étonné que je sois un défenseur aussi acharné du projet d'harmonisation. Surpris, je lui ai demandé ce qu'il voulait dire, les mesures à prendre me paraissant relativement évidentes. «Je vous pose cette question,» m'a-t-il répondu, «parce qu'il règne actuellement une telle confusion dans votre service.» Il a ajouté qu'il devait s'adresser à au moins huit divisions de Statistique Canada pour obtenir les données dont il avait besoin. En outre, il s'est plaint d'un manque fréquent de comparabilité entre sources. Je suis devenu immédiatement un partisan de l'intégration, quoique sans grand effet, mes collègues du Bureau manifestant fort peu d'intérêt pour la collaboration.

Au milieu des années 80, ma division a lancé la revue trimestrielle **Tendances sociales canadiennes** et la série de publications sur les **groupes cibles**, deux

---

[1] Gordon Priest, Directeur, Division de l'intégration et du développement des statistiques sociales, Statistique Canada, Ottawa, (Ontario), Canada, K1A 0T6.

produits qui concrétisaient un effort en vue d'intégrer des données tirées de diverses sources. Devenus nous-mêmes des utilisateurs en raison de ces entreprises, nous avons pu juger des frustrations éprouvées par les clients externes. En effet, rechercher les données qui nous étaient nécessaires et y accéder s'est révélé une tâche gigantesque. Nous avons souvent comparé notre expérience à la lutte pour l'intégration raciale, car nous avions le sentiment très net d'avoir été relégués à l'arrière de l'autobus, et nous avons cherché à comprendre comment cela avait pu se produire.

La situation tenait au fait que la structure organisationnelle de la plupart des bureaux de la statistique se fonde sur la diversité des méthodologies. Nous effectuons des recensements, des enquêtes postcensitaires, des enquêtes sur les ménages, des enquêtes sur les entreprises et nous tirons aussi des données de divers fichiers administratifs. Donc, la collecte des données est, en général, axée sur le véhicule utilisé. Qui plus est, les responsables de chaque véhicule tendent à acquérir des compétences propres en ce qui concerne les systèmes, la méthodologie et les sujets abordés.

1.2 Îlots, silos et tuyaux de poêle

Une telle situation exemplifie ce que Tapscott et Caston (**Paradigm Shift: The New Promise of Information Technology**, New York, McGraw-Hill, 1994) ont appelé «le problème de l'entreprise non contrôlée». Ces auteurs décrivent des îlots de technologie ou d'expertise qui satisfont des besoins particuliers, mais résultent en une fragmentation de l'organisation. Ils précisent que les fonctions de tels îlots sont limitées et spécialisées, et parfois sans aucun rapport avec les objectifs ou les stratégies d'ensemble de l'entreprise. Dans certains cas, on assiste même à une réelle balkanisation, assortie d'énormes obstacles physiques et organisationnels qui donnent lieu à des redondances et à des inefficacités. Ces auteurs attribuent au manque d'intégration la perte d'un nombre important de débouchés.

Dans le document intitulé «Technical Evolution White Paper», Keith Vozel (AT&T) qualifie ces organisations de verticales ou «en tuyau de poêle», leurs éléments ayant tendance à ne s'intéresser qu'à un seul dossier ou à un seul client, sans égard aux besoins ou aux exigences de tiers. De telles organisations sont peu rentables, en ce sens qu'elles collectent des données redondantes ou en double, l'entreprise ne témoignant d'aucune vision collective quant à l'archivage de ces données. D'autres auteurs associent ce type d'organisations à un ensemble de silos auxquels il est difficile d'accéder et entre lesquels les communications sont limitées, voire inexistantes. De telles organisations sont en fait des réservoirs de possibilités inexploitées et d'occasions perdues.

1.3 Le groupement de producteurs et ses conséquences

Selon Bill Bradley, de Santé Canada, Statistique Canada est un bon exemple de ce type d'organisation regroupant des programmes autonomes de production des données. Je pense personnellement que, plutôt que des entreprises, les bureaux de la statistique sont des groupements de producteurs indépendants. Nombre de ces producteurs ont certes bien servi leurs clients attitrés, mais cela n'a pas été sans prix.

2. DISCUSSION

2.1 Manque de méta-information

Les bureaux de la statistique ont, en général, une connaissance *globale* très limitée, voire inexistante, de la nature et de l'ampleur de leurs banques de données et, jusqu'ici, n'ont pas partagé systématiquement avec leurs clients existants ou éventuels le peu de renseignements qu'ils possèdent à ce sujet. Combien de fois n'avons-nous pas entendu un stratège, un décisionnaire ou un chercheur se plaindre d'un manque d'information, alors que les données dont il avait besoin existaient effectivement, mais étaient enfouies, hors de vue, dans une bibliothèque de bandes magnétiques aseptisée et climatisée? Malheureusement, la production de méta-information (c'est-à-dire l'information au sujet des banques de données) dépend en grande partie des divers secteurs de production. Or, la quantité de méta-information archivée varie considérablement d'un secteur à l'autre, et la documentation connexe ne répond ordinairement à aucune norme collective. En outre, les efforts en vue de produire des métadonnées normalisées sont plus souvent déployés pour répondre à des besoins bureaucratiques que pour servir des clients éventuels. Ce manque de méta-information a pour conséquence la sous-utilisation des collections de données. Les clients, aussi bien les membres du personnel des bureaux, qui entreprennent des recherches sur une question ou une population particulière sont essentiellement livrés à eux-mêmes et obligés de communiquer avec *chaque* source pour savoir si elle dispose de données pertinentes. Étant passé par là, je peux affirmer que la tâche est gigantesque, frustrante et, souvent, improductive.

2.2 Discordances

Fait peu surprenant, étant donné la nature de la production indépendante, des discordances entre les véhicules de collecte ou les sources, en ce qui a trait aux concepts, aux définitions, aux systèmes de classification et à la documentation, créent des complications supplémentaires. En effet, chaque secteur de production met au point non seulement sa méthodologie et ses méthodes de traitement et de diffusion, mais aussi un contenu spécialisé qui lui est propre. L'indifférence ou le manque de communication des divers producteurs finit par entraîner des discordances quant aux concepts, aux définitions, aux systèmes de classification et au codage des bases de données. En plus d'être frustrante pour l'utilisateur final, cette situation cause un gaspillage de ressources. Étant donné le manque de normes collectives, à maintes reprises, les chefs de programme mettent au point de la documentation entièrement nouvelle, sans se soucier de ce qui, éventuellement, a déjà été publié par d'autres secteurs d'activité du Bureau.

Je suis certain que nous sommes tous au courant de situations où un ensemble de données en provenance d'une source particulière ne peut être comparé à celui produit par une autre source, même s'ils portent tous deux le même nom. À l'opposé, nous avons les exemples de variables effectivement comparables qui portent des noms différents. À Statistique Canada, nous avons même découvert des cas où des variables portent le même nom dans une langue officielle, mais pas dans l'autre. Et nous avons probablement tous vécu ces situations où, même si une variable retient son intégrité conceptuelle quant on passe d'une source à l'autre, les comparaisons sont impossibles parce que chaque source s'est servie d'un système de classification différent ou a effectué des agrégations non normalisées. Pour terminer, n'oublions pas la pratique insidieuse consistant à utiliser différentes mnémoniques pour la codification des variables qui figurent dans les clichés d'enregistrement destinés à l'extraction de microdonnées. Cette pratique peut causer de graves erreurs de codification quand une personne travaille avec des fichiers multisources.

2.3 Données de sortie contradictoires ou incomplètes

La production de données de sortie indépendantes, axées sur le véhicule de collecte, est un autre legs de l'organisation de type «tuyau de poêle». Elle cause des problèmes manifestes, comme quand les résultats de l'Enquête B contredisent ceux de l'Enquête A publiés antérieurement, en précisant qu'à Peggy's Cove, le nombre annuel de touristes qui sont touchés par des excréments de mouettes est 5 349 au lieu de 316. Après de tels incidents, on assiste généralement à la diffusion d'une nuée de communiqués contenant des notes en bas de page et des restrictions visant à préciser que les résultats d'une des sources sont désaisonnalisés, qu'ils regroupent les grands hérons et les mouettes fautives ou qu'ils ont été arrondis pour prévenir les déductions par recoupement. Dans certains cas, nous nous contentons d'émettre, le rouge au front, le feuillet rose d'errata et d'invoquer, avec embarras, un «problème d'ordinateur». Quoique gênantes, ces situations causent peu souvent des dommages à long terme, car elles sont relativement rares et généralement repérées et corrigées rapidement.

2.4 Données de sortie à source unique biaisées

En revanche, la diffusion d'un ensemble de données résultant de l'analyse de données provenant d'une seule source, sans le bénéfice d'une comparaison avec des données connexes et pertinentes provenant d'autres sources, est un phénomène beaucoup plus préoccupant. Le risque tient au fait que de telles données fournissent seulement des renseignements partiels, donc biaisés et trompeurs. Autrement dit, l'information n'est pas inscrite dans le contexte de notre connaissance générale de la situation. Par exemple, supposons qu'une enquête sur la consommation mondaine d'alcool par les jeunes femmes indique qu'une jeune femme sur 20 a été agressée par un jeune homme après avoir quitté un bar tard dans la nuit. La diffusion de cette information suscite un grand débat public et des pressions afin qu'on adopte une loi obligeant les tenanciers de bar à fermer leur établissement plus tôt. Supposons maintenant qu'on ait réalisé auparavant une enquête générale sur la violence dans la société, y compris la violence au foyer, au lieu de travail et dans la rue. Supposons également que cette enquête confirme les résultats de l'enquête sur la consommation mondaine d'alcool par les jeunes femmes, mais révèle que le taux de jeunes hommes agressés après avoir quitté un bar la nuit est encore plus élevé et, en outre, que les jeunes hommes sont les agresseurs les plus fréquents, non seulement dans les rues, mais aussi au foyer et au lieu de travail. Dans de telles circonstances, le débat public et la recherche d'une solution consécutifs à la publication des résultats de l'enquête sur la consommation mondaine d'alcool prendrait sans doute une toute autre tournure si ces résultats étaient présentés dans leur contexte plus large.

2.5 Conséquences de la production de type «tuyau de poêle»

Pour résumer la situation, rappelons que, n'ayant en général qu'une vue globale imparfaite de l'étendue et de la nature des banques de données à leur disposition, les

membres du personnel des bureaux de la statistique ne parviennent pas à exploiter pleinement ces ressources aux bénéfices des clients. En outre, même quand un client découvre plusieurs sources pertinentes, les données ne sont pas nécessairement comparables en raison de discordances entre ces sources. Enfin, les bureaux de la statistique induisent parfois les clients en erreur en diffusant des données axées sur le véhicule de collecte plutôt que sur l'intégration des données de sortie. Si nous admettons que la production fragmentée de données pose un problème pour les clients, nous sommes obligés d'envisager l'intégration comme l'une des solutions. Autrement dit, nous devons établir un répertoire général des fonds de renseignements, éliminer les discordances et nous assurer que la diffusion des données se fasse dans le contexte de notre pleine connaissance de la situation.

2.6 Des raisons convaincantes d'agir

Voici plusieurs raisons convaincantes d'agir immédiatement. En premier lieu, beaucoup de bureaux font face à des compressions budgétaires à une époque où la demande de renseignements augmente. Quand la conjoncture économique est difficile, il est compréhensible que les stratèges et les décisionnaires des secteurs public et privé veuillent obtenir les données les plus fiables et les plus récentes qui soient, car, dans de telles conditions, les conséquences d'une mauvaise décision ou d'une décision mal éclairée sont beaucoup plus graves. Il incombe donc aux bureaux de la statistique non seulement d'en faire plus avec moins, mais aussi de travailler plus intelligemment, proposition qui inclut l'exploitation aussi complète que possible des données existantes. Or, on ne peut exploiter ce dont on ignore l'existence. Par conséquent, tenir à jour des banques de métadonnées et des renseignements généraux est tout simplement une pratique pleine de bon sens.

En second lieu, la technologie actuelle rend la tâche de gestion des données et des métadonnées infiniment plus aisée qu'elle ne l'était il y a dix ou même cinq ans. En 1980, la Société canadienne d'hypothèques et de logement m'a demandé d'estimer le coût de la création et de la tenue à jour d'une base de métadonnées sur le logement. Selon nos estimations, le coût de la création de la base de données se chiffrait à trois années-personnes et celui de la mise à jour, à environ une année-personne et demie. Inutile de préciser que la base de métadonnées n'a pas été établie. Cette année, nous avons créé une banque de métadonnées sur les statistiques sociales où sont énumérés environ 20 sujets, plus de 1 000 variables et pratiquement 100 sources. Cette banque de métadonnées, dont la création a coûté moins d'une demi-année personne et dont les frais de mise à jour seront négligeables, permet aux clients d'effectuer des recherches selon un mode thématique ou à l'aide de mots-clés. Dans le cas de la méthode thématique, le client parcourt la liste de sujets ou de thèmes (p. ex., démographie, éducation, données ethnoculturelles, santé, travail, etc.). Le choix d'un sujet particulier fait apparaître à l'écran la liste alphabétique de toutes les variables connexes, ainsi que de toutes les sources pour chaque variable. Le choix d'une source particulière fait apparaître des renseignements de référence au sujet de la source, une nouvelle liste thématique de variables pour la source en question, les clichés d'enregistrement des microdonnées, les questionnaires et d'autres renseignements. Les avantages de cette méthode tiennent au fait que le client prend conscience de l'existence de groupes de variables connexes qui peuvent s'avérer utiles et à celui de pouvoir relier les variables par recoupement à plusieurs autres sujets.

En troisième lieu, les clients, particulièrement ceux familiarisés avec l'Internet, connaissent de mieux en mieux les méthodes de recherche de renseignements. Donc, de plus en plus fréquemment, ils veulent avoir la possibilité de s'adresser à un bureau de la statistique, de survoler les banques de données de ce dernier, de préciser les données de sortie qu'ils souhaitent et de télécharger ces dernières, en direct, en temps réel et à peu de frais, voire même sans frais. La mise en place d'une telle capacité de service entraînera incontestablement des dépenses, mais elle pourrait aussi aboutir à la réduction des frais de base (évitement des coûts) et à l'amélioration de la productivité. Par exemple, les bureaux devraient limiter le nombre de produits génériques coûteux et non seulement permettre aux clients de créer leurs propres produits spécialisés, mais les y encourager et les y aider.

3. FUTURES MESURES

3.1 La vision

Il existe un besoin, donc une occasion à saisir. Notre succès tiendra à notre vision commune de l'avenir et à la volonté collective de relever le défi. Notre vision de l'avenir devrait s'articuler autour de trois éléments fondamentaux, à savoir créer la méta-information, éliminer les discordances et passer de la production des données axées sur le véhicule de collecte à celle de données intégrées, mettant l'accent sur le problème (ou la population) étudié.

3.2 Création de la méta-information

La méta-information doit être complète. Elle doit satisfaire le client qui cherche simplement la réponse à une question précise, telle que le nombre de gadgets produits l'année précédente, aussi bien que celui qui veut savoir ce que contiennent les bases de microdonnées, afin de pouvoir effectuer ses propres recherches. Par conséquent, la méta-information doit préciser le contenu des fichiers de microdonnées, le contenu des tableaux de données agrégées, le contenu des rapports analytiques ou descriptifs, et la nature des services spécialisés offerts par le bureau. L'information doit être accessible au moyen d'un instrument facilitant les recherches par mots-clés aussi bien que les recherches thématiques. Idéalement, cet instrument devrait être précédé d'un dictionnaire synonymique permettant la conversion du lexique du client au lexique du bureau. Comme en témoignent nombre des sites les plus utiles de l'Internet, on ne peut sous-estimer l'importance d'un instrument de recherche thématique. La liste des sujets ou des thèmes, ainsi que des variables associées à ces thèmes, rehausse la recherche en révélant des variables dignes d'intérêt dont le client n'avait pas connaissance antérieurement. Toutefois, le résultat devrait être le même, que le client effectue les recherches au moyen de mots-clés ou de façon thématique. Autrement dit, le client doit être orienté vers la *source* de l'information ou des données recherchées.

3.3 Une seule porte d'accès : un seul instrument

Nous savons par expérience que certains clients comparent le Bureau de la statistique à un dédale déconcertant de sources en apparence illogiques. Les nombreux appels débutant par «Je ne sais pas si je m'adresse à la personne qui convient, mais auriez-vous...?» que j'ai reçus au fil des ans témoignent de cette situation. Il ne devrait exister qu'une seule porte d'accès au bureau. À cette porte ne devrait se trouver qu'*un seul* instrument, facile à utiliser, ou du personnel bien informé, muni de l'instrument et capable d'orienter le client vers les sources appropriées. Divers systèmes pourraient soutendre cet instrument, à condition que la présentation et le comportement des commandes demeurent les mêmes.

La porte d'accès pourrait être reproduite à divers emplacements, mais, encore une fois, elle devrait avoir partout la même présentation. Elle pourrait être électronique et complètement automatisée, ou dotée de personnel préposé à l'orientation des clients. Pour ce qui est d'un site Internet, il convient d'être vigilant afin de canaliser l'esprit d'entreprise et de contenir les manifestations du moi qui, par le truchement des «pages d'accueil», ont fait de ce média électronique une maison d'édition à compte d'auteur florissante. On devrait évaluer chacune de ces initiatives en regard du coût de la création et de la mise à jour, ainsi que de la contribution réelle à la recherche du client. Nous devons éviter le piège consistant à élaborer des solutions de type «tuyaux de poêle» pour résoudre les problèmes que cause ce type même d'organisation.

Durant le symposium, au moins deux communications présentées par des collègues de Statistique Canada décriront des éléments de la solution ultime. Il s'agit d'une part, du système IPS, et d'autre part, de Statcan en direct. Ces systèmes représentent des solutions partielles qui devront éventuellement être intégrées à une stratégie collective unique.

3.4 En direct, en temps réel

Très vite, l'Internet a augmenté considérablement les attentes quant à l'obtention de renseignements. Seule une réponse électronique immédiate donne désormais satisfaction. Si l'Internet représente un moyen idéal d'aider les clients à survoler nos métadonnées, la question se pose de savoir comment fournir un produit ou un service réel à ceux qui découvrent ce qu'ils cherchent. À mesure que se développent les marchés à créneaux, les clients se satisfont moins de produits génériques et demandent des données personnalisées, adaptées spécialement à leurs besoins.

Après avoir orienté un client vers une source de données pertinente, il est aussi profitable pour le client que pour le Bureau d'offrir au premier la possibilité de télécharger, en direct et en temps réel, les renseignements recherchés. Ce type de service répond manifestement aux intérêts du client, mais sert aussi ceux du Bureau qui, en plus de compter des clients satisfaits, enregistrera une baisse des coûts de base. Le Bureau épargnera d'autant plus ses ressources que la capacité du client à survoler, à préciser, à codifier ou à télécharger les données sera grande. La technologie actuelle permet d'offrir aux clients le téléchargement, avec facturation automatique, du contenu des fichiers de microdonnées accessibles au public. Il n'est nécessaire de distancer le client des données que dans le cas des fichiers-maîtres confidentiels (qui doivent demeurer derrière des murs coupe-feu et être filtrés pour éviter les divulgations par recoupements). Mais, même dans ces circonstances, rien n'empêche le client de coder la demande à partir de clichés d'enregistrement et de la soumettre au Bureau qui se chargera de produire les données de sortie et d'effectuer les filtrages indispensables pour éviter les divulgations.

La question de la facturation et du coût imputable au

client, quoique fascinante, dépasse largement le cadre de la présente communication et du symposium.

Enfin, au client qui ne possède pas les compétences voulues ou ne dispose pas du temps nécessaire pour télécharger lui-même les données dont il a besoin, on devrait offrir les services de chargés de programmes qui, au moyen des mêmes outils, produiront des données de sortie personnalisées satisfaisant ses besoins particuliers. En outre, à mesure que les archives de données s'ouvriront au monde entier grâce à la méta-information, il est probable que des conseillers du secteur privé profiteront de l'occasion pour offrir à leurs clients des services de consultation, de téléchargement et d'analyse des données.

3.5 Résolution des problèmes de discordance

Il est utopique de croire que toutes les discordances entre sources peuvent être éliminées. Des variations de méthodologie, telles que le fait de poser une question à la porte d'entrée, par téléphone ou grâce à un questionnaire rempli par le répondant, peuvent produire des écarts subtils entre les données. Néanmoins, des efforts concertés permettent d'éliminer les discordances les plus importantes. Il y a quelques années, j'ai participé à des travaux visant à harmoniser des données sur la famille en provenance de neuf ou dix sources. Des négociations entre les secteurs de production ont permis d'éliminer toutes les discordances importantes et la plupart des discordances mineures. Ce genre d'exercice ne se limite toutefois pas à un effort ponctuel, car, à mesure que de nouvelles sources entrent en ligne de compte, de nouvelles discordances apparaissent. Lors des travaux en question, une des tâches les plus gigantesques a consisté simplement à identifier toutes les sources de données sur la famille. En fait, nous avons dû établir un répertoire de ces sources avant de pouvoir déceler les discordances et y remédier. À cet égard, la production de méta-informations facilite le dépistage. Ainsi, durant la production récente de métadonnées sur les statistiques sociales, nous avons relevé de nombreuses discordances qui ont toutes été marquées d'un repère en vue d'un futur examen du problème. La méta-information peut également devenir un exemple de bonnes pratiques de gestion des données et, même, servir de modèle pour la mise au point de documentation normalisée, allant des mnémoniques utilisées dans les clichés d'enregistrement aux définitions, en passant par les systèmes de classification. L'adoption de modèles et de normes permettent aussi d'envisager la réduction des coûts de base, à mesure que de nouvelles sources de données sont mises au point. Les secteurs sources ne peuvent néanmoins se dérober aux discussions et aux négociations destinées à élaborer ces normes. Qui plus est, ils doivent être résolus à éliminer les discordances.

3.6 Multiplication des données de sortie thématiques

La production de méta-information facilitera également l'intégration des données selon un mode thématique. Par le passé, les analystes n'étaient pas toujours au courant de toutes les sources de données existantes. Aujourd'hui, en revanche, à condition de disposer des métadonnées, des instruments de recherche et des systèmes d'extraction appropriés, il n'y a aucune raison de ne pouvoir acheminer toutes les données pertinentes jusqu'à l'ordinateur de bureau. Toutefois, l'analyste doit bien saisir l'importance de l'intégration. Au minimum, les données de sortie agrégées ou tabulaires devraient être accompagnées d'indicateurs attirant l'attention sur des sources de données connexes. Idéalement, l'analyse ou la discussion qui accompagne les données de sortie analytiques ou descriptives devrait englober toutes les données pertinentes. Nous devons prendre conscience du fait que la diffusion de données qui ne reflètent pas notre connaissance globale d'un dossier ou d'une population risquent de nuire autant au client que des réponses non décelées ou des erreurs de traitement. N'est-il pas curieux que le statisticien habituellement si prodigue de notes en bas de page sur les questions de méthodologie demeure aussi muet quand il s'agit de mentionner d'autres sources de renseignements pertinentes au client?

3.7 Initiative organisationnelle

Il reste à savoir si les mesures susmentionnées peuvent être prises sans initiative organisationnelle. Il paraît improbable que des changements surviennent tant que la culture organisationnelle soutiendra la production individuelle plutôt que collective. En effet, les employés des divers secteurs de production ne prendront vraisemblablement aucune initiative, à moins que les critères utilisés pour évaluer leur rendement ne les y incitent. Peut-être certains le feront-ils, créant ainsi un précédent que les autres seront forcés de suivre, sous peine de rester à la traîne. Mais, l'enjeu n'est-il pas trop important pour abandonner de tels développements au hasard d'actions individuelles? Ne risque-t-on pas ainsi des chevauchements et le gaspillage des ressources? Le manque de vision commune, de planification stratégique, d'orientation et de financement ne donne-t-il pas l'impression que l'entreprise n'accorde pas une très grande importance à l'intégration?

Les bureaux de la statistique ne pourront relever les défis que leur pose aujourd'hui la technologie de

l'information, donc en concrétiser les promesses, que s'ils accordent une très haute priorité à l'intégration. Ainsi faudra-t-il sans doute établir et financer au sein de l'organisation un organe centralisé chargé de mener à bien les activités susmentionnées.

4. CONCLUSION

4.1 Le passé

Jusqu'à présent, la production de renseignements statistiques a surtout mis l'accent sur la méthodologie plutôt que sur le contenu thématique. La structure des silos dans lesquels sont collectées et archivées les données a entravé l'intégration de ces dernières en fonction des grandes questions, des populations ou de la distribution géographique, ainsi que les efforts visant à transformer ces données brutes en information. Enfin, ni l'organisation ni, d'ailleurs, le client n'ont réussi à saisir la richesse et la complétude des banques de données.

4.2 L'avenir

Aux yeux du statisticien, la situation idéale consisterait sans doute à effectuer le couplage intégral des enregistrements de toutes les sources de données, donc à réaliser une intégration complète. Malheureusement, les bureaux qui fonctionnent dans une société prête à tolérer une telle manipulation des renseignements personnels sont rares, voire inexistants. Par conséquent, le défi, et la promesse, consistent à passer de la gestion fragmentée à la gestion globale des données. La création de méta-information, l'harmonisation et l'intégration thématique sont des éléments essentiels à la transition des données brutes à l'information. Les bureaux qui refusent de relever le défi que pose la technologie de l'information et de saisir les occasions qui se présentent, et qui continuent à ignorer les attentes de leurs clients, particulièrement ceux qui ont déjà emprunté l'autoroute de l'information et ont aimé l'expérience, seront rapidement perçus comme peu serviables et sans intérêt.

5. BIBLIOGRAPHIE

Bradley, B. (1994). Metadata matters: Standardizing metadata for improved management and delivery in national information systems, document de travail, Ottawa, Santé Canada.

Hammer, M., et Champy, J. (1993). *Reengineering the Corporation,* New York: Harper-Collins.

Nordbotten, S. (1993). Communication inédite. The Statistical Meta Information System Workshop, Luxembourg: Eurostat.

Probst, S. (1995). Discours - programme, Data Warehouse Symposium, Ottawa: Tanning Technology Corporation.

Tapscott, D., et Caston, A. (1994). *Paradigm Shift: The New Promise of Information Technology,* New York: McGraw-Hill.

Vozel, K. (1993). Technical evolution white paper, document de travail, New York: AT&T.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

MÉTA-ANALYSE DE COHORTES MULTIPLES DE MINEURS EXPOSÉS AU RADON

Y. Wang[1], D. Krewski[1,2], J.H. Lubin[3] et J.M. Zielinski[1]

RÉSUMÉ

Le présent article décrit comment effectuer la méta-analyse d'une série d'études de cohorte, au moyen de modèles à effets aléatoires. Plus exactement, on s'est servi d'un modèle de régression à effets aléatoires non linéaire pour décrire le risque moyen pour la population, d'une part, et les risques spécifiques de la cohorte, d'autre part. Les méthodes utilisées ont permis d'ajuster le modèle du risque relatif proportionnel servant à estimer les liens entre l'exposition au radon et la mortalité par cancer du poumon au niveau de la dose-réponse. L'ajustement du modèle à effets aléatoires non linéaire exigeant de nombreux calculs, on a aussi envisagé une méthode de régression à deux degrés pour la méta-analyse. Les résultats obtenus avec le modèle à effets aléatoires et l'analyse de régression à deux degrés sont ensuite comparés aux résultats des méthodes classiques de méta-analyse, dans lesquels l'estimation du risque global repose sur la combinaison linéaire du risque spécifique des cohortes, après pondération (poids inversement proportionnels à la précision de l'estimation).

MOTS CLÉS : Mortalité par cancer; étude de cohortes; équations d'estimation globale; cancer du poumon; produits de filiation du radon; modèle à effets aléatoires.

1. INTRODUCTION

Méta-analyse des études épidémiologiques sur les maladies professionnelles. La méta-analyse est une méthode d'agrégation quantitative (Greenland, 1994). On s'en sert pour faire ressortir les effets globaux des études combinées et les écarts entre les études individuelles. Les méthodes classiques de méta-analyse établissent efficacement l'issue moyenne des études existantes en attribuant aux estimations individuelles des poids inversement proportionnels à l'erreur d'estimation (Greenland et coll., 1992). De nos jours, on a tendance à utiliser des modèles à effets aléatoires pour la méta-analyse (Berlin et coll., 193; Berkey et coll., 1995). Le Conseil national de recherches (1992) préconise cette approche pour la méta-analyse et la recherche des sources de variation au niveau des résultats des études. La supériorité de l'analyse des effets aléatoires sur les techniques classiques de méta-analyse vient du fait qu'on peut tenir compte dans une certaine mesure des sources d'hétérogénéité au-delà de l'erreur d'échantillonnage (Greenland, 1994). L'analyse des effets aléatoires donne une estimation agrégative pour l'ensemble des études et des estimations spécifiques pour chacune d'elles. L'estimation globale illustre les effets fixes, alors que les estimations spécifiques dégagent les effets aléatoires (Moolgavkar et coll., 1995).

Dans cet article, nous nous intéresserons à l'application d'un modèle à effets aléatoires à la méta-analyse des études épidémiologiques sur les maladies professionnelles. Notre travail a été motivé par la nécessité de faire une synthèse des liens entre l'exposition au radon et la mortalité par cancer du poumon dévoilés par 11 études majeures sur les mineurs entreprises au Canada, aux États-Unis et ailleurs dans le monde, mais caractérisées par une grande hétérogénéité (Lubin et coll., 1994). Le radon est un gaz inerte dégagé par le l'uranium lors de sa désintégration radioactive. Les particules alpha émises par les produits de filiation à courte vie du radon expliquent la cancérogénicité du gaz. La méta-analyse des études sur les mineurs

[1] Services des ressources humaines, Santé Canada, Ottawa (Ontario), Canada.

[2] Department of Mathematics and Statistics, Carleton University, Ottawa (Ontario), Canada.

[3] Biostatistics Branch, National Cancer Institute, Bethesda, Maryland, É-U.

Table 1. Caractéristiques de 11 études sur des mineurs[a]

Lieu	Type de Mine	Nombres de mineurs	Durée du suivi	Nombre d'année-personnes	Cas de cancer du poumon
Chine	Étain	17 143	1976-87	175 342	980
Tchécoslovaquie	Uranium	4 284	1952-90	107 868	661
Colorado	Uranium	3 347	1950-87	82 435	329
Ontario	Uranium	21 346	1955-86	380 718	291
Terre-Neuve	Fluorine	2 088	1950-84	48 742	118
Suède	Fer	1 294	1951-91	33 293	79
Nouveau-Mexique	Uranium	3 469	1943-85	58 949	69
Beaverlodge	Uranium	8 486	1950-80	118 385	65
Port Radium	Uranium	2 103	1950-80	52 676	57
Radium Hill	Uranium	2 103	1948-87	51 850	54
France	Uranium	1 785	1948-86	44 043	45
Total		67 746	1943-91	1 151 315	2 736

[a]Cité dans le rapport du NCI (Lubin et coll., 1994).

effectuée par Lubin et ses collaborateurs (1994) a établi hors de tout doute que les descendants radioactifs du radon sont cancérogènes et que l'exposition à ces derniers au niveau observé chez les mineurs accroît les risques de cancer du poumon. Le tableau 1 résume de façon succincte les caractéristiques des différentes études. Bien que ces dernières démontrent toutes une hausse significative des risques de cancer du poumon avec l'intensité de l'exposition au radon, l'estimation du risque relatif excédentaire par niveau opérationnel-mois (RRE/NOM) d'exposition au radon varie sensiblement avec la cohorte. Par conséquent, il est nécessaire de tenir compte de l'hétérogénéité des études. Nous explorons ici les méthodes pour la méta-analyse des études de cohortes. La partie 2 décrit les méthodes d'analyses à effet aléatoires utilisées. À la partie 3, on aborde la question des méthodes d'analyse à deux degrés. Une utilisation de ces méthodes est illustrée à la partie 4. Les conclusions apparaissent à la partie 5.

2. MODÈLES STATISTIQUES

On utilise souvent les méthodes de régression de Poisson pour analyser les données des études de cohortes sur la mortalité (Breslow et Day, 1987). On suppose pour cela que le taux de mortalité au cours d'une période précise et pour un degré d'exposition établi est constant. Les données servant aux analyses de régression sont présentées sous forme d'un tableau d'une année-personne à paramètres multiples comprenant le nombre de décès résultant de la maladie à laquelle on s'intéresse, et des années-personnes d'observations classifiés selon les covariable pertinentes. En vertu du modèle de régression de Poisson, on suppose que le nombre de cas observés suit la distribution de Poisson, selon laquelle la variance est égale à la moyenne. Plus précisément, le nombre de cas prévus est modélisé comme suit :

$$N_{jk}\, r_{jk}\, (x,\, v), \qquad (1)$$

où N_{jk} indique le nombre d'années-personnes courant des risques pour le $j^{\text{ième}}$ état de la $k^{\text{ième}}$ cohorte, et où

$r_{jk}(x,v)$ représente le taux de mortalité associé à la valeur vectorielle des covariables v et des variables confusionnelles potentiels x.

2.1 Modèle du risque relatif proportionnel

Le modèle du risque relatif proportionnel permet d'exprimer le taux de mortalité par le produit suivant :

$$r_{jk}(x, v) = r_{0jk}(x)\, RR_{jk}(v;\alpha_k), \qquad (2)$$

où $r_{0jk}(x)$ correspond au taux de mortalité de fond du $j^{ième}$ état de la $k^{ième}$ cohorte, où $RR_{jk}(v;\alpha_k)$ est le risque relatif connexe et où α_k est un vecteur des paramètres du modèle.

Dans les études sur la mortalité par cancer de travail, on décrit souvent l'association entre les cas et les facteurs de risque au moyen d'une relation linéaire entre l'exposition et la réponse à celle-ci, soit :

$$RR_{jk}(v;\alpha_k) = 1 + \beta_k \times w_j \times \xi_k, \qquad (3)$$

où β_k représente la pente, w_j le niveau d'exposition aux facteurs de risques étudiés et ξ_k le vecteur des covariables qui modifieront la relation entre exposition et réponse. Quand on stratifie l'exposition continue en L groupes, la relation décrite plus haut peut être représentée par le modèle par catégories :

$$RR_{jk}(v;\alpha_k) = 1 + \beta_{l,k} \times \xi_k$$

(l=1, ..., L), où $\beta_{l,k}$ correspond au risque relatif excédentaire spécifique au degré d'exposition. Des méthodes approximatives pour ajuster des modèles non-linéaires aux données seront présentées aux parties 2.3 et 2.4. Ces méthodes sont exactes dans le cas particulier de linéarité.

2.2 Effets fixes et effets aléatoires

On décrit l'hétérogénéité entre les cohortes grâce à un modèle à effets aléatoires (Rutter et Elashoff, 1994), en vertu duquel les effets globaux et la variation entre les cohortes sont caractérisés par des coefficients de régression fixes et aléatoires, respectivement. Plus exactement, pour décrire l'hétérogénéité des cohortes, on décompose le risque relatif excédentaire du cancer du poumon attribuable à l'exposition au radon de la $k^{ième}$ cohorte, β_k, en deux éléments :

$$\beta_k = \beta + b_{\beta,k}, \qquad (4)$$

où β correspond à l'effet fixe de l'ensemble des cohortes et $b_{\beta,k}$, à l'effet aléatoire spécifique à la $k^{ième}$ cohorte. En règle générale, on peut caractériser les paramètres des modèles (3) ou (4) de la façon suivante :

$$\alpha_k = \alpha + b_k, \qquad (5)$$

où α représente la valeur vectorielle des effets fixes applicables à toutes les cohortes et où la valeur vectorielle des effets aléatoires b_k, à moyenne nulle, donne l'écart entre la $k^{ième}$ cohorte et les effets globaux.

2.3 Moments marginaux

En général, $RR_{jk}(v;\alpha_k)$ peut ne pas être une fonction linéaire. Pour calculer l'espérance inconditionnelle et la variance du nombre de cas observés dans le modèle à effets aléatoires, on suppose que l'espérance de tous les effets aléatoires est nulle. Étant donné b_k, la fonction de risque relatif correspond approximativement à la suivante

$$\begin{aligned} RR_{jk}(v;\alpha|b_k) &\sim RR_{jk}(v;\alpha) + \frac{\partial RR_{jk}(v;\alpha)}{\partial\alpha} \times b_k \\ &= RR_{jk}(v;\alpha)(1 + z_{jk}b_k), \end{aligned} \qquad (6)$$

où

$$z_{jk} = RR_{jk}^{-1}(v;\alpha) \times \frac{\partial RR_{jk}(v;\alpha)}{\partial\alpha}. \qquad (7)$$

Les effets aléatoires b_k étant connus, l'espérance conditionnelle et la variance du nombre de décès observés au $j^{ième}$ état de la $k^{ième}$ cohorte dans le modèle de régression de Poisson correspondent à

$$E(y_{jk}|b_k) = N_{jk} r_{jk}(x,v;\alpha|b_k) \qquad (8)$$

et

$$Var(y_{jk}|b_k) = \phi E(y_{jk}|b_k). \qquad (9)$$

où le paramètre de dispersion supplémentaire ϕ décrit la variation excédentaire du nombre de décès observés. Soit $D = Cov(b_k)$, la matrice de covariance des effets aléatoires. Les moments marginaux du $j^{ième}$ état de la $k^{ième}$ cohorte peuvent être exprimés comme suit :

$$E(y_{jk}) = \mu_{jk}(\alpha) \sim N_{jk} r_{ojk}(x)\, RR_{jk}(v;\alpha), \qquad (10)$$

$$Var(y_{jk}) = \sigma_{jk}(\alpha) = \phi\,\mu_{jk}(\alpha) + \mu_{jk}^2(\alpha)\, z_{jk}^T(\alpha) Dz_{jk}(\alpha). \quad (11)$$

et

$$ov(y_{jk}, y_{ik}) = \sigma_{jik}(\alpha) = \mu_{jk}(\alpha)\,\mu_{ik}(\alpha) z_{jk}^T(\alpha) Dz_{ik}(\alpha \quad (12)$$

$(j \neq i)$. La matrice des covariances des effets aléatoires est une matrice non-négative définie inconnue, que l'on doit estimer lors de l'ajustement du modèle. Supposons que $Y_k = (y_1, \ldots, y_{J_k})^T$ est le vecteur des observations dans la $k^{ième}$ cohorte et que $\Omega = diag(D, \ldots, D)$ et $Z_k = diag(z_{1k}, \ldots, z_{J_k k})$. Supposons aussi que $\Lambda_k(\alpha) = diag(\mu_{1k}(\alpha), \ldots, \mu_{J_k k}(\alpha))$ et $\Sigma_k(\alpha) = \{\sigma_{ijk}(\alpha)\}$ représentent deux matrices $J_k \times J_k$ pour lesquelles les paramètres $\mu_{jk}(\alpha)$ et $\sigma_{ijk}(\alpha)$ apparaissent aux équations (10) - (12). On peut exprimer la matrice des covariances des valeurs vectorielles des observations de la $k^{ième}$ cohorte (Y_k) de la manière suivante :

$$Cov(Y_k) = \Sigma_k(\alpha) = \phi\ \Lambda_k(\alpha) + \Lambda_k(\alpha)\ Z_k^T(\alpha)\,\Omega\, Z_k(\alpha)\ \Lambda_k(\alpha). \quad (13)$$

2.4 Équations d'estimation globale

Zeger et Lang, (1988) recourent à des équations d'estimation globale (EEG) pour ajuster les modèles à effets aléatoires aux données longitudinales. Cette approche autorise un certain assouplissement des hypothèses de distribution, et les calculs sont souvent plus simples que l'estimation du maximum de vraisemblance. Une fois que les spécifications de la matrice des covariances provisoires ont été arrêtées, la méthode des EEG procure une estimation cohérente et asymptotiquement normale lorsque les conditions sont légèrement régulières, bien qu'on puisse noter une perte d'efficacité.

Supposons que la $k^{ième}$ cohorte comprenne J_k états $(k=1, \ldots, K)$. Si $\mu_k(\alpha)=E(Y_k)$, les EEG des effets fixes α, la covariance des effets aléatoires $\mathbf{D} = Cov(b_k)$ étant connu, sont

$$\sum_{k=1}^{K} \frac{\partial \mu_k^T(\alpha)}{\partial \alpha} \Sigma_k^{-1}(\alpha)(Y_k - \mu_k(\alpha)) = 0 \quad (14)$$

(Zeger et coll., 1988). On peut résoudre cette équation pour α avec l'itération de Newton-Raphson. Soulignons que l'équation d'estimation (14) n'est pas sans biais puisque $\Sigma_k^{-1}(\alpha)$ est une fonction des paramètres α du modèle (Burnett et coll., 1995). Pour obtenir une estimation non biaisée de α on devrait ajouter un facteur de pénalité à (14) afin que l'espérance de l'équation d'estimations soit nulle.Cette dernière équation d'estimation a reçu le nom de «fonction de quasi-vraisemblance avec pénalité» (Breslow et Clayton, 1993). L'usage du facteur de pénalité augmente néanmoins la somme de calculs nécéssaires à l'ajustement du modèle. Cependant, le biais pourrait être négligeable lorsqu'il s'agit d'échantillons de taille.

Comme le recommandent Zeger et Liang (1988), on se sert de l'approximation (11) pour obtenir une estimation préliminaire de la matrice des covariances des effets aléatoires (D). Bref, on exprime l'équation (11) de la façon suivante :

$$= (z_{jk} z_{jk}^T)^{-1} z_{jk} \left(\frac{E(y_{jk} - \mu_{jk}(\alpha))^2 - \phi\,\mu_{jk}(\alpha)}{\mu_{jk}^2(\alpha)}\right) z_{jk}^T (z_{jk} z_{jk}^T)^{-}$$

L'estimateur de moment

$$\hat{D} = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{J_k}\sum_{j=1}^{J_k} (z_{jk} z_{jk}^T)^{-1} \{z_{jk}[\frac{(y_{jk} - \hat{\mu}_{jk}(\hat{\alpha}))^2 - \hat{\phi}\hat{\mu}_{jk}(\hat{\alpha})}{\hat{\mu}_{jk}^2(\hat{\alpha})}] z_{jk}^T\}\ (z_{jk}(z_{jk}^T)^{-1} \quad (15)$$

permet d'estimer la matrice des covariances des effets aléatoires.

On recourt à l'estimateur de moment

$$\hat{\phi} = \sum_{k=1}^{K} \frac{1}{KJ_k} \sum_{j=1}^{J_k} \{\frac{(y_{jk} - \hat{\mu}_{jk}(\hat{\alpha})^2 - \hat{\mu}_{jk}(\hat{\alpha}) z_{jk} \hat{D} z_{jk}^T}{\hat{\mu}_{jk}^2(\hat{\alpha})}\} \quad (16)$$

pour obtenir une estimation du paramètre de dispersion supplémentaire ϕ (Zeger et Liang, 1986). Pour estimer α et (D, ϕ), on résout d'abord (14) au moyen des méthodes de notation de Fisher, (D, ϕ) étant fixé à sa valeur estimative $(\hat{D}, \hat{\phi})$. On évalue ensuite les équations (14) en prenant la nouvelle estimation α du paramètre $\hat{\alpha}$ et en procédant à une itération jusqu'à ce qu'il y ait convergence.

Lors de ce processus itératif, il est possible d'estimer les effets aléatoires associés à la $k^{ième}$ cohorte en se servant de la valeur estimative de $\hat{\alpha}$ pour compenser les effets fixes des paramètres du modèle de risque relatif dans l'équation (5) et en utilisant la matrice des covariances estimatives des effets aléatoires $\hat{\Sigma}_k$ afin de résoudre l'équation de notation de la quasi-vraisemblable

$$\frac{\partial \mu_k^T(b_k)}{\partial b_k} \hat{\Sigma}_k^{-1} [Y_k - \hat{\mu}_k(b_k)] = 0 \quad (17)$$

pour $\mathbf{b}_k$ (k = 1, ..., K). On peut accommoder la dispersion supplémentaire associée à certaines cohortes de la manière suggérée par McCullagh et Nelder (1989).

On obtient une estimation des effets fixes $\hat{\alpha}$, une estimation des effets aléatoires $\hat{b}_k (k=1,...,K)$ et la matrice des covariances des effets aléatoires $\hat{D}$ en résolvant d'abord (14) selon la méthode de Newton-Raphson, les valeurs retenues pour D et $\mathbf{b}_k$ correspondant respectivement à leurs valeurs initiales $\hat{D}$ et $\hat{b}_k$. Cela fait, on évalue les équations (15) et (16) en donnant à α sa nouvelle valeur estimative $\hat{\alpha}$ et en procédant à une itération jusqu'à ce qu'il y ait convergence. Zeger et Liang (1986) ont remarqué que les estimations convergentes des paramètres (α^*) sont cohérentes et proposent l'estimation robuste de la variance

$$Cov(\alpha^*) = \Gamma_0^{-1} \Gamma_1 \Gamma_0^{-1}, \tag{18}$$

où

$$\Gamma_0 = \sum_{k=1}^{K} \frac{\partial \hat{\mu}_k^T(\alpha^*)}{\partial \alpha^*} \hat{\Sigma}_k^{-1}(\alpha^*) \frac{\partial \hat{\mu}_k(\alpha^*)}{\partial \alpha^*}$$

et

$$\Gamma_1 = \sum_{k=1}^{K} \frac{\partial \hat{\mu}_k^T(\alpha^*)}{\partial \alpha^*} \hat{\Sigma}_k^{-1}(\alpha^*) [Y_k - \mu_k(\alpha^*)]^T [Y_k - \mu_k(\alpha^*)] \hat{\Sigma}_k^{-1}(\alpha^*) \frac{\partial \hat{\mu}_k(\alpha^*)}{\partial \alpha^*}.$$

L'estimation de la variance de $\hat{b}_k$ est donc

$$Cov(\hat{b}_k) = T_k^{-1} P_k T_k^{-1} \tag{19}$$

où

$$T_k = \frac{\partial \mu_k^T(\hat{b}_k)}{\partial \hat{b}_k} \hat{\Sigma}_k^{-1} \frac{\partial \mu_k(\hat{b}_k)}{\partial \hat{b}_k} \tag{20}$$

et

$$P_k = \frac{\partial \mu_k^T(\hat{b}_k)}{\partial \hat{b}_k} \hat{\Sigma}_k^{-1} [Y_k - \hat{\mu}_k(\hat{b}_k)]^T [Y_k - \hat{\mu}_k(\hat{b}_k)] \hat{\Sigma}_k^{-1} \frac{\partial \mu_k(\hat{b}_k)}{\partial \hat{b}_k}. \tag{21}$$

3. ANALYSE DE RÉGRESSION À DEUX DEGRÉS

Selon le nombre de covariables catégoriques, le nombre d'états du tableau d'années-personnes d'une cohorte peut être très élevé. L'ajustement du modèle pourrait donc s'avérer difficile à calculer. Dans la présente partie, nous examinerons les méthodes d'analyse de régression à deux degrés faciles à appliquer malgré un nombre élevé de variables confusionnelles (Whitehead et Whitehead, 1991). Sans perdre de vue la généralisation, nous utiliserons un modèle linéaire simple

$$RR_{jk}(v, \omega; \alpha_k) = 1 + \beta_k \times \omega_j \tag{22}$$

où pour illustrer la méthode à deux degrés les paramètres β_k indiquent le risque relatif excédentaire de la $k^{ième}$ cohorte.

Premier degré. Lors du premier degré , on ajuste le modèle (22) à chaque cohorte. Soient $\hat{\beta}_k$ l'estimation du paramètre β_k du modèle et s_k, l'estimation de la variance de $\hat{\beta}_k$. Les estimations $\{\hat{\beta}_k, s_k; k = 1,...,K\}$ de l'analyse au premier degré sont utilisées pour l'analyse au *deuxième degré*.

Deuxième degré. Soient

$$\bar{\beta} = \frac{\sum_k s_k^{-1} \hat{\beta}_k}{\sum_k s_k^{-1}}, \tag{23}$$

$$\hat{\tau} = \frac{\sum_k s_k^{-1} (\hat{\beta}_k - \bar{\beta}_k)^2 - (K-1)}{\sum_k s_k^{-1} - \frac{\sum_k s_k^{-2}}{\sum_k s_k^{-1}}}, \tag{24}$$

et

$$w_k = \frac{(\hat{\tau} + s_k)^{-1}}{\sum_k (\hat{\tau} + s_k)^{-1}}. \tag{25}$$

L'estimation $\hat{\beta}$ de l'effet global des cohortes est exprimée par

$$\hat{\beta} = \sum_k w_k \hat{\beta}_k \tag{26}$$

La variance de l'effet global estimatif est estimée au moyen de

$$Var(\hat{\beta}) = (\sum_k (\hat{\tau} + s_k)^{-1})^{-1}. \tag{27}$$

L'homogénéité de $\hat{\beta}_k$ dans les cohortes peut être statistiquement testée par

$$\chi^2_{homog} = \sum_k s_k^{-1} (\hat{\beta}_k - \bar{\beta}_k)^2, \tag{28}$$

qui correspond à une distribution chi carré à K - 1 degrés de liberté.

L'estimateur d'érosion de l'effet spécifique à la cohorte β_k est donné par

$$\beta_k^* = \frac{s_k \hat{\beta} + \hat{\tau}\, \hat{\beta}_k}{s_k + \hat{\tau}}, \qquad (29)$$

l'écart avec l'estimation globale étant donné par

$$\hat{\delta}_k = \hat{\beta} - \hat{\beta}_k^*, \qquad (30)$$

à condition que $\hat{\tau}>0$.

La variance estimative de l'écart est exprimée comme suit :

$$Var(\hat{\delta}_k) = \frac{\hat{\tau} s_k}{\hat{\tau} + s_k}. \qquad (31)$$

L'hétérogénéité des cohortes est décrite par τ: des valeurs positives pour τ accroîtront la variance estimative de l'effet global $\hat{\beta}$.

4. ILLUSTRATION

Dans cette partie, nous nous servirons des méthodes décrites aux parties 2 et 3 pour analyser les données de 11 grandes études sur les mineurs (Colorado, Tchécoslovaquie, Chine, Ontario, Terre-Neuve, Suède, Nouveau-Mexique, Beaverlodge, Port Radium, Radium Hill et France). Lubin et coll., (1994) ont déjà effectué une méta-analyse de ces études. Nous n'avons pas l'intention d'effectuer un autre analyse étendue sur ces études, mais voulons simplement illustrer les méthodes.

En tout, les 11 études en question portent sur plus de 2 700 cas de cancer du poumon chez 68 000 mineurs représentant pratiquement 1,2 million d'années-personnes d'observations (Lubin et coll., 1994). Nous inspirant de Lubin et coll., nous avons stratifié l'incidence de fond du cancer du poumon selon l'âge (toutes les études) réparti en groupes de cinq ans (<40, 40-44, 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, ≥ 75 ans), l'exposition à d'autres maladies liée à la profession (disponible pour des études en Chine, en Ontario, du Colorado, du Nouveau-Mexique et en France), un indicateur de l'exposition aux produits de filiation du radon (Beaverlodge) et l'ethnicité (Nouveau-Mexique).

Le modèle du risque relatif

$$E(y_{jk} \mid b_k) = N_{jk} r_{0jk}(x)[1 + (\beta + b_k) \times w] \qquad (32)$$

servira à l'exemple, pour lequel on négligera l'effet modificateur des autres covariables pour raisons de simplicité.

Le paramètre β_k exprime le risque excédentaire par NOM pour l'exposition au radon. On suppose qu'une exposition récente au radon revêt plus d'importance qu'une exposition ancienne, si bien que l'exposition cumulative au gaz dans (22) est répartie en trois éléments, $w = w_{5\text{-}14} + \theta_2 w_{15\text{-}24} + \theta_3 w_{25+}$ ou $w_{5\text{-}14}$, $w_{15\text{-}24}$ et w_{25+} indiquent respectivement l'exposition au radon subie il y a 5 à 14 ans, 15 à 24 ans et plus de 25 ans. Les paramètres θ_2 et θ_3 de l'effet du temps écoulé depuis l'exposition sont traités comme des effets fixes dont la valeur est respectivement estimée à 0,78 et 0,10 d'après les données des 11 études.

La première colonne du tableau 2 donne les valeurs estimatives de β_k pour l'ensemble des études et les études individuelles, selon les méthodes d'analyse des effets aléatoires. La deuxième présente l'estimation du même facteur selon l'analyse de régression à deux degrés. À la dernière colonne du tableau 2, on peut voir les estimations spécifiques à l'étude obtenues par les méthodes classiques de méta-analyse. Selon l'analyse des effets aléatoires, l'estimation globale du risque relatif excédentaire est de 0,49% avec une erreur-type de 0,14%. L'analyse à deux degrés débouche sur une estimation de 0,56% et une erreur-type de 0,15%, et la méthode classique de méta-analyse, sur une estimation globale de 0,57% (erreur-type de 0,018%). L'analyse des effets aléatoires et l'analyse de régression à deux degrés tendent à produire une erreur-type plus importante que l'analyse classique parce qu'elles tiennent compte de l'hétérogénéité des études. L'analyse des effets aléatoires comme l'analyse de régression à deux degrés parviennent à des estimations analogues du RRE spécifique à chaque étude, se rapprochant du RRE estimatif global.

5. CONCLUSION

Dans cet article, les auteurs présentent des méthodes permettant de procéder à la méta-analyse d'une série d'études de cohorte, en vue d'évaluer des risques pour la santé similaires, associés à la profession. Il y est plus spécifiquement question de l'analyse des effets aléatoires et de l'analyse de régression à deux degrés. L'article examine aussi une approche plus classique à la méta-analyse reposant sur la combinaison linéaire des estimations spécifiques à chaque étude, après pondération (les poids étant inversement proportionnels à la précision de l'estimation). Les trois méthodes semblent donner une estimation comparable des effets globaux des études concernées, dans l'exemple servant d'illustration.

Tableau 2. RRE/NOM estimatif en %[a] selon l'analyse des effets aléatoires et les analyses spécifiques à la cohorte

Étude	Analyse des effets aléatoires	Analyse à deux degrés	Analyse spécifique à la cohorte
Combiné	0,49[b]±(0,14)[e]	0,56[c]±0,12[e]	0,57[d]±0,02[e]
Chine	0,25[f]	0,29[g]	0,28[h]
Tchécoslovaquie	0,75	0,67	0,71
Colorado	0,55	0,46	0,44
Ontario	0,52	0,73	0,94
Terre-Neuve	0,74	0,80	1,18
Suède	0,53	0,62	1,68
Nouveau- Mexique	0,64	0,60	2,66
Beaverlodge	0,53	0,64	4,26
Port Radium	0,42	0,42	0,38
Radium Hill	0,50	0,58	6,74
France	0,46	0,35	0,06

[a] RRE/NOM correspond au paramètre β ajusté au moyen de l'équation $RR = 1 + \beta \times w^*$, où $w^* = w_{5\text{-}14} + \theta_2 w_{15\text{-}24} + \theta_3 w_{25+}$ indique l'exposition cumulative au radon, $w_{5\text{-}24}$, $w_{15\text{-}24}$ et w_{25+} correspondent respectivement à l'exposition survenue il y a 5 à 14 ans, 15 à 24 ans et plus de 25 ans. Les valeurs θ_2 et θ_3 ont été estimées à 0,78 et 0,10, respectivement.

[b] Coefficient fixe du modèle des effets aléatoires.

[c] Estimation globale de l'analyse à deux degrés.

[d] Moyenne pondérée des estimations spécifiques à la cohorte (les poids correspondent à la valeur inverse de l'erreur-type des estimations.)

[e] Erreur-type des estimations.

[f] Somme des coefficients fixes et aléatoires du modèle des effets aléatoires.

[g] Érosion estimative spécifique à la cohorte selon l'analyse à deux degrés.

[h] Estimation spécifique à la cohorte.

L'analyse des effets aléatoires utilise toutes les données disponibles des études, contrairement à la méta-analyse classique et à l'analyse de régression à deux degrés qui ne recourent qu'à la synthèse des estimations sommaires des études individuelles. L'avantage de l'analyse des effets aléatoires et de l'analyse de régression à deux degrés sur les méthodes classiques de méta-analyse est qu'elles tiennent compte de l'hétérogénéité des études dans l'estimation du risque global.

Il s'agit d'un avantage particulièrement important quand les résultats de l'étude présentent des écarts appréciables, c'est-à-dire quand le modèle à effets fixes permet difficilement l'ajustement des données (Greenland, 1994). Sur le plan des calculs, l'analyse de régression à deux degrés peut s'avérer plus pratique que l'analyse des effets aléatoires lorsque la base de données est très importante.

6. REMERCIEMENTS

Nous aimerions remercier John D. Boice, Christer Edling, Richard W. Hornung, Geooffrey Howe, Emil Kunz, Robert A. Kusiak, Howard I. Morrison, Edward P. Radford, Jonathan M. Samet, Margot Trimarche, Alistair Woodward, Yao Shu Xiang, et Donald A. Pierce pour la permission accordée d'utiliser les données sur les mineurs de l'exemple à la partie 4.

BIBLIOGRAPHIE

Berlin, J.A., Longnecker, L.M., et Greenland, S. (1993). Meta-analysis of epidemiologic dose-response data, *Epidemiology*, 4, 218-228.

Berkey, C.S., Hoaglin, D.C., Mosteller, F., et Colditz, G.A. (1995). A random-effects regression model for meta-analysis, *Statistics in Medicine*, 14, 395-411.

Breslow, N.E., et Day, N.E (1987). *Statistical Methods in Cancer Research*, Vol. 2: The Design and Analysis of Cohort Studies. Centre international de recherche sur le cancer, Lyon.

Burnett, R.T., Ross, W.H., et Krewski, D. (1995). Nonlinear Mixed Regression Models. *Environmetrics*, 6, 85-99.

Burnett, R.T. (1995). Two-Stage Mixed Regression Model Approach, manuscrit inédit.

Greenland, S. (1994). Invited Commentary: A critical look at popular meta-analysis methods. American *Journal of Epidemiology*, 140, 290-296.

Greenland, S., et Longnecker, M.P. (1992). Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American Journal of Epidemiology*, 135, 1301-1309.

Laird, N.M., et Ware, J.H. (1982). Random-effects models for longitudinal data, *Biometrics*, 38, 963-974.

Lubin, J.H., Boice, J.D., Edling, C., Hornung, R.W., Howe, G., Kunz, E., Kusiak, R.A., Morrison, H.I., Radford, E.P., Samet, J.M., Trimarche, M., Woodward, A., Xiang, Y.S., et Pierce, D.A. (1994). Radon and Lung Cancer Risk: A Joint Analysis of 11 Underground Miners Studies. *National Institutes of Health, NIH Publication*, 94-3644.

Lubin, J.H. (1994). Invited Commentary: Lung Cancer and Exposure to Residential Radon. American *Journal of Epidemiology*, 140, 323-332.

McCullagh, P., et Nelder, J.A. (1989). Generalized Linear Model, Chapman & Hall, New York, (1989).

Moolgavkar, S.H. (1995). When and how to combine results from multiple epidemiological studies in risk assessment, manuscrits inédits.

National Research Council (1992). Combining Information: Statistical Issues and Opportunities for Research, National Academy Press, Washington, D.C.

Rutter, C.M., et Elashoff, R.M. (1994). Analysis of Longitudinal Data: Random Coefficient Regression Modelling, *Statistics in Medicine*, 13, 1211-1231.

Whitehead, A., et Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. Statistics in Medicine, 10, 489-499.

Whittemore, A.S., et McMillan, A. (1983). Lung Cancer Mortality Among U.S. Uranium Miners: A reappraisal, *Journal of National Cancer Institute*, 71, 489-499.

Zeger, S., Liang, K.Y., et Albert, P.S. (1988). Models for longitudinal Data: A General Estimating Equation Approach, *Biometrics*, 44, 1049-1060.

Zeger, S., et Liang, K.Y. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes, *Biometrics*, 42, 121-130.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

COUPLAGE DES DONNÉES POUR CRÉER L'INFORMATION

W. Winkler et F. Scheuren[1]

RÉSUMÉ

Le couplage des enregistrements tirés de deux ou de plusieurs fichiers ne date pas d'hier, surtout au Canada. On a accordé une importance primordiale à l'optimisation de l'étape de couplage et au contrôle des erreurs inévitables (Newcombe et coll., 1959; Fellegi et Sunter, 1969). Les progrès réalisés en matière d'estimation des probabilités de couplage réel (p. ex., Belin et Rubin, 1995; Winkler, 1994) ont été un des éléments moteurs indispensables à l'analyse. Nos travaux récents (Scheuren et Winkler 1993) sur les données appariées en présence d'erreurs de couplage ont porté principalement sur les interactions entre les erreurs de couplage et l'analyse de régression, les ajustements étant conditionnés par l'état de nos connaissances sur le couplage. Dans le présent article, nous poursuivrons cette analyse, mais en portant cette fois une attention particulière à une méthode récursive plus complete. Les deux premières étapes demeurent les mêmes qu'auparavant : 1) estimer les probabilités d'erreurs; 2) corriger les analyses pour tenir compte des erreurs de couplage. L'étape suivante utilise l'analyse comme telle en temps que source additionnelle d'information. À l'intérieur d'un processus récursif, les valeurs abérrantes de la régression sont assimilées à des non-liens puis les probabilités d'appariement sont réajustés et les paramètres de l'analyse de regression sont réestimés.

MOTS CLÉS : Vérification; imputation; couplage d'enregistrements; analyse de régression; processus récursif.

1. INTRODUCTION

Les chercheurs et les décideurs ont souvent besoin d'informations plus complètes que ce qu'ils peuvent trouver dans une seule base de données. L'utilisation des micro-données provenant de deux ou de plusieurs fichiers peut toutefois conduire à des erreurs lorsque le seul moyen d'établir un lien entre des paires d'enregistrements consiste à utiliser des identificateurs qui ne sont pas uniques, comme des noms et des adresses.

Les progrès récents réalisés dans le domaine de la vérification-imputation (**VI**) et du couplage d'enregistrements (**CE**) fournissent selon nous des outils assez puissants pour réaliser des analyses qui, jusqu'à maintenant, étaient infaisables. Deux des raisons de ce changement constituent le fondement des méthodes proposées dans le présent article :

- Premièrement, et en dépit de certains problèmes, les analystes sont généralement en mesure de définir les meilleures façons de vérifier (c.-à-d., corriger) et d'imputer (c.-à-d., compléter) les micro-données existantes. Ceci est vrai, selon nous, tant pour les fichiers de données individuels que pour les fichiers regroupés provenant de diverses sources. Ce qui est nouveau, cependant, c'est que les analystes sont aujourd'hui plus nombreux à utiliser des sous-programmes réutilisables puissants fondés sur des modèles de vérification comme celui de Fellegi et Holt (1976). Ces modèles et ces logiciels facilitent et systématisent le processus de **VI** et remplacent les sous-programmes **SI a ALORS b SINON c** conçus pour des bases de données particulières et qui doivent de ce fait être récrits dans chaque application.
- Deuxièmement, les analystes peuvent effectivement coupler un ensemble représentatif d'enregistrements provenant de fichiers de données distincts grâce aux nouvelles méthodes améliorées proposées à l'origine par Newcombe (Newcombe et coll., 1959) et dont la formalisation mathématique est due à Fellegi et Sunter (1969). Dans un article antérieur (Scheuren et Winkler 1993), nous avons démontré une méthode d'ajustement de **CE** qui permet de réduire le biais dû aux erreurs de couplage des enregistrements dans les analyses de régression.

---

[1] William Winkler, U.S. Bureau of the Census, Washington, DC 20225 et Fritz Scheuren, The George Washington University, Washington, DC 20056.

Dans le présent article, nous présentons une méthode récursive en quatre étapes très puissante, et qui s'utilise avec la même facilité. Pour lancer le processus, nous utilisons une méthode de **CE** améliorée (Winkler 1995; Belin et Rubin, 1995) pour délimiter un ensemble de paires d'enregistrements pour lesquelles le taux estimatif d'erreurs d'appariement est très faible. Nous tentons une analyse de régression. Ensuite, nous utilisons un modèle **VI** élaboré à partir des enregistrements couplés à taux d'erreurs faible, pour vérifier et imputer les valeurs aberrantes de ce qui reste des paires couplées. Une seconde analyse de régression (**AR**) est effectuée et cette fois, les résultats sont réutilisés pour l'étape du couplage de manière que le **CE** puisse être amélioré (et ainsi de suite). Le cycle se poursuit jusqu'à ce que les résultats analytiques désirés cessent de changer. Cette procédure peut être représentée schématiquement comme suit :

Ces trois méthodes sont, bien évidemment, déjà largement utilisées. Ce sur quoi nous souhaitons insister, dans le présent article, c'est une méthode utile pour les intégrer et en accroître de ce fait encore plus l'utilité.

Outre l'introduction, le présent article comporte quatre sections. Dans la section 2, nous présentons un bref aperçu des techniques de vérification-imputation et de couplage des enregistrements. Dans la section 3, nous décrivons les fichiers de données empiriques établis et les analyses de régression réalisées. Dans la section 4, nous présentons les résultats obtenus et dans la section 5, nous présentons nos conclusions et soulignons les aspects qui méritent de plus amples recherches.

2. EXAMEN DES MÉTHODES DE VÉRIFICATION-IMPUTATION ET DE COUPLAGE DES ENREGISTREMENTS

Dans la présente section, nous procédons à un bref examen des méthodes de vérification-imputation (**VI**) et de couplage des enregistrements (**CE**). Nous n'avons pas l'intention de les décrire en détail, mais simplement d'établir un cadre pour notre démonstration. Comme l'analyse de régression (**AR**) est bien connue, nous nous contenterons d'en décrire notre utilisation particulière (Section 3).

2.1. Vérification-imputation

Historiquement, les méthodes de **vérification** des micro-données ont émergé principalement pour corriger les incohérences logiques observées dans les bases de données (p. ex., Nordbotten, 1963). Leur utilisation pour la détection des entrées invraisemblables ou improbables a également été importante (p. ex., Granquist, 1984). Les méthodes d'**imputation** des micro-données ont d'abord servi principalement pour régler les problèmes de données manquantes (p.ex., Little et Rubin, 1987). Les praticiens utilisent déjà les deux méthodes en combinaison depuis longtemps. Les méthodes de vérification-imputation (**VI**) n'ont toutefois pas été entièrement conceptualisées en un seul et même système avant la publication de l'article pionnier de Fellegi et Holt (1976). Il a fallu attendre encore plus longtemps avant qu'on ne réussisse à informatiser complètement le système de Fellegi-Holt. Ce n'est que très récemment, avec l'avènement de la génération actuelle de matériels et logiciels informatiques, qu'on peut se satisfaire des résultats obtenus.

Même si nous porterons uniquement notre attention sur les données continues dans le présent article, les techniques de **VI** conviennent également pour les données discrètes et les combinaisons de données discrètes et continues. Pour les besoins de la démonstration, imaginons un ensemble de données continues. Dans ce cas, la vérification pourrait consister en une série de règles pour chaque enregistrement qui s'exprimeraient comme suit

$$c_1X < Y < c_2X \qquad (2.1)$$

ce qui signifierait que

> **si Y est plus petit que c_1X ou plus grand que c_2X, il conviendra alors de réviser les enregistrements de données.** (2.2)

Dans cette formule, **Y** peut représenter le total des salaires, **X** le nombre d'employés et c_1 et c_2 des constantes telles que $c_1 < c_2$.

Les méthodes de vérification se sont fondées sur des séquences de règles **SI a ALORS b SINON c**, sur des méthodes statistiques elles-mêmes fondées sur des tests de détection des valeurs aberrantes et sur le modèle de Fellegi et Holt (1976), lequel contient généralement d'autres types de modèles pour des cas spéciaux. Les principaux avantages du modèle Fellegi-Holt sont : 1) l'inspection systématique d'un système de vérification de

la cohérence logique avant la réception des données; 2) la détermination de la quantité minimale d'informations qui doivent être modifiées dans un enregistrement de manière à ce que l'enregistrement révisé respecte les règles de vérification; 3) Le modèle se sert comme outil de tables de décision. Les règles de vérification contenues dans ces tables et dans les sous-programmes principaux de vérification sont réutilisables par la suite. Pour les items qui doivent être changés, un algorithme d'imputation est intégré dans le système; il doit satisfaire aux vérifications et peut être utilisé pour remplacer les entrées jugées incohérentes.

Les systèmes Fellegi-Holt généraux actuels, utilisables sur toutes sortes d'ordinateurs, sont le Système généralisé de vérification et d'imputation (SGVI) pour des règles de vérification basées sur des inégalités linéaires (p. ex., Kovar, Whitridge et MacMillan, 1991), le nouveau système de programmes structurés de vérification et de référence économiques (SPEER) du Bureau of the Census pour les régles de vérification utilisant des ratios de données continues (p. ex., Winkler et Draper, 1996) et le système DISCRETE du Bureau of the Census pour la vérification des données discrètes générales (p. ex., Winkler et Petkunas, 1996). L'imputation est réalisée à l'aide de techniques aujourd'hui normalisées (p. ex., Little et Rubin, 1987) qui sont souvent des variantes de la méthode «Hot Deck».

2.2. Couplage des enregistrements

La méthode de couplage des enregistrements cherche à classifier les paires de l'espace issu du produit cartésien $\mathbf{A} \times \mathbf{B}$ provenant de deux fichier A et B en un ensemble M de liens réels et un ensemble U de non-liens réels. Formalisant les notions introduites par Newcombe (p. ex., Newcombe et coll., 1959), Fellegi et Sunter (1969) ont imaginé des rapports **R** de probabilités prenant la forme

$$\mathbf{R} = \Pr(\gamma \in \Gamma \mid \mathbf{M}) / \Pr(\gamma \in \Gamma \mid \mathbf{U}) \qquad (2.3)$$

où γ désigne un schéma particulier de concordance dans un espace de comparaison Γ. Par exemple, Γ pourrait désigner huit schémas particuliers représentant la présence ou l'absence d'une concordance simple concernant le prénom, le nom et l'âge. Par ailleurs, chaque $\gamma \in \Gamma$ pourrait en plus rendre compte de la fréquence relative avec laquelle des noms particuliers, tels que Scheuren ou Winkler, apparaissent. Les champs ainsi comparés (nom, prénom et âge) sont appelés variables d'appariement.

La règle de décision prend la forme suivante:

Si $\mathbf{R} >$ *supérieur*, la paire est ainsi assimilée à un lien.
Si *inférieur* $\leq \mathbf{R} \leq$ *supérieur*, la paire est alors assimilée à un lien possible, sous réserve d'un examen.
Si $\mathbf{R} <$ *inférieur*, la paire est assimilée à un non-lien. (2.4)

Fellegi et Sunter (1969) ont montré le caractère optimal de cette règle de décision du fait que pour chaque paire de limites fixées pour **R**, la région intermédiaire est minimisée pour toutes les règles de décision dans un espace de comparaison Γ donné. Les seuils limites *supérieur* et *inférieur* sont déterminés par les bornes fixées sur l'erreur. Nous appelons le rapport **R** ou toute transformation monotone croissant (typiquement, un logarithme) un poids d'appariement ou un poids total de concordance totale.

Comme les méthodes de **VI**, les techniques de **CE** ont beaucoup progressé grâce à l'avènement d'outils informatiques accessibles et peu coûteux. Au cours des quelque dix dernières années, les techniques de couplage des enregistrements ont fait l'objet de très nombreux articles (p. ex., Jaro, 1989; Newcombe, Fair et Lalonde, 1992). Certains de ces travaux ont découlé d'une série de conférences organisées à partir du milieu des années 1980 (p. ex., Kilss et Alvey, 1985; Carpenter et Fair, 1989). Aux États-Unis, l'attention portée à l'étude du sous-dénombrement dans le recensement décennal de 1990 a aussi beaucoup influé sur ces recherches (p. ex., Winkler et Thibaudeau, 1991). Finalement, l'ouvrage pionnier de Newcombe (1988) a également joué un rôle important dans l'intérêt suscité par cette question.

3. MÉTHODE DE SIMULATION

Pour réaliser nos simulations, nous avons envisagé quatre scénarios d'appariement comme dans notre travail antérieur (Scheuren et Winkler, 1993). L'idée fondamentale consistait à générer des données aux propriétés de distribution connues, d'adjoindre ces données à deux fichiers qui seraient appariés, puis d'évaluer l'effet d'une quantité croissante d'erreurs d'appariement sur les analyses. Nous avons commencé avec deux fichiers (de 12 000 et de 15 000 données) présentant de bonnes informations d'appariement et dont nous connaissions quels étaient les liens réels. En fait, avant l'introduction d'erreurs, environ 10 000 de ces données étaient des liens réels soit une proportion d'environ 83 % pour le fichier plus petit ou fichier de base.

Nous avons ensuite généré des données empiriques aux propriétés de distribution connues, et adjoint ces données aux fichiers. Pendant la conduite des simulations, diverses erreurs ont été introduites dans les variables d'appariement, des quantités différentes de données ont été utilisées pour l'appariement et des écarts plus grands par rapport aux probabilités d'appariement optimales ont été admis. Ces variations sont décrites ci-après et illustrées à la figure 1. Pour chaque scénario de la figure, le poids d'appariement, soit le logarithme de **R**, est porté sur l'axe horizontal, tandis que la fréquence, aussi exprimée en logarithmes, est portée sur l'axe vertical. Les appariements (ou liens réels) sont marquées d'astérisques (*), tandis que les non-appariements (ou non-liens réels) sont marqués de petits cercles (o) :

Bon scénario (figure 1a). Nous avions conclu antérieurement qu'aucun ajustement n'est nécessaire ici pour tenir compte de l'erreur d'appariement. Ce scénario peut se réaliser dans les systèmes conçus pour l'appariement, dont les variables d'appariement sont bonnes et qui utilisent des algorithmes d'appariement sophistiques. Le taux véritable d'appariements erronnés était ici inférieur à 2 %.

Scénario médiocre (figure 1b). Le scénario d'appariement médiocre consistait à utiliser le nom, le prénom et l'initiale, deux variantes de l'adresse, le numéro d'appartement ou d'unité et l'âge. Des fautes typographiques mineures ont en outre été introduites d'une façon indépendante pour un nom sur sept et un prénom sur cinq. Les probabilités d'appariement ont été choisies afin de s'écarter de la valeur optimale, en étant toutefois comparable aux probabilités que pourraient choisir un expert en appariement informatisé. Le taux réel d'appariement erronés était de 6,8 %.

Premier mauvais scénario (figure 1c). Le premier mauvais scénario d'appariement consistait à utiliser le nom, le prénom, une variation de l'adresse et l'âge. Des fautes typographiques mineures ont été introduites indépendamment pour un nom sur cinq et un prénom sur trois. Des erreurs typographiques modérément graves ont été introduites pour le quart des adresses. Les probabilités d'appariement ont été choisies de manière à s'écarter substantiellement de la valeur optimale. On cherchait ici à reproduire la situation d'un praticien appelé à faire un tel choix mais possédant peu d'expérience en cette matière. Le taux réel d'appariements erronnés était ici de 10,1 %.

Second mauvais scénario (figure 1d). Le second mauvais scénario d'appariement consistait à utiliser le nom, le prénom et une variation de l'adresse. Des erreurs typographiques mineures ont été introduites indépendamment pour un nom sur trois et un prénom sur trois. Des erreurs typographiques graves ont été introduites pour le quart des adresses. Les probabilités d'appariement ont été choisies de manière à s'écarter substantiellement de la valeur optimale. On cherchait ainsi à reproduire une situation qui survient avec des listes d'entreprises, lorsque le responsable du couplage a peu d'emprise sur la qualité des listes. Le taux réel d'appariements erronnés était ici de 14,6 %.

Selon les divers scénarios envisagés, notre aptitude à distinguer les liens réels et les non-liens réels varie sensiblement. Dans le cas du bon scénario, nous constatons que la dispersion des points pour les liens et les non-liens réels est presque complète (Figure 1a). Dans le cas du scénario médiocre, les nuages de points correspondants se recoupent modérément (Figure 1b); dans le cas du premier mauvais scénario, le recoupement est important (Figure 1c) et, dans le cas du second mauvais scénario, le recoupement est presque total (Figure 1d).

Les véritables taux d'erreur d'appariement de **CE** peuvent être estimés raisonnablement bien à l'aide de la méthode de Belin et Rubin (1995), sauf dans le cas du second mauvais scénario où cette méthode ne converge pas. En pratique, pour ce scénario, la portion des données pour lesquelles on pourrait différencier les bons des mauvais liens sans procéder à des opérations supplémentaires est presque inexistante. Jusqu'ici, une analyse fondée sur le second scénario n'aurait pas été jugée utile, même dans la meilleure des hypothèses. Toutefois, comme nous le verrons à la Section 4, il est possible d'intervenir utilement même dans ce cas.

Après avoir précisé les situations de concordance ci-dessus, nous avons utilisé le Système SAS pour générer des données issues du modèle $\mathbf{Y} = 4\,\mathbf{X} + \epsilon$ et obtenues par la méthode de moindre carré ordinaire. Les valeurs de **X** ont été choisies pour être réparties uniformément entre 1 et 101, et les termes d'erreur ϵ étaient normaux et homoscédastiques, avec une variance de 4 000 de manière à ce que la régression de **Y** sur **X** ait une valeur R^2 de 78 % dans la population des liens réels. Seuls les résultats du second mauvais scénario sont présentés ici en détail. Ils sont, de très loin, les plus renversants.

Figure 1a. Bon scénario d' appariement

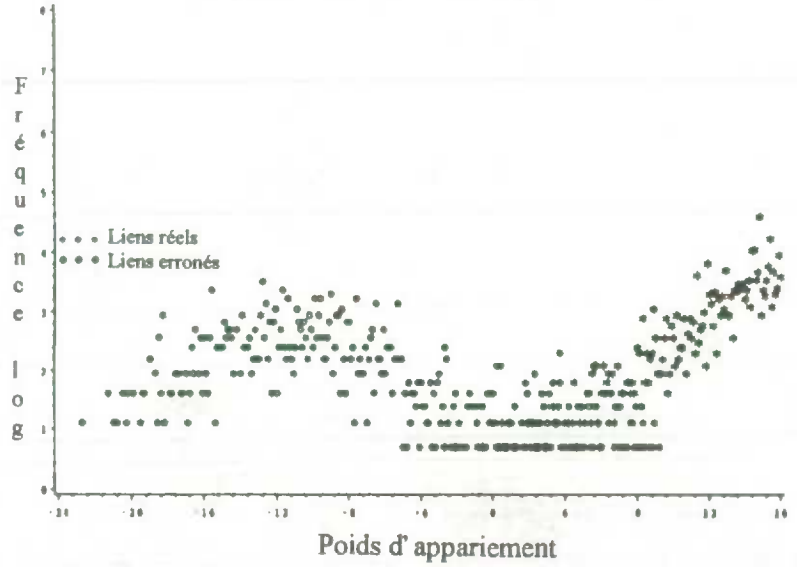


Figure 1b. Scénario d' appariement médiocre

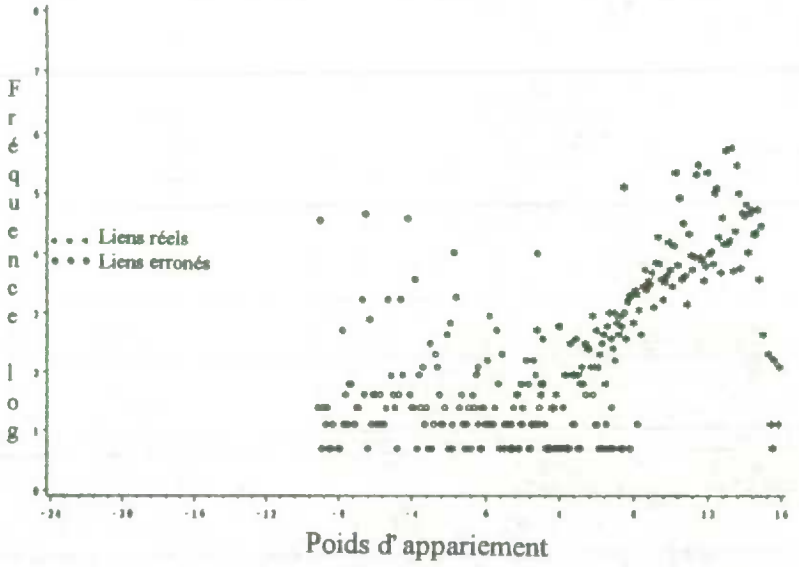


Figure 1c. Premier mauvais scénario d' appariement

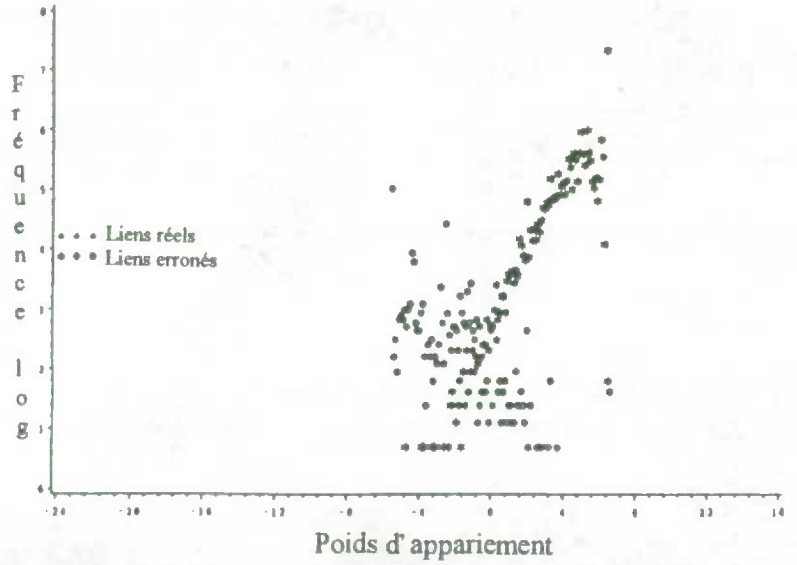


Figure 1d. Second mauvais scénario d' appariement

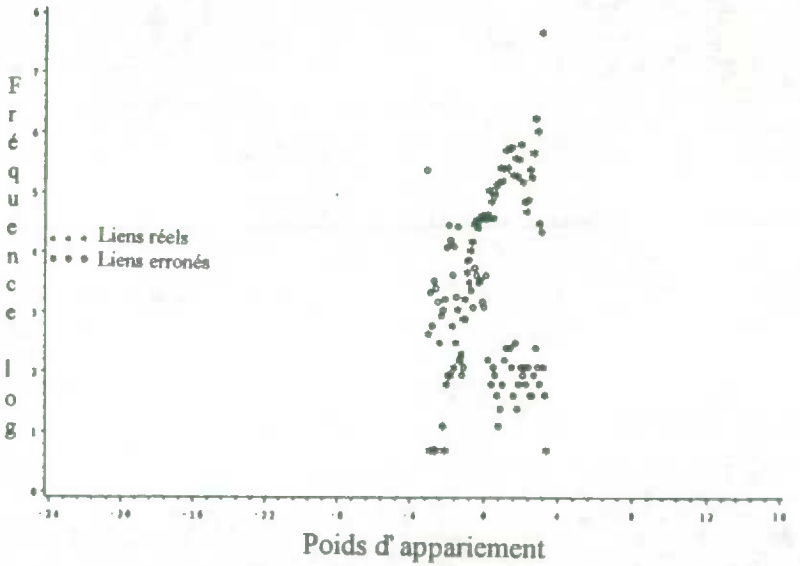


Figure 2. Second mauvais scénario d'appariement, premier essai
Tous les liens erronés et 5% des liens réels donnés de la régression
1606 observations; coefficient bêta = 3,99; R^2 = 0,78

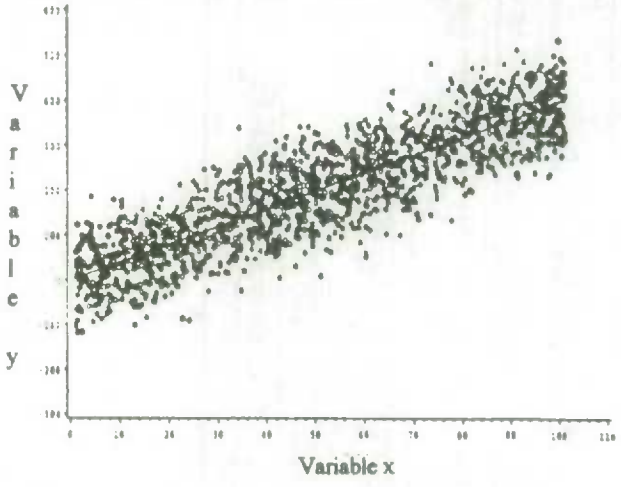


Figure 3. Second mauvais scénario d'appariement, premier essai
Tous les liens erronés et 5% des liens réels donnés observeées
1606 observations; coefficient bêta = 1,18; R^2 = 0,07

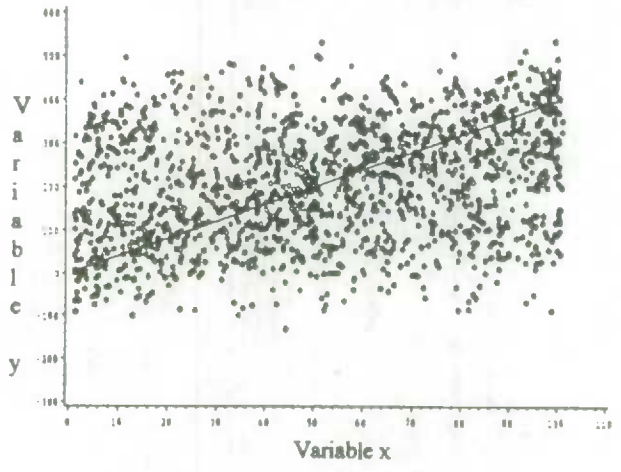


Figure 4. Second mauvais scénario d'appariement, premier essai
Tous les liens erronés et 5% des liens réels, correction des donnés
1606 observations; coefficient bêta = 3,46; R^2 = 0,75

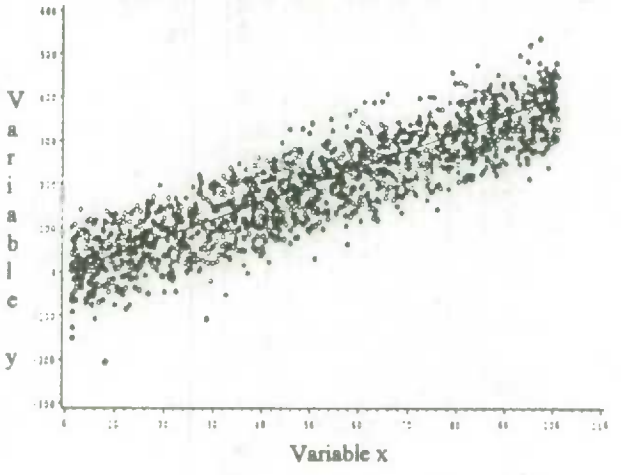


Figure 5. Second mauvais scénario d'appariement, deuxième essai
Tous les liens erronés et 5% des liens réels donnés de la régression
véritable, 1104 observations; coefficient bêta = 3,94; R^2 = 0,77

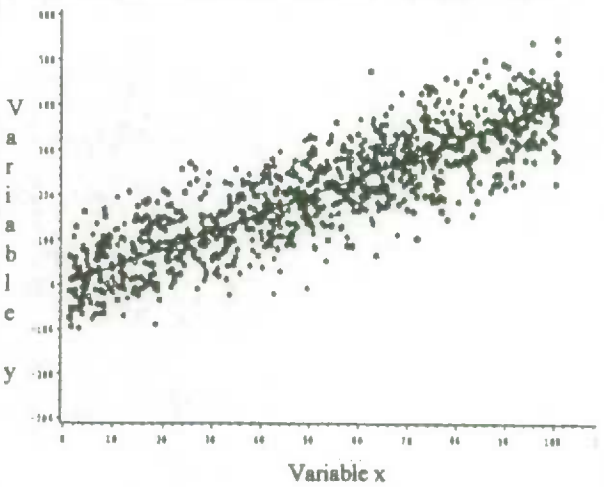


Figure 6. Second mauvais scénario d'appariement, deuxième essai
Tous les liens erronés et 5% des liens réels donnés observées
1104 observations; coefficient bêta = 3,64; R^2 = 0,65

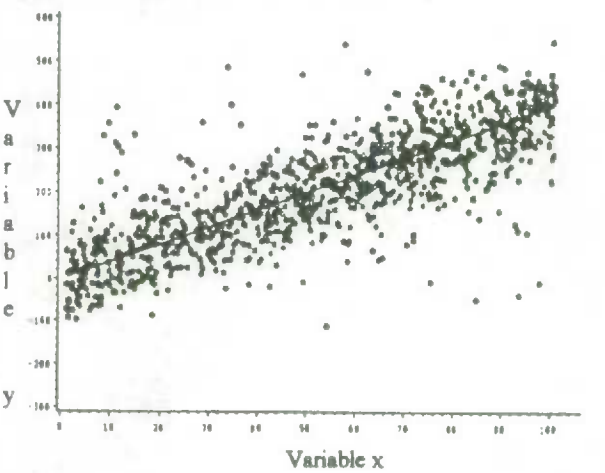
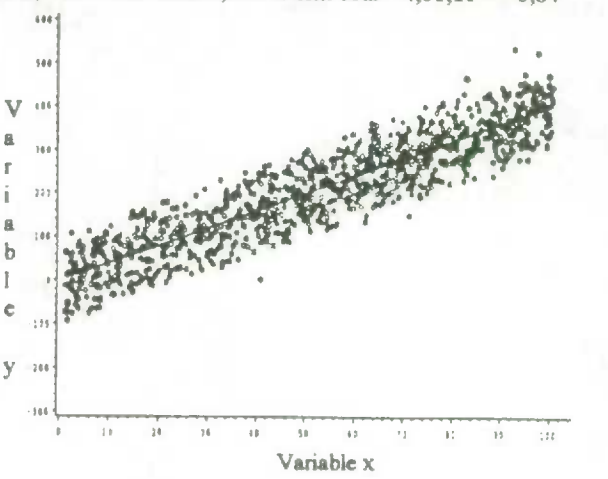


Figure 7. Second mauvais scénario d'appariement, deuxième essai
Tous les liens erronés et 5% des liens réels donnés corrigées - valeurs
aberrantes, 1104 observations; coefficient bêta = 4,01; R^2 = 0,84



4. PROCESSUS RÉCURSIFS ET RÉSULTATS

Nous examinons ici les graphiques obtenus à l'aide du processus récursif pour le second mauvais scénario. Les résultats de la régression sont présentés en deux cycles. En outre, pour faciliter la compréhension, nous présentons les graphiques dans des figures séparées pour chaque étape.

4.1 Résultats du premier cycle

4.1.1 Régression véritable (à titre de référence). Le graphique de la figure 2 illustre le nuage de points des valeurs de **X** et **Y** qu'on obtiendrait s'il n'y avait pas d'erreur de couplage. Il convient de souligner que tous les liens erronnés sont indiqués, mais que seulement 5 % des liens réels sont utilisés. Nous avons procédé ainsi pour empêcher les liens réels de dominer les résultats au point de masquer ce qui se passe. Deuxièmement, dans cette figure et dans toutes les autres, la droite de régression véritable est toujours indiquée à titre de référence. Finalement, la pente de la population véritable, ou coefficient **bêta** (3,99) et la valeur de R^2 (78 %) correspondent aux données affichées.

4.1.2 Régression après l'étape initiale **CE ⇒AR**. Nous présentons à la figure 3 les résultats de l'analyse de régression portant sur les liens tels qu'observés (pas ce qui aurait dû arriver dans des conditions idéales, mais ce qui arrive dans des cas réels). Nous ne constatons sans surprise qu'une faible relation de régression entre **Y** et **X**. La pente observée, ou coefficient **bêta**, s'écarte largement de sa vraie valeur (1,18 au lieu de 3,99). La mesure de l'ajustement est elle aussi touchée, passant de 78% à 7 %.

4.1.3 Régression après l'étape combinée **CE⇒AR⇒VI ⇒AR**. La figure 4 complète notre présentation des résultats du premier cycle de notre processus récursif. Nous avons ici corrigé les données du graphique comme suit. Premièrement, en n'utilisant que les 183 cas assortis d'un poids d'appariement de 3,00+, nous avons cherché à améliorer les mauvais résultats présentés à la figure 3. En utilisant cet ajustement provisoire, nous avons obtenu des valeurs prévues pour tous les cas appariés; ensuite, les valeurs aberrantes assorties d'une variance résiduelle égale ou supérieure à 100 ont été éliminées et la régression a été recalculée avec les paires restantes. Cette nouvelle équation se présentait essentiellement sous la forme $\mathbf{Y} = 3\mathbf{X} + \epsilon$, avec un écart-type de 3 000. En utilisant notre méthode antérieure, (Scheuren et Winkler, 1993), nous avons à nouveau ajusté le coefficient **bêta** de 3,0 à 3,4. Si une paire d'enregistrements appariés produisait une valeur aberrante, les valeurs prévues à l'aide de l'équation $\mathbf{Y} = 3{,}4\mathbf{X}$ étaient alors imputées. Si une paire ne donnait pas de valeur aberrante, on utilisait alors la valeur observée en guise de valeur prévue.

4.2 Résultats du second cycle

4.2.1 Régression véritable (à titre de référence). La figure 5 présente le nuage de points de **X** et **Y**, tel qu'il apparaîtrait s'il pouvait s'agir de liens réels après une seconde étape de **CE**. Dans cette seconde étape, nous avons utilisé les valeurs prévues de **Y** déterminées ci-dessus. Ainsi, le couplage s'appuyait sur plus d'informations et on pouvait donc compter sur un groupe différent de données couplées après cette seconde étape. En particulier, comme nous avons obtenu un couplage bien meilleur, le nombre de liens érronés était moindre. Notre échantillon constitué de l'ensemble de tous les liens érronés et de 5 % des liens réels est donc passé de 1 606, dans les figures 2 à 4, à 1 104 dans les figures 5 à 7. Dans cette seconde itération, la pente véritable, ou coefficient **bêta**, et les valeurs de R^2 sont demeurées, pourtant, essentiellement identiques tant pour la pente (3,94 contre 3,99) que pour l'ajustement (77 % contre 78%).

4.2.2 Régression après la seconde étape **CE ⇒AR**. À la figure 6, nous constatons une très grande amélioration du rapport entre **Y** et **X** fondé sur les liens tels qu'observés après la seconde étape de **CE**. La pente est passée de sa valeur initiale de 1,18 à 3,64. Elle est encore trop faible, mais l'amélioration est importante. L'ajustement a lui aussi été touché, passant de 7 % à 65 %.

4.2.3 Régression après les étapes combinées **CE⇒AR⇒VI⇒AR**. La figure 7 termine notre présentation du second cycle du processus récursif. Dans ce graphique, nous avons corrigé les données comme suit. D'abord, en utilisant seulement les 185 cas présentant un poids d'appariement de 7,00+, nous avons cherché à améliorer encore davantage les résultats obtenus à la figure 6. En utilisant cet ajustement, nous avons obtenu une autre série de valeurs prévues pour l'ensemble des cas appariés. Cette nouvelle équation se présentait essentiellement sous la forme $\mathbf{Y} = 3{,}8\mathbf{X} + \epsilon$, avec un écart-type d'environ 2 000. En utilisant notre méthode antérieure (Scheuren et Winkler, 1993), nous avons procédé à un autre ajustement du coefficient **bêta**, le faisant passer de 3,8 à 4,0. Ici encore, si les paires de données appariées donnaient une valeur aberrante, les valeurs prévues en utilisant l'équation $\mathbf{Y} = 4{,}0\mathbf{X}$ étaient imputées. Si une paire ne donnait pas de valeur aberrante, on utilisait alors la valeur observée en guise de valeur prévue. Le graphique de la figure 7 présente les valeurs ajustées.

5. CONCLUSIONS ET QUESTIONS À EXAMINER PLUS À FOND

En principe, nous aurions pu poursuivre le processus récursif décrit ci-haut. En fait, dans un problème réel, il nous aurait fallu continuer jusqu'à ce que le rapport de **Y** sur **X** (c.-à-d., le coefficient **bêta**, dans cet exemple) cesse de changer de façon appréciable.

Il semble à première vue que nous devrions être heureux des résultats obtenus. Ils nous permettent d'obtenir une réponse passablement raisonnable dans une situation qui peut paraître, au départ, sans issue. Les résultats seraient-ils toujours aussi bons? C'est ce que nous croyons, mais nous nous pencherons sur cette question lors de travaux ultérieurs. Un examen plus approfondi nous permet dès à présent de constater un certain nombre de points où notre méthode s'avère plus faible qu'elle ne le devrait, ou simplement incomplète.

5.1 Surajustement

L'algorithme que nous utilisons a surajusté la relation entre **X** et **Y**. La valeur de R^2 après le second cycle était de 84 %, comparativement à 77 % dans la population. C'est là un problème commun de la régression lorsque les valeurs aberrantes sont éliminées; on souhaitera peut-être s'y résigner et l'ignorer tout simplement. Toutefois, plusieurs solutions sont envisageables à l'étape de la **VI** pour améliorer notre méthode.

Celle qui nous paraît la plus attrayante consiste à reprendre l'idée de Howard Newcombe et utiliser un échantillon de liens erronés connus. Dans notre travail antérieur (Scheuren et Winkler, 1993) portant sur ce problème, l'appariement que nous avons proposé nécessitait d'obtenir deux liens pour chaque cas du fichier de base. Le second lien serait un compromis et serait habituellement assimilable à un lien erronné. Comment ces seconds liens pourraient-ils être utilisés pour régler le problème du surajustement?

Présumons d'abord que l'appariement est suffisamment bon pour que les algorithmes de Belin-Rubin fonctionnent (Belin et Rubin, 1995). Nous pourrons alors calculer une probabilité de lien réel pour chaque appariement. Nous serons ensuite en mesure d'estimer le nombre de liens erronés parmi nos cas les mieux appariés. Ce nombre de cas pourrait être sélectionné à partir du fichier des appariements de rechange - peut-être simplement au hasard ou mieux, d'une manière équilibrée (de manière à ce que, par exemple, les moyennes de **X** et **Y** de ce fichier échantillon de liens erronés concordent avec les valeurs correspondantes du fichier original ou fichier des meilleurs appariements). L'étape ultérieure possible consisterait à apparier l'échantillon de liens erronés connus aux meilleurs liens initiaux et d'éliminer les paires «les plus proches». On pourrait procéder ainsi au lieu de chercher les valeurs aberrantes et d'éliminer toutes celles qui s'écartent d'une distance donnée du centre des points (tel que décrit en 4.1.3).

Même si les algorithmes de Belin-Rubin ne convergent pas au premier cycle, on pourra quand même, à la suggestion de Newcombe, utiliser un fichier de liens érronés lorsque le processus récursif aura donné des appariements de qualité suffisante pour le justifier. Dans l'exemple présent, cela serait devenu possible au second cycle, même si ce ne l'était pas au début.

5.2 Diagnostics

En vertu des hypothèses, jusqu'ici implicites, sous-jacentes à nos simulations, nous avons traité les variables d'appariement et leurs relations d'un fichier à l'autre comme des variables indépendantes de la relation (**X, Y**). Nous avons abordé cette question d'une manière plus détaillée dans notre travail antérieur (Scheuren et Winkler, 1993). Nous nous contenterons ici de souligner que les diagnostics échantillons devraient être examinés pour vérifier cette hypothèse. À chaque étape, il paraît raisonnable de calculer certaines statistiques univariées à partir des cas traités comme s'ils étaient appariés à cette étape : les moyennes et les médianes pour **X** et **Y**, et même les variances et les écarts moyens de **X** et **Y**. Il devrait même être possible d'utiliser ces valeurs pour nous protéger contre le surajustement. Cette question fera l'objet d'études ultérieures.

5.3 Un échantillon de questions ouvertes

Au cours de l'élaboration détaillée de notre méthode, nous avons dû procéder à un certain nombre d'ajustements **spéciaux** de l'étape de **VI**. Comment pourrait-on décider, par exemple, de la limite d'inclusion des cas appariés de manière à obtenir une régression provisoire? Qu'en est-il de l'utilisation d'une trace médiane comme point de départ au moyen duquel on peut identifier les valeurs aberrantes? Quelle raison a motivé notre choix du seuil de démarcation des valeurs aberrantes? Des limites plus souples auraient-elles contribué sensiblement à la réduction du surajustement?

L'étape du **CE** comportait elle aussi des éléments **spéciaux**. Nous avons déjà traité de nos tentatives d'introduire l'algorithme Belin-Rubin le plus tôt possible. Nous devrons nous pencher de façon beaucoup plus attentive sur la façon dont les résultats de l'**AR** ont été utilisés à partir du premier cycle de la seconde étape de **CE**. Par exemple, pourquoi ne pas utiliser la valeur de la

régression ajustée dans tous les cas et non pas seulement pour les valeurs aberrantes?

5.4 Généralisabilitions possibles

Nous avons examiné une régression simple portant sur une variable provenant d'un fichier et une autre variable provenant d'un autre fichier. Qu'arriverait-il si on généralisait cette méthode aux cas de régressions multiples? Nous travaillons actuellement sur cette question et avons bon espoir d'arriver à des résultats utiles.

Les possibilités de généralisation nous paraissent moins certaines dans d'autres cas. Par exemple, qu'arrive-t-il lorsque le rapport entre **Y** et **X** est faible dans la population. Peut-être, dans un tel cas, nous sera-t-il impossible d'améliorer l'appariement suffisamment pour rendre tout le travail décrit ci-haut valable? Qu'arrive-t-il lorsque le recoupement de deux fichiers est très faible (il était élevé dans l'exemple que nous avons utilisé)?

5.5 Technologie et théorie statistiques

Le présent article a porté sur les possibilités technologiques. Notre discussion n'a pas été indépendante des considérations théoriques, mais, à l'inverse, les fondements théoriques des idées explorées n'ont pas encore tous été élucidés. À une étape aussi précoce, l'approche intuitive n'est pas à négliger. Nous ne nous en excusons pas, mais préférons souligner qu'elle permet d'une certaine façon à tous les membres de notre auditoire (ou à nos lecteurs) d'intervenir. Notre travail antérieur avait notamment pour objectif de favoriser la participation des autres. Cet objectif nous tient toujours à coeur.

5.6 Des données à l'information

Lors de notre présentation, nous nous sommes arrêtés beaucoup plus longuement que dans le présent article sur la nécessité d'un processus récursif capable d'intégrer les efforts d'administration (obtenir des liens valables) et les efforts d'analyse (ajuster un rapport de régression). Nous sommes convaincus que la puissance d'analyse que nous procurent aujourd'hui les ordinateurs et les logiciels exige un changement des rôles respectifs des producteurs et des utilisateurs de données. Chacun se doit de devenir plus interactif, de mettre l'accent sur le véritable travail d'équipe. Il y aura un prix à payer pour ce changement, mais le prix sera encore plus lourd si nous continuons à nous entêter dans notre isolement.

BIBLIOGRAPHIE

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

Carpenter, et Fair, M.. (Editors) (1989). *Proceedings of the Record Linkage Sessions and Workshop*, Canadian Epidemiological Research Conference, in Ottawa, Ontario, Canada, August 30-31, 1989, Statistics Canada.

Fellegi, I., et Holt, T.(1976). A systematic approach to automatic edit and imputation, *Journal of the American Statistical Association*, 71, 17-35.

Fellegi, I., et Sunter, A. (1969). A theory of record linkage, *Journal of the American Statistical Association*, 64, 1183-121.

Granquist, L. (1984). On the role of editing, *Statistic Tidshrift*, 2, 105-118.

Jabine, T. B., et Scheuren, F. J. (1986). Record linkages for statistical purposes: methodological issues, *Journal of Official Statistics*, 2, 255-277.

Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida, *Journal of the American Statistical Association*, 84, 414-420.

Kilss, B., et Alvey, W., (Editors) (1985). *Record Linkage Techniques- 1985*, U.S. Internal Revenue Service, Publication 1299 (2-86).

Kovar, J.G., Whitridge, P., et MacMillan, J. (1988). Generalized edit and imputation system for economic surveys at Statistics Canada, *American Statistical Association, Proceedings of the Section of Survey Research Methods*, 627-630.

Little, R.J.A., et Rubin, D.B., (1987). *Statistical Analysis with Missing Data*, New York: John Wiley.

Newcombe, H.B. (1988). *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*, Oxford: Oxford University Press.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., et James, A.P. (1959). Automatic linkage of vital records, *Science*, 130, 954-959.

Newcombe, H., Fair, M., et Lalonde, P., (1992). The use of names for linking personal records, *Journal of the American Statistical Association*, 87 1193-1208.

Neter, J., Maynes, E.S., et Ramanathan, R. (1965). The effect of mismatching on the measurement of response errors, *Journal of the American Statistical Association*, 60, 1005-1027.

Nordbotten, S. (1963). Automatic editing of individual observations, presented at the Conference of European Statisticians, UN Statistical and Economic Commission of Europe.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Scheuren, F., et Winkler, W.E. (1993). Analyse de régression de fichiers de données couplés par ordinateur, *Techniques d'Enquête*, 19, 45-65.

Tepping, B. (1968). A model for optimum linkage of records, *Journal of the American Statistical Association*, 63, 1321-1332.

Winkler, W.E. (1995). Matching and record linkage, in B.G. Cox *et al.* (ed.) *Business Survey Methods*, New York: J. Wiley, 355-384.

Winkler, W.E., et Draper, L. (1996). Application of the SPEER edit system, in *Statistical Data Editing, Volume 2*, Statistical Commission and Economic Commission for Europe, Geneva, Switzerland, to appear.

Winkler, W.E., et Petkunas, T. (1996). The DISCRETE edit system, in *Data Editing*, Volume 2, Statistical Commission and Economic Commission for Europe, Geneva, Switzerland, to appear.

Winkler, W., et Thibaudeau, Y. (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. Census, *Statistical Research Division Technical Report*, U.S. Bureau of the Census.

SESSION 2

Méthodes analytiques

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

STATISTIQUES SOCIO-ÉCONOMIQUES ET POLITIQUE PUBLIQUE: NOUVEAU RÔLE POUR LES MODÈLES DE MICROSIMULATION

M.C. Wolfson[1]

RÉSUMÉ

Les utilisateurs de statistiques socio-économiques veulent habituellement plus de renseignements, et de meilleure qualité. Souvent, on peut répondre à ces besoins simplement par des collectes de Z données plus poussées, qui sont soumises aux contraintes habituelles relatives aux coûts et au fardeau de réponse des répondants. Les utilisateurs, particulièrement aux fins de politiques publiques, continuent de réclamer, et cette demande n'est toujours pas comblée, un système intégré et cohérent de statistiques socio-économiques. Dans ce cas, des données supplémentaires ne seront pas suffisantes; la contrainte la plus importante demeure l'absence d'une approche conceptuelle convenue.

Nous examinons brièvement ici l'état des cadres d'utilisation des statistiques sociales et économiques, y compris les genres d'indicateurs socio-économiques que les utilisateurs pourraient désirer. Ces indicateurs sont premièrement justifiés, en termes généraux, par des principes de base et des concepts intuitifs, ce laisse de côté les détails de leur construction, pour le moment. Ensuite, nous montrons comment une structure cohérente de tels indicateurs peut être assemblée.

Une conséquence fondamentale est que cette structure exige un réseau coordonné d'enquêtes et de processus de collecte de données, ainsi que des normes supérieures de qualité de données. Ceci, à son tour, implique une décomposition des systèmes sur mesure qui caractérisent la majeure partie du travail d'enquête des agences statistiques nationales (ex. des «chaînes de production» de données parallèles, mais généralement non reliées). De plus, les données émanant du réseau d'enquêtes doivent être intégrées. Puisque les données en question sont dynamiques, la méthode proposée dépasse la correspondance statistique et s'étend aux modèles de microsimulation. Enfin, ces idées sont illustrées avec les résultats préliminaires du modèle LifePathss actuellement en cours d'élaboration à Statistique Canada.

MOTS CLÉS: Microsimulation; statistiques sociales; cadres statistiques.

1. INTRODUCTION

Il est tout à fait raisonnable de s'attendre que les services de statistique nationaux brossent un tableau fiable, cohérent et pertinent des processus socio-économiques (lire Garonna, 1994; OCDE, 1976). Dans une large mesure, ils y parviennent grâce au Système de comptabilité nationale (SCN). On sait néanmoins que maintes faiblesses, qui plus est sérieuses, affligent le SCN, en particulier sur le plan des préoccupations sociales et des politiques. Ces faiblesses expliquent le désir de concevoir des séries ou des ensembles d'indicateurs sociaux susceptibles de recevoir la sanction internationale.

---

[1] Michael C. Wolfson, Statistique Canada et Institut des recherches avancées, 24 étage, Immeuble R.H. Coats, Ottawa,(Ontario), K1A 0T6, Canada, internet:wolfson@statcan.ca.
Le présent document propose le travail d'une équipe principalement constituée des membres du Groupe de la modélisation socio-économique de la Direction des études analytiques de Statistique Canada. Geoff Rowe et John Armstrong ont mené l'analyse empirique tandis que Steve Gribble a adapté le logiciel ModGen qui sert d'environment au modèle de microsimulation LifePathss et a assuré en permenence la direction de l'équipe. Les erreurs ou les incohérences qu'on pourrait relever dans le présent document doivent m'être entièrement imputées. Document rédigé pour les 50e assises de l'Institut international de statistique, Beijing, 1995. Communication sollicitée sur l'évolution des attentes des utilisateurs concernant la qualité des statistiques dans les secteurs public et privé.
L'auteur assume seul la responsabilité des opinions dans le présent document qui ne représente pas nécessairement le point de vue de Statistique Canada.

Jusqu'à présent, les efforts déployés pour créer des indicateurs sociaux d'une ampleur et d'une cohérence similaires à celles du SCN, et acceptés partout dans le monde comme celui-ci, se sont soldés par un échec. Pour cette raison, sans doute devra-t-on envisager des approches fondamentalement différentes si on veut satisfaire la demande de statistiques socio-économiques, notamment répondre aux besoins qui motivent depuis longtemps les travaux poursuivis dans le domaine des indicateurs sociaux.

Au sens large, on a envisagé trois grandes stratégies pour donner un cadre statistique à la sphère sociale. Une solution consisterait à prolonger le SCN, principalement sous forme de matrices de comptabilité sociale (MCS; lire Pyatt, 1990) ou de comptes satellites (lire Vanoli, 1994; Pommier, 1981). La seconde stratégie proposée prévoit la construction d'un canevas adapté à la statistique sociale - le plus connu et le mieux développé étant le système de statistiques sociodémographiques de Stone (SSSD; ONU, 1975; Stone, 1973). La troisième approche abandonne la structure et la cohérence d'un cadre explicite et les remplace par un ensemble spécial d'indicateurs statistiques faisant l'assentiment de tous. Une parfaite illustration de cette approche est le jeu d'indicateurs sociaux recommandé par l' OCDE (Moser, 1973; OCDE, 1976).

Aucune de ces trois stratégies n'a été mise en oeuvre à suffisamment grande échelle dans les pays industrialisés pour qu'on dispose de la base nécessaire à la production de données comparables à l'échelon international. Trois raisons expliquent cet échec. La première à trait à la faisabilité; la seconde résulte du manque d'intérêt de la part des gouvernements ou des organismes qui procurent les services de statistique; enfin, la troisième se rapporte au peu d'attention qu'engendre le sujet. (Manifestement, ces raisons partagent des liens entre elles.) L'expérience sur les indicateurs sociaux (OCDE, 1982) a néanmoins permis de dresser une liste opérationnelle et détaillée d'indicateurs - fruit d'un consensus entre experts et hauts fonctionnaires du gouvernement des pays membres. Malheureusement, les mécanismes de collecte des données qui faciliteraient la comparaison au niveau international n'ont pas encore été implantés pour bon nombre des indicateurs sur lesquels on s'est entendu (p.ex. vie saine, emploi du temps, revenu). Pourtant, on ne peut parler d'impossibilité technique puisque certains pays recourent déjà à des systèmes de données à des fins analogues.

Reste donc comme principale explication de cet échec un amalgame d'insignifiance et de priorités mal définies - entre autres la réticence d'investir les ressources voulues dans la collecte de statistiques. Le peu d'intérêt que soulèvent les indicateurs sociaux se situe plus au niveau de la comparabilité internationale qu'à celui de l'utilité locale. En effet, la plupart des pays utilisent déjà des statistiques sociales très variées; la difficulté est qu'il est généralement impossible de les comparer entre nations. Le manque d'intérêt pour un ensemble de statistiques sociales comparable à l'échelon international pourrait tout simplement venir de l'importance nettement plus grande accordée aux préoccupations de nature économique (comme en témoignent les efforts, couronnés de succès, visant à créer un SCN d'envergure internationale), comparativement aux préoccupations d'ordre social, du moins quand on les compare. Une autre raison pourrait être que quelques indicateurs sociaux (p.ex. espérance de vie, taux de chômage) permettent déjà une comparaison entre pays, et qu'on les juge suffisants.

Une des raisons soulignant l'utilité de données comparables sur le plan international est que les pays partagent des liens étroits dans le secteur en question. En ce qui concerne l'économie, les flux financiers internationaux rapprochent manifestement beaucoup les pays où on retrouve de surcroît des courants de pensée analogues sur la théorie macro-économique. Pareilles affinités sont à l'origine du SCN. Les liens pourtant tangibles au niveau social paraissent peut-être plus lâches (bien qu'on puisse douter de cette explication, étant donné les importants flux culturels et intellectuels qu'engendrent les médias de masse et les tendances similaires, observées partout dans le monde, relatives au terrorisme, au chômage, à l'éclatement de la famille, à la dénatalité, à l'inégalité grandissante des salaires, etc.). De plus, faute d'une théorie commune, la comparabilité internationale des données repose sur une base plus fragile -remarque aussi valable pour les statistiques individuelles (à savoir, répartition du revenu entre les ménages) que pour la manière dont on regroupe diverses statistiques sociales (quand on le fait). Bref, l'hésitation des pays à investir dans la constitution des systèmes de collecte des données essentiels aux indicateurs sociaux de l'OCDE pourrait bien venir d'un manque d'intérêt pour la comparabilité d'une série particulière d'indicateurs.

Quoi qu'il en soit, l'incapacité d'implanter le SSSD ou les comptes satellites dans le domaine social doit avoir d'autres raisons qu'un simple manque d'intérêt pour des données comparables à l'échelon international; le peu d'intérêt que soulève le cadre théorique sous-entendu par ces dernières doit aussi faire partie du problème. Les causes du mal sont ici plus profondes, car il est difficile de trouver des structures générales

cohérentes dont l'envergure et la mise en oeuvre se rapprochent même de loin du SCN pour les statistiques sociales, y compris à l'intérieur d'un pays.

L'échec des stratégies précitées décennie après décennie donne à penser que l'élaboration d'une structure cohérente pour les statistiques sociales doit être envisagée sous un nouvel angle. Des statistiques plus cohérentes contribueraient à répondre aux besoins d'une population d'utilisateurs toujours plus nombreux. D'une part, elles établiraient la base essentielle à la satisfaction des exigences du généraliste (par exemple, un aperçu sommaire des grandes tendances) et, d'autre part, elles permettraient une organisation des statistiques sociales complexes susceptible de venir en aide aux utilisateurs spécialisés, chez qui l'existence de plusieurs estimations hétérogènes du même élément, selon la provenance des données, pourrait semer la confusion (on comprend pourquoi).

2. PRINCIPES ÉLÉMENTAIRES

Avant de chercher de nouvelles approches à la construction d'un modèle pour les statistiques sociales, il convient de se fixer une série d'objectifs quantitatifs fondamentaux. En voici trois auxquels devrait se rallier, espérons-le, la majorité.

a. **Issue générale** -- Un des principaux objectifs consiste à déterminer si la situation empire ou s'améliore. Les gens sont-ils mieux lotis que l'année ou la décennie antérieures ? Répondre à une telle question s'avère difficile, principalement parce qu'aucune approche sommaire permettant de jauger le bien-être des individus ne fait l'assentiment général. Le revenu, l'état de santé, le niveau de scolarité et la privation sociale sont autant d'éléments qui entrent dans une telle mesure. Cependant, on ne s'entend pas sur les autres facteurs dont il faudrait tenir compte, ni sur la façon de les combiner pour obtenir un indice général. En outre, le concept même de l'issue recherchée ne fait pas l'objet d'un consensus dans certains domaines comme la santé et l'éducation.

Une entente partielle de ce genre a d'importantes répercussions sur l'élaboration d'un modèle statistique. La première est qu'on a besoin d'une marge de manoeuvre. Si certains paramètres permettant de mesurer le bien-être global engendrent un consensus, il est essentiel de les inclure au programme statistique sous-jacent. Cependant, puisqu'il n'existe pas une seule et unique «bonne manière» de les combiner, les utilisateurs devraient disposer d'une certaine latitude à cet égard. On devrait pouvoir forger différents indices sommaires à partir des éléments fondamentaux - tant pour des aspects comme la santé et l'éducation, que sur un plan plus général.

Une deuxième implication est que les statistiques sociales ne devraient être ni asservies aux statistiques économiques, ni s'en dissocier, comme l'ont fait jusqu'à présent les trois grandes approches stratégiques au problème. En effet, la situation économique se trouve manifestement au coeur même de toute mesure générale servant à déterminer si la situation des gens s'améliore ou non. Par conséquent, on envisagera de préférence un cadre pour des statistiques socio-économiques plutôt que des statistiques purement sociales. Ainsi, contrairement à ce qu'en pensent certains spécialistes des comptes nationaux (lire Vanolli, 1994), bâtir un modèle pour les statistiques sociales à partir des principes qui régissent le SCN laisserait à désirer. On doit plutôt songer à des approches innovatrices qui tireraient parti de nouvelles prémisses et engloberaient «l'économie sociale» dans son ensemble. De cette façon, le SCN deviendrait un important élément d'un plus vaste «système de statistiques socio-économique» (Wolfson, 1994; Ruggles et Ruggles, 1973).

Une troisième implication s'ajoute aux deux précédentes, soit que les méthodes sommaires qui permettent de comparer deux économies sociales ou davantage dans le temps ou entre divers pays ne doivent pas se résumer à une agrégation linéaire. Ainsi l'intérêt est un concept qui se prête parfois mal à l'agrégation à partir d'un seul numéraire, comme le fait le SCN avec les valeurs monétaires. Heureusement, les recherches sur certains sujets comme les méthodes servant à comparer la distribution du revenu entre les ménages, les distributions qui recourent de façon plus générale à des techniques graphiques (lire Easton et McCulloch, 1990) et les bases de données à architecture souple autorisant l'application des mathématiques à théorie figée, reposant sur des pointeurs, aux ensembles complexes de microdonnées longitudinales multivariées - désormais aisément utilisables grâce aux progrès de l'informatique - démontrent qu'une agrégation analogue à celle du SCN n'est pas essentielle. De fait, de telles approches se complètent et appuient le deuxième objectif que voici.

b. **Diversité** -- Un autre objectif fondamental de mesure consiste à aider les utilisateurs à percevoir la diversité de l'économie sociale. Le terme «diversité» englobe l'hétérogénéité sous de nombreuses formes -- par exemple, ne pas se borner aux agrégats et aux moyennes pour apaiser une critique qui revient constamment au sujet du SCN, soit que ce dernier ne dit

rien des nantis, des démunis et de la répartition inégale du revenu. La diversité se reflète aussi dans la dispersion du niveau de scolarité et de la structure des ménages au sein de la population d'un pays.

Saisir cette diversité a des implications fondamentales pour la statistique. Essentiellement, un tel exercice requiert des bases de microdonnées explicites. Modèle statistique prédominant, le SCN a été conçu avant la révolution de l'ordinateur. En ce sens, il nuit à l'exercice de réflexion créateur au sujet d'un cadre utile pour les statistiques sociales et socio-économiques. À l'ère pré-informatique où la structure de base du SCN a été conçue, l'agrégation n'était pas seulement un fondement théorique, c'était une nécessité. Aujourd'hui, grâce aux nouvelles techniques d'exploitation des bases de données, l'agrégation n'entrave pas que l'expression précise de la diversité, elle devient inutile sur le plan pratique.

(L'idée de bases de microdonnées explicites dans le cadre de projets statistiques d'une telle envergure ne date pas d'hier : lire Organisation des Nations Unies, 1979; Ruggles, 1981. Pourtant, le principe de «l'agrégation» a tout envahi, comme on a pu le constater avec les efforts déployés par l'OCDE pour mettre au point des indicateurs sociaux. Cette organisation a jugé bon de définir une série «de désagrégations fondamentales des principaux indicateurs sociaux» a priori ; OCDE 1977. Le fait qu'on se soit entendu à cet égard n'est pas dépourvu d'utilité, mais un tel exercice n'est pas essentiel quand les analystes disposent déjà de bases de microdonnées appropriées et comparables à l'échelon international.)

c. **Et si ?** -- Le troisième objectif de mesure fondamental consiste à établir la base qui permettra de poser des questions du genre «et si ?» et d'y répondre correctement. Il existe deux raisons élémentaires à cela. La plus évidente est que les services gouvernementaux qui forgent les politiques et les décideurs du secteur privé, principaux utilisateurs des statistiques socio-économiques, s'efforcent justement de trouver une réponse à de telles questions, par exemple comment le revenu disponible se répartirait-il si on apportait telle ou telle modification aux politiques fiscales/de transfert ? Ou encore, combien dépenserait-on pour tel ou tel produit dans cinq ans si les tendances actuelles se maintenaient ?

Une raison moins apparente mais tout aussi importante est qu'en réalité, les indicateurs statistiques constituent la réponse à ces questions. L'espérance de vie en est la meilleure illustration. L'espérance de vie en 1990, par exemple, répond à la question hypothétique suivante: «combien de temps vivrait une cohorte de naissances si tous ses membres étaient constamment exposés au taux de mortalité observé en 1990 ?» L'espérance de vie n'est pas une donnée que l'on peut observer directement comme le nombre de décès selon l'âge et le sexe. Il s'agit plutôt du résultat d'une simulation numérique étroitement associée au nombre de décès et de personnes à risque pris respectivement comme numérateur et dénominateur après désagrégation.

Si l'espérance de vie est un élément artificiel impossible à mesurer directement, il s'agit aussi d'un concept intuitif, facile à saisir et pouvant servir de canevas à des séries d'indicateurs apparentées. On le constate le plus clairement dans le secteur de la santé où un groupe spécial de chercheurs du REVES (Réseau pour l'espérance de vie en santé; Mathers et Robine, 1993) s'est constitué dans le but de mettre au point une telle série d'indicateurs et de susciter un consensus à leur sujet.

Lorsqu'on rassemble tous les éléments du débat, trois objectifs de quantification fondamentaux se dégagent:

- indicateurs généraux permettant de déterminer dans quelle mesure la situation de la population s'améliore;
- capacité d'illustrer la diversité et l'hétérogénéité de la population;
- instruments permettant de poser des questions du genre «et si ?» et d'y répondre.

Pour atteindre ces objectifs, le modèle statistique doit satisfaire aux exigences suivantes:

- il doit être souple;
- il doit englober les aspects social et économique;
- il doit reposer sur des bases de microdonnées explicites;
- il doit recourir aux techniques contemporaines d'informatique et d'exploitation des bases de données;
- il doit intégrer des modèles de simulation étroitement associés aux données.

À partir de telles prémisses, quels pourraient être les principaux éléments d'un modèle de statistiques socio-économiques? On peut formuler les remarques suivantes:

- à un moment quelconque dans le temps, la meilleure façon de représenter la population consiste à prélever un échantillon d'individus, chacun caractérisé par un jeu d'attributs et de relations données;

- les attributs comprennent le revenu, le niveau de scolarité, la consommation, divers paramètres de l'état de santé et les tendances relatives au temps consacré à diverses activés;
- Les relations se rapportent aussi bien aux liens de parenté classiques qu'à la cohabitation (à savoir, dans les bases de données ou sur le plan graphique-théorique, on peut représenter les relations de ce genre au moyen de pointeurs différents désignant d'autres personnes, qui font aussi partie de la base de données);
- le terme «relation» couvre également les liens avec les grandes institutions sociales, soit l'école, le travail et les programmes gouvernementaux. Les prises de contact, les relations ou les transactions entre individus et grandes institutions peuvent faire partie du jeu d'attributs personnels. Il pourrait s'agir de pointeurs se rapportant à la description des institutions - école, lieu de travail et programmes gouvernementaux - avec lesquelles le sujet a affaire;
- la base de données peut ensuite aisément être perçue comme une hiérarchie composée d'unités de type variable, à savoir individu, famille nucléaire, famille étendue et ménage;
- chaque unité (sujet, famille ou ménage) peut être décrite par un ou plusieurs attributs sommaires, par exemple le revenu disponible, les heures de loisir ou le degré de satisfaction personnelle;
- cela fait, il est possible d'évaluer la diversité de la population par application de statistiques sommaires à la distribution mixte multivariée des unités (p.ex., coefficient de Gini, quantiles);
- dans le temps, la meilleure façon de représenter la population consiste à utiliser une série de biographies, donc l'équivalent d'une vaste et longue enquête longitudinale à échantillon constant;
- grâce à une telle représentation longitudinale, la généralisation du concept de l'espérance de vie permettra de bâtir un ensemble cohérent d'indicateurs sommaires - notamment par division de l'espérance de vie en périodes cumulatives passées dans divers stades de l'existence.

Un tel modèle socio-économique intégrerait essentiellement un échantillon complet de micro-données longitudinales, véritable microcosme de la population réelle et de ses liens avec les principales institutions sociale et économiques. On pourrait aisément concevoir un vaste assortiment d'indicateurs statistiques à partir d'un tel microcosme, en réalité fait sans beaucoup plus d'efforts qu'enfoncer la touche <enter> présente sur tous les claviers d'ordinateur pour lancer l'algorithme approprié et faire analyser les données du microcosme par le logiciel.

De par leur construction, de tels indicateurs sommaires seraient cohérents, car ils dériveraient d'une base de microdonnées identique. Ils ne cacheraient pas la diversité et l'hétérogénéité de la population, la base de microdonnées étant toujours accessible pour une analyse approfondie (au déclic de la souris, par exemple, si on pense à l'aspect fonctionnel de l'informatique contemporaine).

La principale question que l'on peut se poser est la suivante: d'où viendrait ce microcosme ? Pour des raisons très pratiques (coût, fardeau pour les répondants et préoccupations relatives au respect de la vie privée), il ne pourrait venir d'une enquête longitudinale générale sur les ménages. En outre, on ne disposerait pas de cinquante ans ou davantage pour compléter une telle enquête, car au bout de ce laps de temps, de nombreuses choses auront changé de facon radicale. La conclusion inévitable est que le microcosme en question doit être artificiel.

La synthèse du microcosme prolongerait celle de la cohorte que sous-entendent déjà certains indicateurs comme l'espérance de vie. La méthodologie utilisée différerait néanmoins, car l'approche reposant sur une agrégation partielle ou la constitution de cellules, inhérente à la table de survie sous-jacente, est incompatible avec les bases explicites de microdonnées. On devrait plutôt recourir à une microsimulation.

De fait, on propose un hybride des idées de Stone sur le SSSD (ONU, 1975) et des bases de microdonnées intégrées explicites qu'a envisagées un groupe d'experts international subséquent (ONU, 1979). La première étape consiste à admettre que le SSSD repose implicitement sur des microdonnées longitudinales. Stone (1973) donne l'explication suivante :

> «Bien sûr, si l'on recueille les statistiques au moyen d'un ensemble de registres compatibles reliés entre eux ou, mieux encore, au moyen d'un jeu de données générales et individuelles, continuellement mises à jour (à savoir microdonnées longitudinales), toute la question d'un ordre séquentiel (c.-à-d. représentation des données en fonction de périodes de temps finies, chaînes du premier ordre de Markov) n'a plus grande importance puisque l'information peut être combinés de toutes les façons voulues dans une vaste base de données informatisée. Néanmoins, s'il est possible que de telles méthodes de collecte des données statistiques voient le jour dans l'avenir, on n'y recourt pas

encore pour l'instant, à de très rares exceptions près, si bien qu'il est sensé d'aborder la systématisation des statistiques sociales à partir de méthodes de collecte qui nous sont plus familières.» [TRADUCTION] (p.152, c'est nous qui écrivons en italique)

Vu sous cet angle, nous voici dans l'avenir. Inutile donc désormais de se plier aux contraintes de l'algèbre matricielle de Stone, des hypothèses restrictives du premier ordre de Markov et des «méthodes de collecte des données qui nous sont familières».

La deuxième étape ne fait que pousser plus loin l'idée de la création d'une base de données intégrée (BDI) synthétique par les méthodes de rapprochement statistique, énoncée si clairement il y a près de vingt ans par le groupe de travail de l'ONU sur les BDI (ONU, 1979). Les membres de ce groupe avaient reconnu la grande utilité des microdonnées à très grand nombre de variables et les limites pratiques à la collecte directe de telles données. Ils avaient donc recommandé que les microdonnées souhaitées soient synthétiques, même s'il fallait pour cela recourir à des enregistrements artificiels. Dans ces premiers travaux sur les BDI, on parlait généralement de microdonnées transversales.

Les microdonnées longitudinales synthétiques forment le trait d'union entre ces deux grands courants de pensée -- un modèle similaire au SSSD de Stone articulé sur la dynamique de l'analyse longitudinale et le rapprochement statistique artificiel de microdonnées. La différence est que la création de microdonnées longitudinales synthétiques exige plus que des techniques de rapprochement statistique, qui se prêtent mal à la combinaison de séries de microdonnées longitudinales disjointes. Les microdonnées longitudinales synthétiques doivent plutôt être générées par un modèle recourant à la microsimulation dynamique (idée qui, elle non plus, ne date pas d'hier; lire Ruggles, 1981). En un mot, ce ne sont pas les particularités des observations individuelles que l'on «apparie» dans les séries de données longitudinales, mais les tendances qui ressortent du comportement dynamique des séries d'observations dans chaque ensemble de microdonnées (ainsi qu'on le constatera plus loin).

Par ailleurs, la synthèse du microcosme par la microsimulation signifie qu'il ne faudra pas débourser grand-chose pour répondre aux questions du genre «et si? ». Ainsi, une fois qu'on aura bâti une table de survie, calculer la réduction de l'espérance de vie attribuable à une cause précise ne nécessitera que très peu de travail supplémentaire. La construction du microcosme par microsimulation crée une situation analogue. Dès qu'on aura investi dans la genèse artificielle du microcosme «de base», la synthèse de «variantes» de ce microcosme s'avère relativement simple.

Enfin, comme on se rendra compte dans la description qui suit, cette approche microanalytique du cycle de vie signifie qu'on n'a plus besoin de choisir entre une comptabilité sociale chronologique ou démographique, ainsi qu'il en était question dans Juster et Land (1981). L'approche élaborée dans le présent document couvre ces deux aspects.

3. IMPLICATIONS À L'ÉGARD DE LA COLLECTE DES DONNÉES

Le fait d'envisager un modèle de statistique socio-économique dans les grandes lignes décrites précédemment a d'importantes conséquences sur les aspects conceptuels et opérationnels des mécanismes de collecte des données pour les organismes nationaux chargés d'une telle mission. Ces conséquences pourraient néanmoins ne pas être très onéreuses (comparativement à ce que coûte la collecte des données primaires) et, dans la plupart des cas, sont relativement bénignes:

- les procédés de collecte des données ne peuvent être fait «sur mesure», ni être isolés les uns des autres;
- l'existence de définitions et de principes communs (à savoir, définitions identiques du niveau de scolarité et des méthodes servant à l'établir) constitue une forme de coordination entre les méthodes de collecte des données;
- un autre type de coordination consiste à veiller à ce que les données se chevauchent de la manière appropriée, soit à prévoir la nécessité d'un rapprochement statistique artificiel (ou de méthodes équivalentes);
- une microanalyse des données brutes s'avère beaucoup plus exigeante qu'une agrégation au niveau de la qualité des données.

En fait, tout cela signifie que les systèmes de collecte des données doivent être planifiés conjointement et que les normes applicables à la qualité des microdonnées doivent être plus sévères.

Une planification conjointe n'est pas une nouvelle exigence. L'élaboration du SCN a elle aussi exigé une certaine coordination au niveau de l'acquisition des données, ne serait-ce que pour s'assurer qu'on couvrait tous les secteurs de l'économie d'une manière quelconque. Cette forme de coordination coûte toutefois

considérablement moins cher que celle nécessitée par la microsimulation, parce qu'il est possible d'éliminer les incohérences des systèmes de collecte des données révélées par le SCN à un palier «macroscopique». On modifie de vastes agrégats, sans se soucier si les changements suscitent d'autres incohérences entre divers agrégats du SCN et les microdonnées originales. Pour la microsimulation, par contre, la cohérence interne entre les séries de données originales au niveau microscopique est capitale.

La nécessité de données de qualité à l'échelon microscopique n'est pas neuve. On se heurte surtout à cette difficulté chaque fois qu'il faut créer un ensemble de microdonnées pour l'usage public. Sachant que les utilisateurs examineront et analyseront les données à la loupe (pour étudier les valeurs «aberrantes» de la régression, par exemple), ces dernières font l'objet de modifications et d'imputations extensives. Les fichiers de microdonnées sur le recensement de la population soulèvent des difficultés analogues quant à la qualité des données au niveau microscopique, bien que ces difficultés soient moindres, car même si le public ne peut les consulter, ces fichiers peuvent faire l'objet de demandes générales spéciales pour des tableaux de corrélation.

La qualité des microdonnées deviendra encore plus préoccupante avec un modèle de micro-analyse intégrée comme celui que nous allons décrire. S'assurer que chaque élément d'un ensemble de microdonnées est plausible et cohérent à l'interne par un processus de correction et d'imputation est une chose, veiller à ce que la totalité de multiples jeux de microdonnées soient cohérents est bien différent (par exemple s'assurer qu'une enquête sur la santé et une autre sur les incapacités donnent les mêmes distributions d'incapacités en fonction de leur gravité selon l'âge et le sexe ou qu'une enquête longitudinale sur la dynamique du travail procure une estimation transversale de la participation de la population active qui concorde avec celle obtenue dans le cadre de l'enquête principale sur la population active, voire qu'une série chronologique de données administratives sur le nombre d'inscriptions dans les écoles coïncide avec les données du recensement sur le niveau de scolarité, selon l'âge et le sexe).

Pareille exigence de cohérence réciproque met en relief un problème soulevé par Wilk (1987), à savoir la faiblesse relative des méthodes statistiques pour résoudre les erreurs qui ne résultent pas de l'échantillonnage. Ainsi, le refus de répondre ou l'existence d'un biais quelconque lors des enquêtes auprès des ménages entraîne habituellement une sévère sous-déclaration de certaines sources de revenu. En général, les méthodes de correction et d'imputation classiques ne résolvent ce problème qu'en partie (c.à-d. on ne change pas les sources du revenu faussement signalées comme inexistantes). Les chercheurs qui élaborent les modèles de microsimulation pour les politiques fiscales/de transfert sont les seuls à avoir résolu ce problème, par nécessité (Citro et Hanushek, 1991; Bordt et coll., 1990; Wolfson et coll.., 1989). En outre, la correction des données recueillies auprès des ménages ne présente à toutes fins pratiques aucune utilité pour l'arrondissement des réponses (à savoir, mentionner le revenu à la centaine ou au millier du dollars le plus près), même si on possède la preuve qu'un tel comportement de la part des répondants fausse autant certaines statistiques (à savoir les quantiles) que l'erreur classique d'échantillonnage (Rowe et Gribble, 1994).

Enfin, on admet de plus en plus l'importance des enquêtes longitudinales, de toute évidence essentielles à une description de la dynamique et au déchiffrage des enchaînements de causalité. Pour utiliser les microdonnées longitudinales à ces fins, on devra recourir à des méthodes d'inférence plus complexes que celles dont se servent couramment les organismes qui s'occupent de statistiques, notamment la régression des risques plutôt que les tableaux de corrélations. Il pourrait s'ensuivre un examen encore plus critique des données.

4. LE PROJET LIFEPATHS

Nous allons maintenant passer à une illustration des points généraux qui précèdent. Le projet LifePaths est un projet ayant pour but la construction d'un modèle expérimental de statistique socio-économique. Ce projet se poursuit sous la direction de Statistique Canada au nom du ministère du Développement des ressources humaines du Canada, le nouveau «superministère» responsable, entre autres, du bien-être social, des pensions, de l'assurance-chômage et des politiques sur le marché du travail.

L'objectif fondamental du modèle LifePaths est de produire une série de vues multiples mais cohérentes de la situation socio-économique des Canadiens. Le modèle a été conçu en fonction des caractéristiques générales que nous venons de voir, soit la capacité de dégager la tendance générale, de refléter la diversité et de répondre aux questions du genre « et si ? ». Plus concrètement, le projet porte sur le temps que les Canadiens consacrent à diverses activités comme le travail, l'éducation, la

famille, les programmes gouvernementaux et les loisirs.

La généralisation des tables de survie pour la population active est l'une des vues ou l'un des aspects principaux envisagés. Le tableau 1, par exemple, présente non seulement l'espérance de vie classique des cohortes de Canadiens de sexe masculin nés à différentes années, mais aussi l'âge moyen auquel ces sujets entreront dans la population active rémunérée et en sortiront. En examinant une série de cohortes de naissances (période), chacune représentant des décennies qui se succèdent, l'analyse illustre très clairement certaines tendances à long terme, soit consacrer plus de temps aux études, prendre sa retraite plus tôt et travailler généralement moins longtemps, dans le cas des hommes. La dernière colonne révèle les effets de ces tendances sur le coût des régimes de pension publics. (Notons que malgré leur ancienneté relative, ces données brossent apparemment le tableau estimatif le plus récent de la vie active).

Le projet LifePaths étend les résultats de ce tableau fondamental sur la vie active dans plusieurs directions. Il approfondit l'analyse des tendances annuelles relatives au travail sans se limiter à une simple division entre années de vie active et inactive. Ainsi, il tient compte du travail à temps partiel, de la prolongation des congés payés et des vacances, de la modification du nombre typique d'heures de travail par semaine, des périodes de chômage ou de sortie de la population active de moins d'un an, des périodes où les personnes travaillent tout en poursuivant leurs études et de l'intérêt croissant pour le travail autonome. De plus, les aspects temporels du travail sont combinés à ses aspects économiques, notamment au revenu.

D'autres grandes catégories d'activités entrent aussi en ligne de compte. L'une d'entre elles est la poursuite des études; une autre, le milieu familial (à savoir, le fait de vivre seul ou avec d'autres membres de la famille). On peut dire que le modèle couvre les interactions avec les grandes institutions sociales, soit le travail, l'école et la famille. Le modèle LifePaths réunit donc les séquences «actives» (études et travail rémunéré) et «passive» (suite de groupements familiaux auxquels un individu adhère durant le cours de sa vie, p. 145), comme le dit Stone (1973) dans sa proposition de comptabilité démographique pour le SSSD, combinaison difficile à réaliser en pratique avec les méthodes matricielles qu'utilise cet auteur.

Le modèle planifie la participation à quelques grands programmes sociaux, entre autres les prestations d'assistance sociale (AS), d'assurance-chômage (AC) et d'indemnisation pour les accidents du travail (AT). En règle générale, on tient aussi mieux compte de l'emploi du temps, grâce aux données des enquêtes pertinentes - le régime de comptabilité du temps proposé, notamment par Juster et ses collaborateurs (1981). Les principales catégories d'activité n'intègrent donc pas seulement le travail et les études, mais aussi les tâches domestiques non rémunérées, l'hygiène personnelle, les soins dispensés à autrui, le sommeil, les déplacements urbains, la télévision, les autres loisirs passifs, les loisirs actifs, l'interaction avec les membres de la famille et d'autres formes de socialisation.

Tableau 1 - Espérance de vie et espérance de vie active des hommes de 15 ans (cohortes historiques constantes)

Âge	entrée dans la population active	âge moyen à la retraite	âge au décès	années de travail	années de retraite	années de travail par année de retraite
1921	16,5	63,7	67,6	47,2	3,9	12,1
1931	17,0	64,0	68,4	47,0	4,4	10,7
1941	17,2	64,1	69,1	46,9	5,0	9,4
1951	17,5	63,9	70,4	46,4	6,5	7,1
1961	18,2	64,0	71,2	45,8	7,2	6,4
1971	19,8	63,3	71,3	43,5	8,0	5,4

Source: Gnanasekaran et Montigny (1975) et Wolfson (1979)

Le modèle LifePaths englobe toutes ces activités humaines, en adoptant un point de vue qui épouse le cycle de vie dans son entièreté, d'une manière cohérente et intégrée - si bien qu'on combine et enchasse les approches de comptabilité chronologique et démographique qu'analysent Juster et Land (1981). L'élaboration du modèle LifePaths est un exercice de haute voltige et les résultats présentés ici doivent être considérés comme expérimentaux.

Côté méthodologie, le modèle LifePaths innove sur plusieurs plans.

En premier lieu, aucun ensemble de données particulier ne renferme toute l'information requise, par exemple des données détaillées sur les activités humaines sous les angles économique et social. Les ensembles existants et ceux qui pourraient s'y ajouter ne sont que partiels et fragmentaires. En outre, tel qu'indiqué précédemment, pour des raisons d'ordre pratique (coût, fardeau pour les répondants et protection de la vie privée), on ne pourra jamais recourir aux données entièrement intégrées des enquêtes sur les ménages. On devra inévitablement appliquer des processus d'intégration synthétique à des ensembles de données multiples.

Deuxièmement, le modèle doit couvrir le cycle de vie complet des sujets. Le faire avec les données longitudinales dont on dispose actuellement exigerait des décennies d'enquêtes subséquentes au terme desquelles bon nombre de choses auront changé. L'idée fondamentale consiste donc à généraliser le concept de l'espérance de vie pour la période et la table de survie sous-jacente. Par conséquent, l'analyse concentrera sur des cohortes réalistes mais hypothétiques.

Un troisième élément de l'objectif primaire a trait à l'expression détaillée de la diversité ou de l'hétérogénéité de la population, donc la capacité d'observer un phénomène distributif comme la répartition inégale du revenu. Une telle capacité exige comme base des microdonnées explicites. Puisqu'on ne peut recueillir de données sur la vie réelle d'un échantillon représentatif d'individus, les microdonnées nécessaires doivent être artificielles. Elles doivent cependant être assez réalistes pour qu'on ne puisse pas vraiment les distinguer des ensembles partiels de caractéristiques provenant des données réelles sur des échantillons de la population, notamment des enquêtes longitudinales.

Toutes ces contraintes signifient que le modèle LifePaths doit avoir pour âme un modèle de microsimulation, en d'autres termes, un aperçu réaliste mais artificiel de la vie de l'individu.

5. LES DONNÉES SYNTHÉTIQUES

Avant de vous présenter nos résultats initiaux, il est important d'expliquer dans quelle mesure le modèle LifePaths repose sur des données synthétiques et comment les résultats synthétiques reflètent raisonnablement la réalité.

Un être humain vit généralement 75 ans, environ. Toutefois, face à la rapidité relative avec laquelle évoluent maintes activités humaines, procéder à des observations socio-économiques cohérentes et soutenues pendant un tel laps de temps est à toutes fins pratiques irréalisable. Les statistiques reconnnues depuis des décennies n'existaient tout simplement pas il y a 75 ans (qu'on songe au taux de chômage, au PIB par habitant et aux mesures concernant les loisirs). De même, il se peut fort bien que dans 75 ans d'ici, en 2070, les statistiques fondamentales, dont l'importance est acquise de nos jours, soient remplacées par d'autres valeurs que l'on peut à peine imaginer aujourd'hui.

Pourtant, les indicateurs statistiques qui reflètent les processus couvrant la vie humaine dans son entièreté suscitent beaucoup d'intérêt. Le plus connu est sans doute l'espérance de vie. Néanmoins, il en existe d'autres comme la proportion de mariages qui devraient se terminer par un divorce, le nombre d'emplois différents qu'une personne pourrait connaître durant sa carrière professionnelle, la pertinence ou non des pensions publiques face aux revenus antérieurs à la retraite et la partie de la vie moyenne qu'une personne passe en santé ou malade. Il existe manifestement des indicateurs qui s'appliquent à la vie humaine, et ceux-ci sont plus ou moins largement acceptés. Le modèle LifePaths les généralise.

Même si on ne le l'admet pas de manière générale, l'espérance de vie est une statistique «artificielle». En un sens, on peut la comparer à une déclaration que l'on ferait sur la destination d'une automobile dont on connaît la position et la vitesse, mais pas l'accélération. L'espérance de vie repose sur le taux de mortalité spécifique selon l'âge (et le sexe). Comme c'est le cas pour la vitesse du véhicule, elle s'appuie donc sur des données réelles. Cependant, l'espérance de vie (période) s'applique à un individu hypothétique qu'on retire du passage du temps et qui passera le reste de sa vie exposé au taux de mortalité en vigueur au début des années 1990. Bref, on ne tient pas compte de l'accélération positive ou négative du taux de mortalité.

On sait, bien sûr, que les taux de mortalité ont généralement diminué au cours des dernières décennies et on s'attend largement à ce que cette tendance se poursuive. Par conséquent, bien qu'en soi elle néglige la

tendance suivie par les taux de mortalité, l'espérance de vie, par les tendances qui lui sont propres, nous renseigne fort commodément sur les changements des taux de mortalité sous-jacents, car elle suit dans le temps une forme de moyenne pondérée du taux de mortalité spécifique selon l'âge (et le sexe). Évidemment, on pourrait toujours examiner le taux de mortalité spécifique selon l'âge sous-jacent, mais essayer de saisir l'évolution ne serait-ce que d'une centaine de valeurs est une tâche passablement complexe. (Or, ces valeurs deviennent beaucoup plus nombreuse dès qu'on répartit les taux de mortalité selon le sexe, l'état civil et l'âge.) L'espérance de vie garde néanmoins son utilité comme indicateur, précisément parce qu'elle comprime des centaines de données en un indicateur intuitivement accessible, indicateur dont la variation dans le temps épouse raisonnablement celle des taux de mortalité spécifiques selon l'âge sous-jacents.

Le modèle LifePaths est conçu pour être entièrement analogue. Néanmoins, il repose sur une multitude de processus et de descriptions statistiques des états marquant la transition de l'individu entre différents stades de la vie. Par exemple, outre la mortalité, le modèle tient explicitement compte des paramètres démographiques comme l'état civil et les transitions connexes que sont le passage du célibat au concubinage ou au mariage, voire la rupture du couple par la séparation ou le divorce. Pareillement, on a tenu compte d'autres classifications de la situation socio-économique, notamment le fait de travailler ou de poursuivre des études, en fonction des données réelles sur les taux récents de distribution et de transition pertinents.

Pour réussir une telle généralisation de l'espérance de vie, on a dû généraliser le concept sous-jacent de la table de survie. Le degré de précision le plus élevé de la table de survie est le groupe de sujets - par exemple défini par le sexe et l'âge. On présume que les membres du groupe sont homogènes. Pareil degré de précision ne suffit pas pour le modèle LifePaths. En effet, on doit explicitement tenir compte des sujets hétérogènes que caractérisent de nombreux attributs si l'on veut effectuer le meilleur usage du modèle et illustrer aussi précisément que possible les résultats de l'analyse des schémas de comportements dynamiques observés avec les vastes ensembles de microdonnées longitudinales.

En un sens, non négligeable, tout cela signifie que le modèle LifePaths donne des résultats beaucoup plus réalistes que l'espérance de vie obtenue avec la table de survie classique. Ainsi, le modèle ventile les taux de mortalité d'après l'état civil, plus l'âge et le sexe. À son tour, l'état civil repose sur un ensemble complexe de facteurs comme le niveau de scolarité, les antécédents de fécondité et la durée des séjours au sein de la population active.

D'un autre côté, étant donné l'aspect artificiel des «données», le modèle produira inévitablement des résultats plus explicites qu'une table de survie. En effet, alors que la table de survie classique s'appuie sur un groupe de sujets, ces derniers demeurent implicites en soi - on se borne à compter le nombre d'individus de la cellule ou de la catégorie, soit selon l'âge et le sexe. Le modèle LifePaths, quant à lui, tient explicitement compte de la vie de chaque élément.

Dans ce cas, quel sens devrait-on accorder à un résultat du modèle, par exemple la ventilation de l'espérance de vie entre le nombre d'années qu'une personne peut s'attendre à consacrer au travail et aux études ? Un tel résultat devrait donner une interprétation analogue à celle de l'espérance de vie traditionnelle - c'est-à-dire constituer une sorte de sommaire du taux démographique récent. Les résultats du modèle LifePaths illustrent ce qui se produirait si les taux de transition les plus récents entre différents états socio-économiques (dépendant des attributs des sujets hétérogènes) demeuraient constants.

6. RÉSULTATS INITAUX

Le modèle LifePaths se compose essentiellement d'un échantillon du cycle de vie complet (synthétique) de plusieurs sujets. Cette base de microdonnées longitudinale articulée sur un échantillon de vies est malheureusement beaucoup trop complexe pour permettre une analyse directe. C'est pourquoi nous ne présenterons ici que quelques «vues» sommaires du microcosme sous-jacent, vues obtenues dans la plus pure tradition d'une analyse démographique.

Soulignons que ces «vues» se limitent à des indicateurs scalaires comme le PIB; elles dévoilent simultanément plusieurs paramètres démographiques fondamentaux. Néanmoins, on ne doit pas voir là une lacune, comme l'impossibilité de passer à une mesure générale unique avec le SCN. Chaque «vue» doit plutôt être perçue comme une illustration de la puissance des logiciels d'infographie contemporains, qui permettent une évaluation plus poussée de l'économie sociale qu'un simple indice.

Débutons par une des représentations démographiques les plus simples, la pyramide des âges. La figure 1 montre la pyramide des âges d'une table de survie de base. Le nombre de sujets de sexe féminin apparaît à droite, sur l'axe horizontal, et le nombre de sujets de sexe masculin, à gauche, jusqu'à l'âge de 100

ans, indiqué sur l'axe vertical. Un tel diagramme repose sur les probabilités de transition pour la période (fin des annéees 1980 et début des années 1990), sur lesquelles on reviendra. Ainsi qu'on peut s'y attendre, la courbe de survie des femmes diminue plus lentement que celle des hommes quand l'âge augmente, illustrant l'espérance de vie supérieure des femmes (ou plus exactement la raison à l'origine de ce phénomène). (La coche à 99 ans est attribuable à l'intervalle, qui correspond à ≥ 99 ans.)

La figure 1 répartit la même population en trois catégories socio-économiques - soit les «travailleurs», les personnes aux «études» et les «autres». La vie d'un «étudiant» débute avec la première année, de telle sorte que les enfants en garderie et à la maternelle font partie du groupe «autres». Puisque le modèle LifePaths suit chaque sujet tout au long de sa vie, on a dû prendre certaines décisions arbitraires lorsqu'une personne poursuit plusieurs activités la même année. Plus exactement, pour qu'une personne fasse partie des «travailleurs», elle doit travailler au moins 15 heures par semaine et consacrer la majeure partie de l'année au travail, au même rythme. Une personne qui passerait donc 5 mois à étudier, 4 à travailler au moins 15 heures par semaines et les trois derniers mois à travailler hebdomadairement moins de 15 heures (ou pas du tout) se retrouverait dans le groupe «aux études» cette année-là; si on substitue les périodes de 5 et de 4 mois cependant, elle compterait parmi les personnes «au travail». (L'utilisateur du modèle a tout pouvoir sur ces définitions.) Le diagramme révèle qu'à de rares exemptions près, chacun poursuit des études à l'âge de 8 ans et que quelques sujets commencent à quitter le milieu scolaire à 16 ans; que la majorité des gens ont terminé leurs études à 20 ans, mais qu'une poignée d'hommes et de femmes les poursuivent au-delà de la vingtaine.

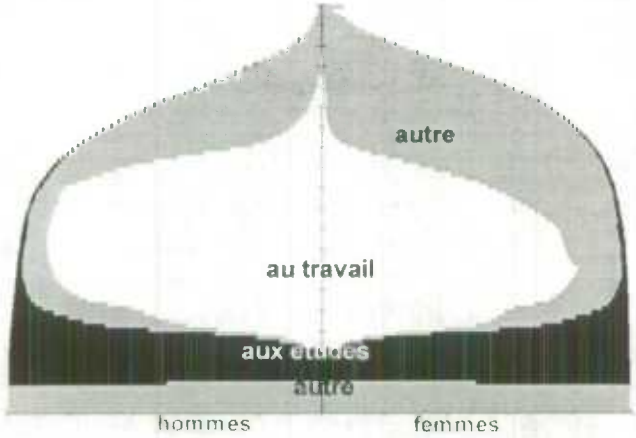


Figure 1 - Population du modèle LifePaths (années-personnes) selon le type d'activité, l'âge et le sexe.

Personne ne semble passer directement des études au travail, mais nous en reparlerons en examinant un autre diagramme. On notera peut-être une proportion surprenante de sujets dans le groupe «autres», qui comprend les personnes sans emploi et celles qui ne font pas partie de la population active (à savoir, ménagères, pensionnés). Comme c'était prévisible, toutes proportions gardées, la probabilité que les hommes travaillent à un âge quelconque est plus élevé que pour les femmes. La courbe des femmes de 20 à 25 ans détenant un emploi s'affaisse malgré la tendance relative à une plus grande participation à la population active, car cette période représente les principales années de procréation. Ensuite, la courbe augmente légèrement entre 25 et 35 ans. Dans le cas des hommes, le taux de participation diminue considérablement pour le groupe des 60 à 65 ans.

La figure 1 correspond à la «séquence active» dont parle Stone (c'est-à-dire le passage des études au travail), alors que la figure 2 donne un aperçu de la «séquence passive». La figure 2 reprend la pyramide des âges de la figure 1 et se rapporte exactement à la même population synthétique sous-jacente du modèle LifePaths, mais ventile les individus en fonction d'une autre dimension, soit l'état civil. Par définition, tous les sujets de moins de 18 ans sont des «enfants» à moins qu'ils soient mariés ou aient eux-mêmes un enfant. Lorsqu'un couple se sépare, on suppose que les enfants restent avec leur mère. Cette hypothèse explique pourquoi il y a des parents seuls de sexe féminin mais pas de sexe masculin. (Les versions ultérieures intégreront des données plus réalistes sur les ententes de garde.)

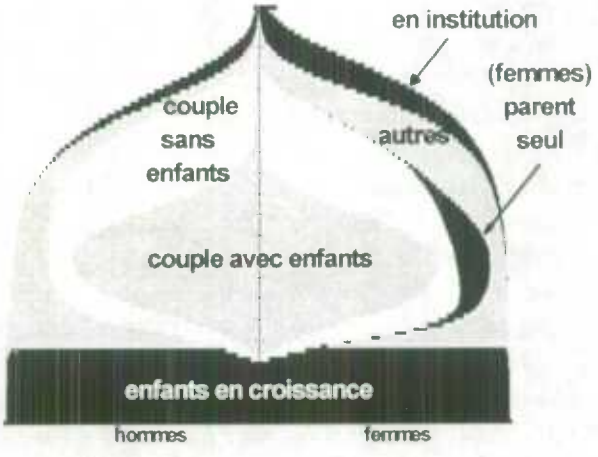


Figure 2 - Population du modèle LifePaths (années-personnes) selon l'état civil, l'âge et le sexe

Quand on compare la courbe des gens mariés (couples avec ou sans enfants) des deux sexes, on constate que celle des hommes est décalée de quelques années vers le haut. Ce résultat reflète la tendance générale selon laquelle le mari a quelques années de plus que son épouse. Le diagramme révèle aussi qu'il existe plus de veuves que de veufs. On peut y voir la conséquence de la différence d'âge moyenne positive entre mari et femme, et de la plus grande espérance de vie des personnes de sexe féminin. Enfin, le diagramme montre que beaucoup de femmes vivent en institution (principalement des foyers de soins infirmiers ou des établissements de traitement des maladies chroniques), encore une fois en raison de leur plus grande longévité et de la plus forte prévalence des problèmes de santé à un âge avancé, doublées au fait que les hommes souffrant d'une incapacité similaire ont souvent une épouse en mesure de prendre soin d'eux à la maison.

Les figures 1 et 2 ne donnent qu'un aperçu des possibilités du modèle LifePaths, des «vues» (tableaux de corrélation dans le cas présent) du microcosme complet sous-jacent, soit un ensemble de microdonnées longitudinales pour une cohorte de naissances artificielle du «début des années 1990». On peut classer le même ensemble de microdonnées longitudinales sous-jacent de façon à obtenir la figure 3 indiquant le passage entre divers états, plutôt que la population de chaque état. Dans ce cas, le graphique de la figure 3 reproduit les passages d'un état à l'autre pour les sujets de la figure 1. L'axe horizontal correspond au nombre de personnes qui passent par telle ou telle transition chaque année, une fois de plus sous forme de pyramide des âges. L'axe vertical commun donne l'âge, les femmes se retrouvant à la droite, sur l'axe horizontal, et les hommes, à la gauche. (Dix-huit pour cent de la population se retrouvent aux extrémités de l'axe horizontal, si bien qu'une cohorte de 100,000 sujets autorise jusqu'à 9 000 transitions aussi bien pour les hommes que pour les femmes, par année).

La première transition se fait du groupe «autre» (petite enfance ou pré-maternelle) à celui «aux études». La figure 1 révèle que tous les enfants de sexes masculin et féminin effectuent ce passage à l'âge de 6 ou de 7 ans. La grande transition suivante survient au terme des «études», le passage au groupe des «travailleurs» atteignant son point culminant vers l'âge de 20 ans, aussi bien pour les hommes que pour les femmes. Un plus petit nombre de sujets, qui atteint également un sommet vers 20 ans, accomplit la transition entre les «études» et une «autre» activité. On se rappelera que cette dernière catégorie correspond aux années-personnes qui n'ont consacré la majeure partie de leur année (soit légèrement plus du tiers) ni aux études, ni à un travail les occupant plus de 15 heures par semaine.

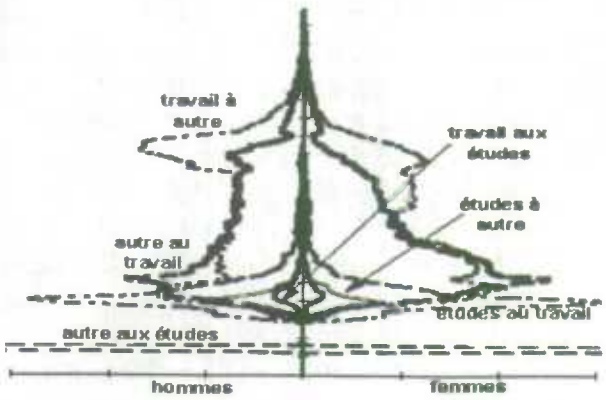


Figure 3 - Transitions de la population du modèle LifePaths (années-personnes), selon l'âge et le sexe .

Du début de l'âge adulte à la soixantaine, les principales transitions surviennent entre les stades «travail» et «autre». Précisons que les passages correspondent à une valeur brute et non à une valeur nette. On remarque d'ailleurs que le passage net entre les catégories «travail» et «autre» (établi par comparaison des transitions brutes) change de direction vers le groupe «autre» pour les sujets de sexe féminin de 40 à 45 ans, alors qu'il ne varie pratiquement pas pour les hommes jusqu'à l'âge de 50 ans. Enfin, un nombre maximal de personnes prennent leur retraite lorsqu'elles atteignent de 55 à 65 ans, le pic étant plus prononcé dans le cas des hommes.

Outre le nombre de personnes recensées dans les différentes catégories d'activité et le nombre de sujets passant d'un groupe à l'autre, le modèle LifePaths permet de produire des tableaux montrant la durée de séjour, c'est-à-dire le temps que les individus passent dans tel ou tel stade de vie. Le tableau 1, plus haut, tendait déjà dans cette direction puisqu'il donnait une première estimation de l'espérance de vie active. Une autre capacité du modèle, d'une grande importance compte tenu des microdonnées qui en forment implicitement la base, concerne la production de distributions à une ou deux variables du séjour pour tel ou tel groupe - par exemple la distribution combinée du nombre d'années passées aux études et au travail pour les hommes et les femmes. (Faute d'espace nous ne pourrons présenter ces diagrammes.)

La figure 4 présente une autre image obtenue grâce à la simulation de base du modèle - il s'agit cette fois d'une classification des activités selon un axe horizontal différent. Au lieu d'indiquer le nombre d'années-

personnes tirées de la table de survie d'une cohorte de naissances pour une période donnée, comme les figures 1 et 2, la figure 4 prend comme axe horizontal les activités principales, soit le nombre d'heures consacrées à chacune d'elles pendant une semaine normale (168 heures), selon le sexe et l'âge. Ainsi, on remarque que les hommes passent en moyenne près de 40 heures par semaine au travail entre l'âge de 30 à 55 ans.

En surface, la figure 4 imite exactement les données que l'on pourrait obtenir directement d'une enquête sur l'emploi du temps. Sur le plan de la validation, les données des deux sources devraient correspondre étroitement. Pourtant, ce diagramme a été produit par simulation au moyen du modèle LifePaths et diffère quelque peu des données de l'enquête sur l'emploi du temps sous-jacente, principalement parce qu'on a accru la cohérence des données. Ainsi, le taux de participation annuel à la population active selon l'âge et le sexe de la figure 4 est cohérent avec les taux sous-entendus à la figure 1, de même qu'avec les tendances démographiques de la figure 2, en raison de la façon dont le modèle est construit.

Une des impressions qui se dégage du diagramme est qu'en moyenne, les hommes et les femmes passent une partie relativement faible de leur vie à travailler contre rénumération - domaine d'élection du SCN. Quand on l'examine sous l'angle du nombre moyen d'heures qu'on y consacre chaque semaine (plutôt qu'en fonction du fait qu'on travaille plus de 15 heures par semaine pendant plus du tiers de l'année, comme à la figure 1), on se rend compte que le travail ne monopolise qu'une très petite fraction de la vie (la vie éveillée, s'entend). Bien sûr, le travail non rémunéré, de même que les aspects de l'hygiène personnelle et des loisirs, présentent aussi une grande importance pour l'économie, mais le SCN n'en tient pas compte au-delà de la valeur monétaire agrégative du taux de consommation personnel par produit.

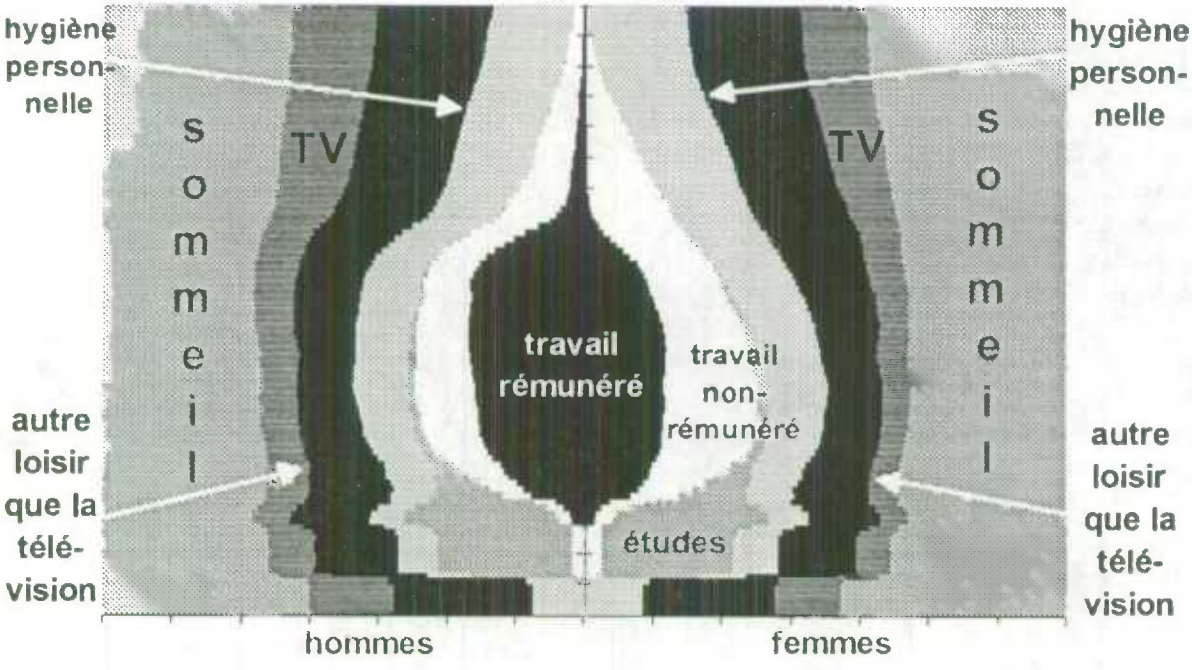


Figure 4 - Emploi du temps (nombre d'heures moyen par semaine) selon le modèle LifePaths, par grande activité selon l'âge et le sexe.

La même figure révèle les limites des ratios de dépendance démographiques classiques - qui reposent sur l'utilisation du nombre brut de personne d'âge productif (à savoir de 20 à 64 ans) comme dénominateur. L'examen de la figure 4 révèle que ces ratios sous-estiment manifestement la dépendance économique de nombreuses personnes face à la société. Le même diagramme donne à penser qu'il faut reproduire de façon plus explicite les mécanismes qui permettent au reste de la population d'acquérir son pouvoir d'achat, surtout par le temps consacré au «travail productif». Ces mécanismes comprennent les transferts intra-familiaux ainsi que les programmes fiscaux et de transfert des gouvernements. D'une manière plus générale, en combinant l'emploi du temps aux paramètres démographiques plus classiques, le modèle LifePaths permet de construire une série cohérente de tableaux statistiques qui reflètent beaucoup mieux l'activité sociale et économique.

Les diagrammes LifePaths comme celui de la figure 4 montrent clairement que la vie ne se résume pas à ce que le SCN permet de capturer avec son orientation économique. Il s'ensuit qu'une publication régulière des résultats statistiques de ce genre pourrait avoir une incidence appréciable sur l'élaboration des politiques publiques, car on replacerait les paramètres économiques dans un cadre plus général et attirerait l'attention sur les effets beaucoup plus étendus des politiques en matière de chômage, retraite, redistribution du revenu, éducation, garde des enfants, désinstitutionalisation et semaine de travail, pour n'en citer que quelques-unes.

Soulignons encore une fois que les résultats du modèle LifePaths ont toujours essentiellement une valeur d'illustration. La base de microdonnées longitudinales synthétique sous-jacente exige encore des améliorations. Comme on le verra dans la partie qui suit, ces données reposent sur une série d'enquêtes et d'analyses récentes, donc sur des données réelles, mais les analyses concernées ne portent toujours en partie que sur des résultats préliminaires.

7. MÉTHODES SOUS-JACENTES

Le modèle LifePaths que l'on vient d'illustrer exploite surtout deux séries de données récentes et les résultats de près d'une décennie de recherches sur les modèles de microsimulation connexes. Les deux séries de données précitées sont celles de l'Enquête sociale générale (ESG) de 1992, en vertu de laquelle on a posé aux répondants des questions détaillées sur leur emploi du temps au cours des 24 dernières heures, et l'Enquête sur l'activité (EA), qui a servi à recueillir des données longitudinales détaillées sur la dynamique du marché du travail entre 1988 et 1990. Le modèle de microsimulation LifePaths est un hybride de ces deux enquêtes, du modèle de microsimulation DEMOGEN (Wolfson, 1989) adapté à l'environnement du tout nouveau logiciel de microsimualtion ModGen C++, et du nouveau modèle de calcul du remboursement en fonction du revenu applicable aux prêts pour l'éducation post-secondaire mis au point par le ministère du Développement des ressources humaines du gouvernement canadien.

La présente partie survole très rapidement les processus qui ont débouché sur la synthèse de la cohorte de naissances du modèle LifePaths, âme même du modèle. En règle générale, cette synthèse nécessite, d'une part, une architecture générale raccordée à divers mécanismes économiques et socio-démographiques et, d'autre part, une analyse détaillée des données en mesure de produire une description statistique de chaque processus de manière empirique (par exemple, la dynamique du comportement).

Comme la table de survie classique, le modèle LifePaths débute avec une population spécifique, par exemple 100 000 naissances. Contrairement à elle cependant, il suit chaque sujet toute sa vie jusqu'à la mort. (Une table de survie suit un groupe d'individus que l'on suppose homogènes.) À divers moments dans le temps, chaque sujet a la possibilité d'effectuer un changement dans sa vie (transition). Compte tenu de la série de paramètres existants, il pourrait s'agir du passage au marché du travail ou d'un changement d'état civil. Le nombre de transitions possibles dépend de la gamme d'états explicitement envisagés. Dans la version actuelle du modèle, les membres de la cohorte sont caractérisés par les attributs de base qui suivent, à chaque moment de leur vie:

- âge -- en tant que variable continue
- fécondité -- âge à la naissance des enfants, présence d'enfants au foyer familial
- état civil -- célibat, concubinage ou mariage, séparation ou divorce
- situation relative à l'emploi -- y compris, participation à la population active et type d'emploi (nombre d'heures par semaine, nombre de semaines dans l'année)
- éducation -- année et type d'établissement si le sujet est encore aux études, niveau de scolarité
- revenu -- rémunération horaire, hebdomadaire et annuelle

- emploi du temps -- catégories de la figure 4 avec désagrégation plus poussée
- participation aux programmes -- notamment assistance sociale, assurance-chômage, régime de pension public
- paramètres du conjoint -- y compris âge, niveau de scolarité, expérience du marché du travail .

Ces attributs fondamentaux, permettent également d'en obtenir d'autres, secondaires, fort variés, comme les variables qui apparaissent aux figures 1 à 4.

Connaissant cette série d'attributs, nous pouvons maintenant décrire les processus qui ont servi à générer la trajectoire de chaque attribut dans le modèle. En voici une brève description.

Démographie - Dans le modèle, la fécondité est une conséquence de la conception, elle-même modélisée sous forme d'une suite de taux de probabilité finis et constants, subordonnés à l'âge, à l'état civil et au nombre antérieur de naissances vivantes. Les principales sources de données sont les enregistrements de naissances, plus l'Enquête sur la famille de 1983. De cette façon, on peut tenir compte du biais attribuable à la conception durant le célibat ou le concubinage, suivie par le mariage avant l'accouchement. Le taux de mortalité est associé à l'âge, au sexe et à l'état civil et s'appuie sur les avis de décès. Dans les deux cas, c'est le recensement de la population qui sert de dénominateur.

La formation et la dissolution du couple sont illustrées par une série de fonctions de probabilité. Partant du célibat , on note des probabilités concurrentes de passer au concubinage ou au mariage. La rupture du couple génère des probabilités de séparation et de divorce qu'on estime séparément pour les hommes et les femmes et qui dépendent des antécédents, d'une manière assez complexe. Par exemple, la «probabilité» qu'une femme s'engage dans une union est positivement corrélée à celle d'être enceinte et atteint sa valeur la plus élevée peu après l'entrée au sein de la population active. La probabilité d'une séparation est plus forte pour les femmes si le couple n'a pas d'enfants en bas âge à la maison, si la femme s'est mariée lorsqu'elle était adolescente et si elle a travaillé récemment.

Éducation - Les taux de transition illustrant le passage de l'école primaire au cours secondaire ont été bâtis de manière à être conjointement aussi cohérents que possible avec les taux de fréquentation scolaire des enfants d'âge pertinent, obtenus lors des recensements de 1986 et de 1991. Le passage aux études post-secondaires (collège, institut technique, université) repose sur les taux de probabilité estimatifs qui dérivent de l'Enquête nationale auprès des diplômés (END), des données administratives sur le nombre d'inscriptions dans les écoles et de l'Enquête sur l'activité (EA) lorsque les jeunes quittent leur emploi pour retourner aux études et poursuivre celles-ci.

Travail - L'expérience sur le marché du travail est simulée en deux grandes étapes: le fait d'avoir ou non un emploi et la rémunération provenant de l'emploi en question. Dans le premier cas, on estime les passages à la vie active et inactive grâce à l'EA pour les hommes et les femmes pris séparément, ainsi que séparément pour un premier emploi, un second, les emplois subséquents, et l'abandon du marché du travail. L'entrée initiale sur le marché du travail est représentée par une distribution du temps d'attente, alors que les autres transitions sont illustrées par des fonctions de probabilité à variables multiples. Le sexe et le niveau de scolarité sont d'importants facteurs en ce qui concerne la période d'attente qui précède l'obtention du premier emploi. La probabilité d'une réinsertion dans la population active dépend du sexe, du niveau de scolarité et de la durée de la période courante de non-emploi; dans le cas des femmes, elle dépend aussi de l'existence d'enfants en bas âge, ce paramètre ayant un effet à la baisse supplémentaire.

La rémunération repose sur la situation relative à l'emploi décrite précédemment et sur des modèles distincts pour le nombre d'heures de travail hebdomadaires et le salaire horaire. Après l'entrée initiale sur le marché du travail, on attribue au hasard un nombre d'heures de travail hebdomadaires à partir d'une distribution selon l'âge, le sexe et le niveau de scolarité. Cette distribution s'appuie sur les données combinées de l'END, de l'EA et de l'Enquête sur les finances des consommateurs (EFC - enquête annuelle sur la distribution du revenu dans les ménages). La variable «heures hebdomadaires» est ensuite corrigée d'après l'âge, le sexe, le nombre d'heures de travail hebdomadaires de l'année antérieure et le niveau de scolarité. Chaque sujet reçoit un rang-centile pour le salaire horaire avec le nombre d'heures de travail hebdomadaires. Le taux de rémunération horaire est subséquemment «repris» en fonction de distributions selon l'âge, le sexe et le niveau de scolarité. Le rang-centile est corrigé annuellement d'après le classement ordinal qui «dérive» de l'EA.

Emploi du temps - L'Enquête sociale générale (ESG) de 1992 a permis d'interroger environ 9000 personnes, uniformément réparties selon l'âge, le sexe, la journée de la semaine et le mois de l'année, sur leur emploi du temps pendant 24 heures. Elle a aussi servi à recueillir des données de base sur le niveau de scolarité,

la situation relative à l'emploi et l'état civil. Après analyse approfondie des données, on a créé un module LifePaths qui impute à chaque journée-personne simulée un vecteur représentant le temps consacré à chaque activité durant une période de 24 heures, y compris au niveau d'agrégation le plus élevé des catégories indiquées à la figure 4. (On a formulé des hypothèses spéciales pour les enfants de moins de 15 ans et les personnes agées résidant en institution car ils n'étaient pas touchés par l'ESG.)

L'analyse statistique révèle que l'âge, le sexe, le jour de la semaine, l'état civil, l'existence de jeunes enfants, le niveau de scolarité et l'activité principale (à savoir, études, travail rémunéré ou travail autonome, autre) sont tous associés significativement à ces tendances vectorielles. On s'est donc servi des attributs produits par d'autres processus du modèle LifePaths pour l'imputation. Le processus d'imputation a également été conçu pour imiter les tendances variables relatives à l'emploi du temps observées chez les sujets qui présentaient les mêmes attributs, essentiellement grâce à une distribution des résidus vectoriels d'une analyse de régression à variables multiples.

8. VALIDATION ET QUANTITÉ DES DONNÉES

Valider le modèle LifePaths est fondamentalement impossible, tout simplement parce que le modèle crée un échantillon à partir d'une cohorte de naissances hypothétique. Par conséquent, on ne pourra jamais en comparer les résultats avec la «réalité». Néanmoins, en raison de sa construction, le microcosme artificiel de vies devrait reproduire les principales distributions communes marginales qui lui servent de point de départ, par exemple le taux de participation à la population active, le taux de fécondité, le taux de mortalité, le taux d'union et de dissolution des couples, le taux d'inscriptions à l'école et la distribution du revenu provenant d'un travail, selon l'âge et le sexe.

On a constamment vérifié les comparaisons de ce genre durant la construction du prototype du modèle décrit ici. Dans une large mesure, il existe une bonne concordance. Les principales discordances surviennent lorsque les sources de données sous-jacentes manquent elles-mêmes de cohérence, signe qu'il existe des erreurs dans les données originales. En fait, le modèle LifePaths fournit un cadre en partie analogue à celui du SCN, pour les microdonnées socio-économiques, cadre qui rend les données de diverses sources cohérentes, donc met en relief les incohérences.

9. CONCLUSION

Nous avons débuté en soulignant les besoins des utilisateurs pour des renseignements plus complets et plus cohérents sur le plan socio-économique et proposé une explication à l'échec des efforts déployés antérieurement à l'échelon international pour satisfaire ces besoins. Une nouvelle approche a été suggérée, approche supposant un usage beaucoup plus important de base de microdonnées multivariées et de méthodes de microsimulation. Les particularités fondamentales de cette nouvelle approche, entre autres sa cohérence et sa complétude, ont été mises en relief grâce aux résultats préliminaires venant du modèle en cours d'élaboration à Statistique Canada.

L'espace ne nous permet pas d'illustrer les autres caractéristiques du modèle par des graphiques, notamment les microdonnées explicites qui en forment la base et permettent d'analyser la diversité. Des recherches plus poussées sont nécessaires pour dégager d'autres grandes caractéristiques comme les indicateurs sommaires (à savoir, distribution du revenu durant la vie) et les simulations du type «et si ?». Les résultats présentés dans ce document constituent néanmoins une importante «preuve par construction» de la faisabilité pratique et technique d'une telle approche.

La même approche fait apparaître des lacunes et des faiblesses au niveau des données de statistique socio-économique existantes, examinées sous l'angle de la micro-analyse. L'approche LifePaths susciterait des exigences beaucoup plus sévères à l'égard de la cohérence et de la qualité des enquêtes ainsi que des méthodes de collecte de données socio-économiques. Dans la mesure où l'on reconnaît les avantages d'une méthode semblable au modèle LifePaths pour l'analyse des statistiques socio-économiques, pareille méthode pourrait constituer la base d'un exercice quelconque de planification stratégique pour les organismes nationaux qui dispensent des services de statistique.

BIBLIOGRAPHIE

Bordt, M., Cameron G., Gribble S., Murphy B., Rowe G., et Wolfson M. (1990). The Social Policy Simulation Database and Model: An Integrated Tool for Tax/Transfer Policy Analysis, *Canadian Tax Journal*, 38, 48-65.

Citro, C.F., et Hanushek E.A. (1991). *Improving Information for Social Policy Decisions, The Uses of Microsimulation Modeling*, National Academy Press, Washington, D.C.

Easton, G.S., et McCulloch R.E. (1990). A Multivariate Generalization of Quantile-Quantile Plots, *Journal of the American Statistical Association*, juin, 88, 410, Theory and Methods, 376-386.

Garonna, P. (1994). Statistics facing the concerns of a changing society, *Statistical Journal of the United Nations ECE*, 11, 2, 147-156.

Gnanasekaran, K.S., et Montigny, G. (1975). *Tables de vie active des hommes au Canada et dans les provinces*, 1971, Statistique Canada, 71-524F au catalogue occasionel, Ottawa.

Juster, F.T., et Land K.C. (1981). Social Accounting Systems: An Overview in F.T. Juster et K.C.Land (sous la dir. de), *Social Accounting Systems -- Essays in the State of the Art*, Academic Press, New York.

Juster, F.T., Couran, P.N., et Dow, G.K. (1981). The theory and measurement of Well-Being: A Suggested Framework for Accounting and Analysis in Juster, F.T. et .Land, K.C. (sous la dir. de), *Social Accounting Systems -- Essays in the State of the Art*, Academic Press, New York.

Mathers, C., et Robine, J.-M. (1993). Health expectancy indicators: a review of the work of REVES to date, in J-M Robine, C.D.Mathers, M.B.Bone, I.Romieu (sous la dir. de), *Calculation of Health Expectancies: Harmonization, Consensus Achieved and Future Perspectives*, INSERM / John Libby Eurotext Ltd., 226.

Moser, Sir C. (1973). Social Indicators -- Systems, Methods and Problems, *Review of Income and Wealth*, Series 19, 2, juin, 133-141.

OCDE (1976). *Measuring Social Well-Being*, Paris.

OCDE (1977). Basic Disaggregations of Main Social Indicators, D.F.Johnston, *Special Studies No. 4, The OECD Social Indicator Development Programme*, Paris.

OCDE (1982). *The OECD List of Social Indicators*, Paris.

Pommier, P. (1981). Social Expenditure: Socialization Expenditure? The French Experience with Satellite Accounts, *Review of Income and Wealth*, décembre.

Pyatt (1990). Accounting for Time Use, *Review of Income and Wealth*, Series 36, 1, mars, 33-52.

Rowe, G., et Gribble, S. (1994). Income Statistics from Survey Data: Effects of Respondent Rounding, à venir dans Proceedings of the American Statistical Association, Section on Government Statistics.

Ruggles, N. et R. Ruggles (1973), A Proposal for a System of Economic and Social Accounts, in M. Moss (sous la dir. de), *The Measurement of Economic and Social Performance*, National Bureau of Economic Research, New York.

Ruggles, R. (1981). The Conceptual and Empirical Strengths and Limitations of Demographic and Time-Based Accounts, in F.T.Juster, et K.C. Land (sous la dir. de), *Social Accounting Systems -- Essays in the State of the Art*, Academic Press, New York.

Stone, R. (1973). A System of Social Matrices, *Review of Income and Wealth*, Series 19, 2, juin, 143-166.

Organisation des Nations Unies (1975). *Towards a System of Social and Demographic Statistics* (SSDS), Studies in Methods, Series F, 18, ST/ESA/STAT/SER F/18, New York.

Organisation des Nations Unies (1979). *The Development of Integrated Data Bases for Social, Economic, and Demographic Statistics* (IDBs), Studies in Methods, Series F, 27, ST/ESA/ STAT/SER F/27, New York.

Vanoli, A. (1994). Extension of National Accounts: opportunities provided by the implementation of the 1993 SNA, *Statistical Journal of the United Nations ECE*, 11, 3, 183-191.

Wilk, M.B. (1987). *The Concept of Error in Statistical and Scientific Work*, document présenté à la U.S. Bureau of the Census Third Annual Research Conference, Baltimore.

Wolfson, M.C. (1979). Épargner pour la retraite: mais combien ?, II, 18 dans Le système de retraite au Canada: problèmes et possibilités de réforme, Groupe d'étude sur la politique de revenu de retraite, ministère des Finances, Ottawa.

Wolfson, M.C. (1989). Divorce, Homemaker Pensions, and Lifecycle Analysis, *Population Research and Policy Review*, 8: 25-54.

Wolfson, M.C., Gribble, S., Bordt, M., Murphy, B., et Rowe, G. (1989). The Social Policy Simulation Database and Model: An Example of Survey and Administrative Data Integration, *Survey of Current Business*, 69, 36-40.

Wolfson, M.C. (1994). Implications of Evolutionary Economics for Measurement in the SNA, Towards a System of Social and Economic Statistics, document présenté à la vingt-troisième conférence générale de l'International Association for Research in Income and Wealth, St.Andrews, Nouveau-Brunswick, 21-27 août 1994, miméographié, Statistique Canada, Ottawa.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

ÉLABORATION, UTILISATION ET MODIFICATION DES FONCTIONS D'ÉVALUATION DES RISQUES POUR LA SANTÉ : L'ÉTUDE DE FRAMINGHAM

R.B. D'Agostino[1]

RÉSUMÉ

Les fonctions d'évaluation des risques pour la santé sont des fonctions ou des modèles mathématiques qui permettent d'établir un lien entre les variables des facteurs de risque et la probabilité qu'un événement tel qu'une maladie coronarienne se déclare. L'étude de Framingham a ouvert la voie à l'élaboration des fonctions d'évaluation des risques de maladies cardiovasculaires. Dans le présent article, nous examinons les étapes de la mise au point de ces fonctions ainsi que certaines de leurs utilisations. Nous présentons en outre certaines des modifications récentes qu'on leur a apportées pour répondre à certaines préoccupations de nature mathématique et pratique.

MOTS CLÉS : Modèles de prévision; facteurs de risque; profils des risques; études épidémiologiques.

1. INTRODUCTION

L'étude de Framingham est une vaste étude épidémiologique prospective et continue de cohortes commencée en 1948 et dont l'objectif principal est d'enquêter sur les rapports qui existent entre les maladies cardiovasculaires (MCV) et les facteurs de risque tels que l'âge, le sexe, la pression artérielle, le taux de cholestérol, le tabagisme, l'hématocrite, l'obésité et le diabète (D'Agostino et Kannel, 1990). Le vocable MCV englobe la maladie coronarienne (infarctus du myocarde, angor instable et angor stable), l'accident cérébrovasculaire, l'insuffisance cardiaque, la claudication intermittente et les cas de décès de cause cardiovasculaire ou cardiaque. L'étude porte sur 5 209 sujets âgés de 28 à 62 ans (2 336 hommes et 2 879 femmes). Les sujets sont soumis tous les deux ans à un examen médical et à une entrevue qui servent à mettre à jour les facteurs de risque de MCV et les informations portant sur l'état et l'évolution des maladies cardiovasculaires depuis la dernière visite. En outre, les sujets font l'objet d'une surveillance constante qui permet de noter et d'obtenir des informations sur les décès ainsi que sur l'évolution des MCV.

Au fil des années, l'étude de Framingham a permis de mettre au point des **modèles de prévision mathématiques** qui établissent des liens entre les facteurs de risque et la probabilité de survenue de MCV ou de certains types particuliers de ces maladies comme l'insuffisance coronarienne. On donne aujourd'hui à ces modèles ou à ces fonctions le nom de **fonctions d'évaluation des risques pour la santé**. Dans les pages qui suivent, nous présenterons un bref aperçu de l'élaboration et de l'utilisation de ces fonctions et nous décrirons certains des progrès récents qui s'y rapportent.

2. PÉRIODE INITIALE (1948 - 1976)

La première décennie de l'étude a été consacrée à l'accumulation de données et les articles publiés portaient principalement sur le plan d'enquête et sur les objectifs (Dawber, Meadors et Moore, 1951; Dawber, Kannel et Lyell, 1963). Les fonctions de prévision ont été mentionnées pour la première fois dans un article au milieu des années 1960 (Truett, Cornfield et Kannel, 1967).

Nous présentons au tableau 1 deux des fonctions originales. Il s'agit de fonctions discriminantes linéaires

[1] Ralph B. D'Agostino, professeur de mathématiques, de statistique et de santé publique, Boston University, 111 Cummington Street, Boston, MA 02215, É.-U.

de Fisher qui établissent un lien entre les principaux facteurs de risque de MCV et le fait q'une première maladie coronarienne (MC) se déclare. Les facteurs de risque sont: l'âge en années (AGE), le taux de cholestérol total (CHOL), la pression artérielle systolique (PAS), le poids relatif métropolitain (PRM), l'hématocrite (HEM), le tabagisme (CIG) et l'hypertrophie du ventricule gauche telle que mesurée à l'électrocardiogramme (HVG). La variable PRM est calculée en divisant le poids réel du sujet par un poids dit idéal relevé dans les tableaux des poids idéals de la compagnie d'assurance La Métropolitaine. Tous les facteurs de risque sont significatifs au niveau p = 0,05, à l'exception de ceux marqués d'un (NS).

Tableau 1

Fonctions discriminantes linéaires de Fisher reliant les facteurs de risque au fait qu'une première maladie coronarienne (MC) se déclare à l'intérieur d'une période de 12 ans.

Les sujets ne montraient aucun signe de MC au premier examen. Certains d'entre eux ont souffert de MC au cours des 12 années qui ont suivi ce premier examen (HOMMES : n = 2 187 dont 258 cas de MC; FEMMES : n = 2 669 dont 129 cas de MC)

	Hommes		Femmes	
	Coefficients			
Constante	-10,8986		-12,5933	
AGE	0,0708		0,0765	
CHOL	0,0105		0,0061	
PAS	0,0166		0,0221	
PRM	0,0138		0,0053	(NS)
HEM	-0,0837	(NS)	0,0355	(NS)
CIG	0,3610		0,0766	(NS)
HVG	1,0459		1,4338	

NS = Non significatif au niveau p = 0,05. Toutes les autres variables sont significatives au moins au niveau p = 0,05.

Ces fonctions peuvent servir à des fins de classification comme suit. Désignons par F la fonction discriminante linéaire de Fisher définie par

$$F = A + B_1 * X_1 + B_2 * X_2 + \ldots + B_K * X_K \qquad (1)$$

où $X_1, \ldots, X_K$ représentent les valeurs des facteurs de risque et $B_1, \ldots, B_K$ sont les coefficients dont les valeurs numériques sont fournies au tableau 1. Pour un sujet donné, on obtient les facteurs de risque du tableau 1 et on calcule la valeur F de l'équation (1). La règle de classification est :

Si F ≥ 0, le sujet est classé MC
Si F < 0, le sujet est classé non-MC

Les chercheurs ont en outre observé qu'il était possible d'estimer la probabilité qu'une maladie coronarienne se déclare au cours des 12 années suivant le premier examen en déterminant l'exponentielle de la fonction F de (1). Cette exponentielle permet notamment d'estimer la probabilité conditionnelle qu'une MC se déclare, compte tenu des données des facteurs de risque $X_1, \ldots, X_K$. Cette probabilité s'exprime symboliquement comme suit :

$$P(MC \mid X) = [1 + \exp(-F)]^{-1} \qquad (2)$$

On désigne souvent la fonction telle que présentée en (2) sous le nom de forme logistique de la fonction.

Les chercheurs de Framingham craignaient cependant que l'utilisation de la théorie de l'analyse discriminante de Fisher pour l'estimation des coefficients de régression de (1) et de (2) fournisse des estimations biaisées et inappropriées puisque cette méthode part de l'hypothèse que le vecteur des facteurs de risque X appartient à une distribution normale multivariée. Or, il est clair que cette hypothèse n'est pas respectée puisque des variables dichotomiques comme le tabagisme (CIG) et l'HVG sont incluses dans les modèles du tableau 1. On a donc décidé de s'écarter de l'analyse discriminante et d'estimer les coefficients B conditionnels aux valeurs observées de X. On a ainsi obtenu une **régression logistique**, et mis au point une méthode des moindres carrés pondérés (Walker et Duncan, 1976) aux fins de l'estimation.

Nous présentons au Tableau 2 les fonctions de prévision — que les chercheurs de Framingham ont appelé à l'époque fonctions du profil de risque — qui établissent un lien entre les facteurs de risque et le fait qu'une première MCV se déclare au cours d'une période de 8 ans (tiré de Kannel, McGee et Gordon, 1976).

Tableau 2

Fonctions de régression logistique reliant les facteurs de risque au fait qu'une première maladie cardiovasculaire (MCV) se déclare à l'intérieur d'une période de 8 ans (tous les facteurs de risque sont significatifs, $p < 0.05$).

	Hommes	Femmes
	Coefficients	
Constante	-19,7710	-16,4598
AGE	0,3743	0,2666
AGE^2	-0,0021	-0,0012
CHOL	0,0258	0,0161
PAS	0,0157	0,0144
CIG	0,5583	0,0395
HVG	1,0529	0,8745
GLUC	0,6020	0,6821
CH*AGE	-0,0004	-0,002

Dans les fonctions de profil (ou fonctions d'évaluation des risques pour la santé) du tableau 2, la variable « hématocrite » (HEM) du tableau 1 a été exclue. On a par contre ajouté le carré de l'âge (AGE^2) et l'interaction entre l'âge et le cholestérol total sérique (CH*AGE). L'ensemble des variables des facteurs de risque du tableau 2 est devenu la norme pour beaucoup des fonctions prédictives de Framingham. Les sujets abordés ci-dessus ont fait l'objet d'articles importants dont deux méritent d'être mentionnés : Halperin, Blackwelder et Verter (1971) et Gordon, Kannel et Halperin (1979).

3. UTILISATION DE MESURES MULTIPLES SUR UN SUJET - MÉTHODE DES MESURES RÉPÉTÉES GROUPÉES (1968-1989)

À mesure que l'étude de Framingham s'est poursuivie, les séries d'examens réalisés tous les deux ans ont donné lieu à l'accumulation d'une vaste quantité de données. Ces données ont permis de mettre à jour les facteurs de risque d'un sujet particulier et de les incorporer dans des régressions logistiques. La méthode statistique des mesures répétées groupées a été mise au point à cette fin (Cupples, D'Agostino, Anderson et Kannel, 1988). Il s'agit essentiellement d'une méthode d'examen personnel dont le déroulement est exposé au tableau 3 ci-après.

Tableau 3

Illustration de la méthode des observations répétées groupées

Des observations sur les facteurs de risque sont recueillies tous les deux ans.

MÉTHODES D'EXAMEN PERSONNEL

	Examen t	t+1	t+2	échantillon
non-IC	100	92	83	275 total
IC	5	8	6	19 IC
suivi impossible	3	1	2	

ANALYSE sur 275 sujets et 19 événements

Au temps t, il y avait 100 sujets exempts de MC. Entre les temps t et t+1, une MC s'est déclarée chez cinq de ces sujets et 3 autres sont devenus introuvables. Il restait donc 92 sujets pour l'examen au temps t+1. De ce nombre, 8 ont été atteints de MC et une personne ne répondait plus à l'appel entre les temps t+1 et t+2, ce qui a laissé 83 sujets pour l'examen au temps t+2. De ces 83 sujets, 6 ont par la suite été atteints d'IC et 2 n'ont pas répondu à l'appel. Le nombre d'examens personnels réalisés dans le cadre de cette étude a donc atteint 275 (100+92+83), et le nombre d'événements observés a été de 19 (5+8+6). Dans le contexte de l'étude de Framingham, deux années se seraient écoulées entre chaque série d'examens. Les valeurs 275 et 19 servent de données pour une régression logistique, par exemple, mettant en rapport les facteurs de risque et le fait qu'une MC se déclare au cours des deux années qui suivent l'examen. Notez que les facteurs de risque utilisés dans l'analyse seraient ceux obtenus lors de l'examen le plus récent.

La méthode des mesures répétées groupées est devenue une méthode standard de l'étude de Framingham (Shurtleff, 1974; Cupples et D'Agostino, 1987). D'Agostino et coll. (1990) ont démontré que lorsqu'on l'utilise avec la régression logistique, cette méthode, appelée en l'occurrence régression logistique groupée, est liée asymptotiquement (c'est-à-dire, pour les grands échantillons) à la régression des risques proportionnels de Cox avec covariables dépendantes dans le temps (D'Agostino et coll., 1990).

Robert Abbott et Daniel McGee (1987) ont utilisé cette méthode des mesures répétées groupées, mais en

choisissant cette fois des fenêtres de 8 ans. Ils ont produit des fonctions d'évaluation des risques pour la santé correspondant à un certain nombre d'événements de MCV, y compris l'infarctus du myocarde, la MC, les décès provoqués par la MC, la claudication intermittente, les accidents cérébrovasculaires et les MCV.

Ces fonctions sont devenues extrêmement populaires et les demandes d'élaboration de fonctions destinées à des fins particulières, présentées aux responsables de l'étude de Framingham, se sont multipliées. Par exemple, le *Carter Center* de la *Emory University* a demandé aux responsables de l'étude de mettre au point des fonctions pour le calcul de la mortalité tandis qu'un certain nombre de compagnies pharmaceutiques réclamaient des fonctions traitant explicitement de la pression artérielle ou du taux de cholestérol.

Pendant la même période, d'autres articles publiés dans le cadre de l'étude de Framingham ont clarifié encore davantage la notion des fonctions prédictives (Gordon et Kannel, 1982; Kannel et McGee, 1987).

4. UTILISATION DES ANALYSES DU TEMPS PRÉCÉDANT L'ÉVÉNEMENT (1987 - 1993)

4.1 Décision de produire des fonctions réglementaires

Vers le milieu des années 1980, on a commencé à craindre que le nombre de fonctions d'évaluation des risques pour la santé mises au point ne soit trop élevé, et que l'élaboration de ces fonctions ne s'appuie sur aucune méthode systématique et ne fasse l'objet d'aucun contrôle approprié. En outre, on possédait dorénavant des méthodes statistiques nouvelles, telles que la régression des risques proportionnels de Cox, qui semblaient convenir parfaitement à ces fonctions prédictives mais qui avaient pourtant jusque-là été ignorées. Ces méthodes étaient supérieures à la régression logistique du fait qu'elles pouvaient tenir compte du temps écoulé avant l'événement ainsi que de l'abandon de certains sujets et d'autres formes de censure. Les responsables de l'étude ont donc décidé de générer des fonctions « réglementaires » qui feraient usage des nouvelles méthodes.

4.2 Les fonctions professionnelles et l'*American Heart Association*

La décision susmentionnée a eu pour première conséquence de produire des fonctions destinées à l'usage professionnel (Wolf et coll., 1991; Anderson et coll., 1991) dont la première a été une fonction prédictive des accidents cérébrovasculaires (Wolf et coll., 1991). Cette méthode utilise le modèle des risques proportionnels de Cox et peut servir à l'estimation de la probabilité d'un premier accident cérébrovasculaire sur une période atteignant jusqu'à 14 ans en s'appuyant sur les facteurs de risque mesurés au point de départ (c'est-à-dire, au temps 0). Le modèle mathématique utilisé ici s'exprime par la formule

$$S(t) = 1 - S_0(t)^{\exp(F)} \qquad (3)$$

où $S(t)$ désigne la fonction de survie au temps t (c'est-à-dire la probabilité que le sujet survive au moins jusqu'au temps t) et

$$F = A + B_1 * X_1 + B_2 * X_2 + \ldots + B_K * X_K \qquad (4)$$

La fonction $S_0(t)$ s'appelle fonction de survie sous-jacente ou moyenne et correspond à la probabilité de survie au temps t d'un sujet dont les valeurs du risque au temps 0 correspondent aux valeurs moyennes calculées pour l'ensemble des facteurs de risque.

La deuxième fonction destinées à l'usage professionnel avait pour objet l'évaluation des risques de MC. Malheureusement, l'hypothèse de la proportionnalité de la régression de Cox ne se vérifiait pas dans ce cas et il a donc fallu employer un modèle de défaillance accélérée de Weibull incorporant une composante de non-proportionnalité. Le modèle mathématique de cette fonction prend la forme suivante :

$$S(t) = EXP[-EXP\{(\ln(t) - F)/J\}] \qquad (5)$$

où $S(t)$ est la fonction de survie au temps t, et F est une fonction linéaire comme celle donnée en (4)

et

$$\ln(J) = A + C * F \qquad (6)$$

La modélisation de J en tant que fonction des facteurs de risques X répond aux exigences de la non-proportionnalité. Cette fonction peut servir à l'estimation de la probabilité d'un premier cas d'IC au cours de la période de 4 à 14 ans suivant la mesure des facteurs de risque.

La fonction d'évaluation des risques de MC était une fonction unique utilisant 8 facteurs de risque (sexe, âge, pression artérielle systolique, cholestérol total, cholestérol des LPHD, tabagisme, présence de diabète et HVG) et les transformations de ces facteurs. On a

également élaboré des fonctions d'évaluation du risque d'accident cérébrovasculaire spécifiques au sexe, qui utilisaient les facteurs de risque susnommés, exception faite du cholestérol total et du cholestérol LPHD, en plus de la manifestation antérieure d'une MCV (sauf les accidents cérébrovasculaires), de fibrillation auriculaire et de l'utilisation de médicaments antihypertenseurs.

L'*American Heart Association* a diffusé les fonctions d'évaluation des risques d'insuffisance coronarienne et d'accident cérébrovasculaire aux médecins, afin que ces derniers fasse l'évaluation de leurs patients présentant des facteurs de risque de MCV. La présentation est conçue de manière qu'on puisse obtenir les estimations de la probabilité directement sur une période de dix ans lorsqu'on connaît les facteurs de risque. On peut alors les comparer au risque moyen d'un sujet du même âge et du même sexe. Ces fonctions peuvent également servir à estimater les effets d'une intervention puisqu'on peut les utiliser pour évaluer le changement de la probabilité découlant de la modification d'un facteur de risque. Par exemple, si une personne cesse de fumer ou si sa pression artérielle systolique est réduite de 20 mmHg.

4.3 Fonctions accessibles au public pour l'*American Heart Association*

L'*American Heart Association* a également demandé aux chercheurs de l'étude de Framingham de produire des fonctions d'évaluation des risques pour la santé plus facilement compréhensible pour des non-professionnels. De telles fonctions ont été élaborées pour les MCV et les accidents cérébrovasculaires. Celle qui est destinée aux MCV est connue sous le nom de **RISKO**.

5. MODÈLES DU SECOND ÉVÉNEMENT (1994 À AUJOURD'HUI)

Jusqu'ici, tous les modèles d'évaluation des risques pour la santé que nous avons décrits ont visé un premier événement, un premier cas de MCV, de MC, ou d'accident cérébrovasculaire. Les travaux actuels portent également sur l'élaboration de modèles concernant les seconds événements. Par second événement, nous entendons un événement qui frappe un sujet qui en a déjà subi un premier; par exemple, une personne qui souffre d'un second infarctus du myocarde. Les chercheurs de l'étude de Framingham ont porté leur attention sur les personnes qui avait survécu au stade aigu de l'événement initial. Nous présentons ci-après un ensemble de fonctions de régression des risques proportionnels de Cox spécifiques au sexe pour la prévision d'une IC ou d'un accident cérébrovasculaire pour les sujets chez lesquels on a déjà diagnostiqué une MCV :

	Hommes	Femmes
	Coefficients	
ln(AGE)	1,006029	2,383863
ln(C-T/C-LPHD)	0,957359	0,750673
CIG	0,0	0,764838
ln(PAS)	1,278776	1,157384
DIAB	0,229595	0,799036

La variable C-T/C-LPHD est le rapport du cholestérol total sur le cholestérol LPHD et la variable DIAB est la variable dichotomique de la présence ou de l'absence du diabète. Toutes les autres variables ont été définies antérieurement.

6. VALIDATION DES FONCTIONS DE FRAMINGHAM

Il existe une abondante documentation sur la validation des fonctions d'évaluation des risques pour la santé de Framingham. Ces documents sont trop nombreux pour être énumérés dans le présent article. Toutefois, il nous paraît utile de mentionner deux articles récents où l'on affirme que ces modèles peuvent être transposés validement dans d'autres contextes (Laurier et coll., 1994; Grover et coll., 1995).

7. RÉSUMÉ ET QUESTIONS ACTUELLEMENT À L'ÉTUDE

L'étude de Framingham a permis de produire un ensemble de fonctions d'évaluation des risques pour la santé d'une grande utilité. De nombreux travaux sont toujours en cours. Nous énumérons ci-après quelques-unes des questions qui font toujours l'objet de recherches.

1. Des fonctions d'évaluation des risques pour la santé liés aux maladies cardiovasculaires, à la maladie coronarienne et à la mortalité due à ces maladies, ainsi qu'aux accidents cérébrovasculaires pour des périodes de suivi atteignant de 2 à 14 ans ont été mises au point et sont actuellement utilisées ou prêtes à l'être.

2. De nouvelles méthodes statistiques perfectionnées ont été incorporées dans l'élaboration de ces fonctions.

3. Les fonctions liées aux accidents cérébrovasculaires ont intégré l'utilisation de médicaments contre l'hypertension. On travaille actuellement à l'intégration de l'utilisation de tels médicaments dans les fonctions liées à la MC.

4. Des modèles approuvés destinée à l'usage professionnel ont été élaborés pour l'évaluation des risques de MC et d'accidents cérébrovasculaires.

5. Des modèles accessibles aux non-professionnels ont également été élaborés pour les MCV et les accidents cérébrovasculaires.

6. Des modèle de Framingham pour les événements secondaires ont récemment été mis au point. Les fonction préliminaires seront publiées par l'*American College of Cardiology*. Les travaux à ce sujet ne font que commencer.

7. On travaille à l'intégration dans les modèles de l'utilisation de substances telles que les triglycérides et le fibrinogène.

8. On travaille actuellement à l'élaboration de fonctions pour des événements différents tels que le cancer.

9. Les fonctions d'évaluation des risques pour la santé de l'étude de Framingham ont été utilisées avec succès dans des contextes différents de celui initialement visé par l'étude.

BIBLIOGRAPHIE

Abbott, R.D., et McGee, D. (1987). Section 37: the probability of developing certain cardiovascular disease in eight years at specific values of some characteristics, in Kannel, W.B., Wolf, P.A. et Garrison, R.J. (éd.), *The Framingham Study, an Epidemiological Investigation of Cardiovascular Disease*, DHHS PHS NIH Pub. 87-2284.

Anderson, K.M., Wilson, P.W.F., Odell, P.M., et Kannel, W.B. (1991a). An updated coronary risk profile: a statement for health professional. *Circulation*, 83, 356-362.

Anderson, K.M., Odell, P.M., Wilson, P.W.F., et Kannel, W.B. (1991b). Cardiovascular risk profiles. *American Heart Journal*, 121, 293-298.

Cupples, L.A., et D'Agostino R.B., (1987). Section 34: some risk factors related to the annual incidence of cardiovascular disease and death using repeated biennial measurements: Framingham heart study, 30-year follow-up, in Kannel, W.B., Wolf, P.A. et Garrison, R.J., (éd.), *The Framingham Study, an Epidemiological Investigation of Cardiovascular Disease*, DHHS PHS NIH Pub. 87-2703 (NTIS PB870177499), Washington, DC.

Cupples, L.A., D'Agostino, R.B., Anderson K., et Kannel W.B. (1980). Comparison of baseline and repeated measure covariate techniques in the Framingham heart study, *Statistics in Medicine*, 7, 205-218.

D'Agostino, Ralph B., et Kannel, W. B. (1990). Epidemiological background and design: the Framingham study, *Proceedings of the American Statistical Association Sesquicentennial Invited Papers Sessions - 1989 & 1988*, 707-718.

D'Agostino, R.B., Lee, M., Belanger, A., Cupples, L.A., Anderson, K., et Kannel, W.B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham heart study, *Statistics in Medicine*, 9, 1501-1515.

D'Agostino, R.B., Wolf, P.A., Belanger, A.J., et Kannel, W.B. (1994). Stroke risk profile: adjustment for antihypertensive medication, *Stroke*, 25, 40-43.

Dawber, Thomas R., Kannel, W.B., et Lyell, L. (1963). An approach to longitudinal studies in a community: the Framingham study, *Annals of the New York Academy of Sciences*, 107, 539-556.

Dawber, Thomas R., Meadors, Gilcin F., et Moore, Felix E. (1951). Epidemiological approaches to heart disease: the Framingham study, *American Journal of Public Health*, 41, 279-286.

Gordon, T., et Kannel, W.B. (1982). Multiple risk functions for predicting coronary heart disease: the concepts, accuracy and application, *American Heart Journal*, 103, 1031-1039.

Gordon, T., Kannel, W.B., et Halperin, M. (1979). Predictability of coronary heart disease, *Journal of Chronic Diseases*, 32, 427-440.

Grover, S.A., Coupal, L., et Xiao-Ping, H. (1995). Identifying adults at increased risk of coronary disease, *Journal of the American Medical Association*, 274, 801-806.

Halperin, M., Blackwelder, W., et Verter, J. (1971). Estimation of the multivariate risk function: a comparison of the discriminant function and maximum likelihood approach, *Journal of Chronic Diseases*, 24, 125-128.

Kannel, W.B., et McGee, D.L. (1987). Composite scoring - methods and predictive validity: insights form the Framingham study, *Health Services Research*, 22, 499-535.

Kannel, W.B., McGee, D., et Gordon, T. (1976). A general cardiovascular risk profile: the Framingham study, *American Journal of Cardiology*, 38, 46-51.

Laurier, D., Chau, N.P., Cazelles, B., Segond, P., et groupe PCV-METRA. (1994). Estimation of CHD risk in a French working population using a modified Framingham model, *Journal of Clinical Epidemiology*, 47, 1353-1364.

Shurtleff, D, (1974). Section 30: some characteristics related to the incidence of cardiovascular disease and death: Framingham Study 18-year follow-up, in Kannel W.B., Gordon T. (éds), *The Framingham Study: an Epidemiological Investigation of Cardiovascular Disease*, US Government printing Office, DHEW publication (NIH) 74-599.

Truett, J., Cornfield, J., et Kannel, W. (1967). A multivariate analysis of the risk of coronary heart disease in the Framingham study, *Journal of Chronic Diseases*, 20, 511-524.

Walker, S.H., et Duncan, D.B. (1967). Estimation of the probability of an event as a function of several independent variables, *Biometrika*, 54, 167-179.

Wolf, P.A., D'Agostino, R.B., Belanger, A.J., et Kannel, W.B. (1991). Probability of stroke: a risk profile from the Framingham study, *Stroke*, 22, 312-318.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

INTERPRÉTATION DES TESTS MULTIVARIÉS

D. R. Thomas[1]

RÉSUMÉ

Nous abordons dans le présent article certaines des méthodes décrites dans la documentation spécialisée pour l'interprétation des tests MANOVA significatifs. Nous insistons sur des mesures dont la mise en oeuvre nécessite uniquement le recours à des progiciels standards tels que SAS et SPSS. Les mesures recommandées dans le présent article sont les coefficients de rapports discriminants proposés par Thomas (1992). On peut les utiliser pour évaluer l'importance relative de variables de réponses individuelles pour un test multivarié significatif, ainsi que pour déterminer plus facilement les concepts sous-jacents associés aux fonctions discriminantes individuelles. Nous donnons pour terminer des exemples de leur application.

MOTS CLÉS : Importance de la variable; MANOVA; fonctions discriminantes; interprétation.

1. INTRODUCTION

1.1 Description du problème

Dans le présent article, nous abordons les problèmes d'interprétation des tests multivariés significatifs découlant des comparaisons de moyennes de groupes, c'est-à-dire de l'analyse multivariée de la variance (MANOVA). Beaucoup d'articles de statistiques décrivent en détails l'élaboration et les propriétés de répartition des méthodes destinées à tester l'égalité des moyennes d'un ensemble de mesures de réponses provenant de deux ou de plusieurs groupes. Ces tests MANOVA classiques et les valeurs de p correspondantes sont facilement accessibles aux spécialistes dans des progiciels tels que SAS et SPSS. Les spécialistes s'intéressent également aux méthodes qui leur permettront d'interpréter ces tests MANOVA déclarés significatifs. Par exemple, ils pourraient souhaiter évaluer l'importance relative des variables de réponses pour un test MANOVA significatif, ou encore interpréter les combinaisons linéaires des variables de réponses, appelées fonctions discriminantes, qui sont associées à chaque test MANOVA. Les articles portant sur l'analyse multivariée, en particulier ceux rédigés par des statisticiens, s'intéressent beaucoup moins à cet aspect de l'analyse multivariée de la variance. Lorsque cette question est abordée, on recommande habituellement d'examiner les coefficients de fonctions discriminantes individuelles, ou d'examiner les corrélations qui existent entre les variables de réponses individuelles et les fonctions discriminantes. Thomas (1992) a présenté un survol du débat actuel, dans les secteurs des sciences du comportement, concernant les méthodes d'interprétation des tests MANOVA. Certaines des questions principales de ce débat seront résumées dans le présent article, et nous tenterons de déterminer les lacunes des méthodes actuelles. Les méthodes de rechange pour la mesure de l'importance des variables proposées par Thomas (1992) et par Thomas et Zumbo (1995) seront décrites et illustrées à l'aide d'ensembles de données réelles. Nous démontrerons que ces mesures, appelées coefficients de rapports discriminants (CRD), peuvent servir à mesurer l'importance des variables en plus d'identifier les concepts sous-jacents qui peuvent donner lieu aux fonctions discriminantes linéaires. Nous décrirons des travaux récents inédits au cours desquels on a procédé à la rotation de fonctions discriminantes multiples afin de maximiser la « structure simple » des vecteurs de CRD. Nous proposerons un exemple visant à démontrer que cette technique peut clarifier l'interprétation des concepts qui différencient les groupes d'étude.

1.2 Objet principal

Dans le présent article, nous porterons une attention particulière aux procédures informelles ou ayant trait à

[1] D. Roland Thomas, School of Business, Carleton University, Ottawa (Ontario) Canada K1S 2T9.

l'analyse des données. En d'autre mots, il ne sera pas question des écarts-types des divers coefficients abordés. En outre, nous présumerons que toutes les observations recueillies sont indépendantes et réparties de façon identique à l'intérieur des groupes; il ne sera donc pas non plus question de données pondérées ni de données tirées d'échantillons en grappes, même s'il reste possible d'étendre la méthode à de tels échantillons complexes. Exception faite du travail expérimental portant sur la rotation des fonctions discriminantes, on s'attachera surtout aux méthodes d'interprétation qui peuvent être mises en application par les praticiens n'ayant accès qu'au logiciel MANOVA standard fourni dans les progiciels SAS, SPSS et d'autres progiciels statistiques semblables. Soulignons finalement que le présent article a été écrit à l'intention de praticiens qui n'ont pas une connaissance étendue de l'analyse multivariée. Une certaine connaissance de la notation matricielle de base ainsi qu'une familiarité avec le test de t à deux groupes et le test de F de l'ANOVA devraient suffire.

2. COMPARAISONS MULTIVARIÉES : DEUX GROUPES

2.1 Contexte statistique

Le contexte statistique nécessaire sera abordé dans la présente section à l'aide d'un exemple à deux groupes. Désignons par $y_1, y_2, \ldots, y_p$ les variables de réponses p qui peuvent être représentées sous forme d'un vecteur y de $p \times 1$ observations. Imaginons deux groupes à partir desquels on échantillonne indépendamment n_1 et n_2 observations, et désignons par y_{jk} le vecteur de la $k^{ième}$ observation ($k = 1, \ldots, n_j$) du $j^{ième}$ groupe ($j = 1,2$). En vertu des hypothèses classiques, nos observations des deux groupes appartiendront à des distributions normales multivariées ayant μ_1 et μ_2 pour moyennes respectives, et une matrice de covariance commune Σ. La comparaison de groupes multivariés consiste donc à tester l'égalité des vecteurs moyens, μ_1 et μ_2 c'est-à-dire à tester l'hypothèse de différence nulle du test multivarié:

$$H_0: \mu_1 = \mu_2, \qquad (1)$$

où $\mu_j = (\mu_{1j}, \mu_{2j}, \ldots, \mu_{pj})'$, $j = 1, 2$, et où les éléments μ_{ij}, $i = 1, \ldots, p$ représentent les moyennes des variables de réponses individuelles, c'est-à-dire, $\mu_{ij} = E(y_{ijk})$. E désigne ici l'espérance en vertu du modèle multivarié normal présumé.

2.2 Exemple 1

Cet exemple se fonde sur des données provenant d'une enquête de Sénéchal, LeFevre, Hudson et Lawson (1995) portant sur les rapports entre les habiletés verbales des enfants et leur contexte d'alphabétisation. Les sujets de l'étude étaient des enfants âgés de quatre à six ans. Pour les besoins de cet exemple, ils seront répartis en deux groupes : le premier jouissant de bonnes habiletés verbales (n_1 =60), et l'autre moins doué à cet égard (n_2= 59). Nous présumerons que les deux groupes sont conformes au modèle statistique décrit plus haut, c'est-à-dire que les problèmes de mauvaise classification seront ignorés. Dans l'étude originale, les données ont été analysées à l'aide d'une analyse de régression et il n'a donc pas été nécessaire de les regrouper. Quatre des variables de l'étude originale seront examinées, chacune représentant une mesure d'un aspect particulier du contexte d'alphabétisation des enfants :

PRINTEXP	Une mesure de l'exposition du principal adulte dispensateur de soins à la documentation écrite, mesurée par la proportion des titres d'ouvrages reconnus dans une liste donnée;
KNTTLSKB	Une mesure des connaissances du principal adulte dispensateur de soins en matière de livres d'enfants, mesurée ici encore par la proportion des titres reconnus dans une liste;
NUMKBKS	Le nombre de livres d'enfants au foyer;
NUMKREAD	Le nombre hebdomadaire de périodes de lecture de l'enfant.

L'analyse décrite ci-après a en fait porté sur les racines carrées de ces variables, puisque les distributions des variables ainsi transformées étaient plus proches d'une distribution normale que celles des variables originales. Les noms des variables ne seront pas changés.

2.3 Test T^2 de Hotelling

Le test approprié de l'hypothèse (1) est le test T^2 de Hotelling. En ce qui a trait aux données sur les habiletés verbales, le test T^2 donne une valeur de p inférieure à 0,001, ce qui, compte tenu d'une différence intergroupe estimative de 0,92 (voir Stevens 1992, page 178), laisse conclure à une grande différence entre les vecteurs des moyennes de groupes. Des tests univariés séparés (tableau 1) laissent constater des différences de groupes significatives entre les moyennes de chacune des variables de réponses individuelles.

Tableau 1. Tests univariés pour les groupes à habiletés verbales grandes ou faibles

Variable	valeur de t	d. de L	valeur de p
PRINTEXP	3,78	117	< 0,001
KNTTLSKB	3,96	117	<0,001
NUMKBKS	3,72	117	<0,001
NUMKREAD	3,22	117	0,002

Les tests T^2 et les mesures utilisées pour interpréter leurs résultats sont plus faciles à comprendre lorsqu'on pose le problème multivarié en des termes univariés. Pour l'exemple des habiletés verbales, imaginons une nouvelle variable Z représentant la combinaison linéaire des quatre variables à l'étude, lesquelles, pour des raisons de commodité, sont respectivement désignées par les symboles y_1 à y_4. Ainsi,

$$Z = a_1y_1 + a_2y_2 + a_3y_3 + a_4y_4, \qquad (2)$$

où les paramètres a_i représentent des poids fixes. Les indices supplémentaires représentant les groupes et les sujets qui les composent ont été omis de l'équation (2) pour des raisons pratiques. Un vecteur de pondération spécifique $\boldsymbol{a}=(a_1,a_2,a_3a_4)'$ génère une valeur de Z pour chaque sujet dans chacun des deux groupes. Désignons par $t(\boldsymbol{a})$ la valeur de t des deux groupes correspondant à cet ensemble particulier de valeurs de $\boldsymbol{a}$. Par exemple, lorsque $\boldsymbol{a}$ = (0,25, 0,25, 0,25, 0,25)', $t(\boldsymbol{a})$ correspond à une valeur de t calculée en utilisant la moyenne des notes obtenues par chaque sujet pour les variables y_1 à y_4. À chaque valeur de $\boldsymbol{a}$ correspond une valeur spécifique de $t(\boldsymbol{a})$, ou une valeur correspondante de $t^2(\boldsymbol{a})$ qu'on utilisera de préférence puisqu'il est plus pratique de travailler avec des valeurs positives. Il paraît naturel de chercher le vecteur de pondération $\boldsymbol{a}$ qui donne la valeur maximale de $t^2(\boldsymbol{a})$, et la valeur maximale de $t^2(\boldsymbol{a})$ ainsi obtenue est la valeur T^2 de Hotelling. Les poids qui donnent la valeur maximale sont appelés coefficients discriminants et la combinaison linéaire Z correspondant aux poids donnant les valeurs optimales est habituellement appelée fonction linéaire discriminante de Fisher. Comme les valeurs de t n'ont pas d'unités, il est clair que la valeur de T^2 ne sera pas touchée si le vecteur **a** des poids discriminants est multiplié par une constante k. On choisit habituellement une valeur telle que $\boldsymbol{a}^E = k\boldsymbol{a}$ satisfera à l'équation $\boldsymbol{a}^{E'}\boldsymbol{S}_E\boldsymbol{a}^E = 1$, un procédé auquel on a donné le nom de normalisation (Thomas et Zumbo, 1995). La valeur $\boldsymbol{S}_E$ est l'estimation mise en commun à l'intérieur de chaque groupe de la matrice de covariance commune Σ. Comme nous le démontrerons dans la prochaine section, les éléments du vecteur de coefficients discriminants $\boldsymbol{a}^E$ et les coefficients apparentés sont fréquemment utilisés aux fins de l'interprétation des tests T^2 significatifs.

3. INTERPRÉTATION D'UN T^2 SIGNIFICATIF

3.1 Importance relative des variables

Pour beaucoup d'analystes, la première étape de l'interprétation d'un test multivarié significatif consiste à déterminer l'apport relatif, ou l'importance des variables individuelles. L'expression « importance relative » ou le simple mot « importance », lorsqu'il s'agit de variables, sont souvent utilisés dans la documentation, mais il est rare qu'ils soient définis avec précision, comme l'ont fait remarquer Kruskall et Majors (1989). L'étude de Pratt (1987) sur le développement axiomatique d'une mesure unique de l'importance de la variable pour la régression multiple constitue l'exception qui confirme la règle. Dans le contexte de l'analyse multivariée de la variance, Huberty et Wisenbaker (1992) ont proposé une définition selon laquelle l'importance de la variable comprendrait l'apport aux valeurs de la fonction discriminante et l'apport aux effets de mise en commun des variables. Nous aborderons plus tard, dans la présente section, les mesures de l'importance des variables que ces auteurs recommandent, en même temps que les méthodes de mesure de rechange proposées par Thomas (1992) et étudiées d'une manière plus approfondie par Thomas et Zumbo (1995).

3.2 Mesure de l'importance de la variable

On convient généralement que les valeurs de t des tests univariés du tableau 1 ne fournissent pas de mesures multivariées utiles de l'importance des variables de réponses individuelles pour un test multivarié significatif puisqu'elles ne permettent pas de tenir compte des corrélations qui existent entre ces variables. On a donc proposé un certain nombre de méthodes multivariées pour la détermination de l'importance de la variable dans l'analyse multivariée de la variance, dont plusieurs seront décrites dans la présente section. Nous nous attacherons en particulier au test T^2 à deux groupes pour aborder plus tard, dans la section 4, le cas des groupes multiples.

3.3 Coefficients discriminants

Il est tout naturel de songer aux coefficients discriminants lorsqu'on recherche des mesures multivariées de l'importance de la variable. Par exemple, si le coefficient a_2 de l'équation (2) est « grand » dans un certain sens, il sera permis de supposer que la deuxième variable est importante puisqu'elle est assortie d'un poids important dans la fonction discriminante qui donne une discrimination maximale des groupes. Toutefois, comme des variables différentes peuvent être mesurées selon des échelles différentes, les coefficients discriminants doivent être considérés avec prudence. Pour les rendre comparables, on a coutume de les réduire, par exemple, en écrivant l'équation (2) sous la forme suivante :

$$Z=(a_1k_1)(y_1/k_1)+\ldots+(a_4k_4)(y_4/k_4), \qquad (3)$$

où les valeurs k_i ont la même échelle que la $i^{ième}$ variable. Les membres ak_i de l'équation (3) sont appelés coefficients discriminants réduits (CDR) et désignés par $b_i = a_i\ k_i, i = 1, \ldots, p$. Le choix des paramètres d'échelle k_i a fait l'objet d'un important débat dans la documentation spécialisée, débat dont Thomas et Zumbo (1995) on récemment présenté un compte rendu. Le choix le plus commun est l'écart-type de y_i, c'est-à-dire, $k_i=(S_{E\,ii})^{½}, i=1,\ldots,p$, où $S_{E\,ii}$ désigne le $i^{ième}$ élément diagonal de la matrice de covariance S_E de l'échantillon intra-groupe. Thomas et Zumbo (1995) ont désigné par $b^{E\,E}$ le vecteur correspondant des CDR, le premier membre de l'exposant représentant la réduction et le second membre représentant la normalisation. Ils ont en outre proposé un ensemble de CDR de rechange, désigné par $b^{T\,T}$, où la normalisation et la réduction sont fondées sur des quantités « totales » au lieu des quantités intra-groupes utilisées pour la définition de $b^{E\,E}$. Ces CDR sont préférables aux $b^{E\,E}$ (voir Thomas et Zumbo, 1995), mais ils doivent être calculés séparément à partir des données fournies par les progiciels standards. Nous présentons au tableau 2 les formes réduites et non réduites correspondant aux données sur les habiletés verbales.

Tableau 2. Coefficients discriminants pour la comparaison des groupes aux habiletés verbales grandes et faibles

Variable	a^E	b^{EE}	b^{TT}
PRNTEXP	2,133	0,440	0,419
KNTTLSKB	2,365	0,352	0,337
NUMKBKS	1,070	0,342	0,325
NUMKREAD	0,410	0,287	0,269

Beaucoup d'auteurs présument implicitement que les CDR convenablement réduits mesurent l'apport aux valeurs de la fonction discriminante. Rencher et Scott (1990) recommandent explicitement que les valeurs absolues des CDR réduits intra-groupes, b^{EE}, soient utilisées pour évaluer l'importance relative des variables pour une discrimination à deux groupes, c'est-à-dire, pour un test T^2 significatif. Thomas et Zumbo (1995) soutiennent cependant que les CDR ne constituent pas des mesures idéales de l'importance de la variable, quelle que soit la méthode de réduction utilisée. Des méthodes de réduction différentes peuvent conduire à des classements différents de l'importance, une observation qui est à l'origine du débat antérieur dans la documentation spécialisée. Pour cette raison et pour d'autres, Thomas et Zumbo (1995) recommandent que les CDR soient remplacés, pour la mesure de l'importance, par les coefficients de rapports discriminants (CRD) proposés par Thomas (1992).

3.4 Coefficients de structure

La plupart des progiciels fournissent les valeurs des corrélations des échantillons intra-groupes entre chaque variable de réponse et chaque fonction discriminante. Ces corrélations, qu'on appelle coefficients de structure (CS) dans la documentation spécialisée, ont également été proposées pour la mesure de l'importance des variables. Toutefois, leur utilisation à cette fin a été discréditée lorsqu'on a observé que, dans le cas où ily a deux groupes, le vecteur des coefficients de structure r^E est proportionnel au vecteur des valeurs de t découlant des tests univariés des variables de réponses individuelles. Les valeurs des CS correspondant à l'exemple des habiletés verbales sont consignées au tableau 3. On peut constater que ces valeurs sont proportionnelles aux valeurs de t présentées au tableau 1. Ainsi, les coefficients de structure ne fournissent pas d'informations de type multivarié et ne constituent donc pas des mesures utiles de l'importance des variables.

3.5 Test de F avec exclusion d'une variable

Les valeurs de F du test avec exclusion d'une variable, désignées par $F_{(i)}, i=1, \ldots, p$, ont été recommandées à l'origine par Huberty (1984) pour la mesure de l'importance des variables dans l'analyse multivariée de la variance. Elles ont par la suite été adoptées par Huberty et Wisenbaker (1992) pour opérationaliser leur interprétation de l'« apport aux effets de groupement » dans l'importance des variables. Pour la $i^{ième}$ variable de réponse, on peut obtenir la valeur de $F_{(i)}$ à partir d'une analyse de la covariance de la variable

y_i, toutes les autres variables de réponses *p-1* étant assimilées à des covariables. La valeur $F_{(i)}$ équivaut donc au test de Rao (1973, page 551) pour les information supplémentaires fournies par y_i. $F_{(i)}$ constitue certainement une mesure valide de l'importance de la variable. Toutefois, dans les cas à groupes multiples où il risque d'exister plus d'une fonction discriminante significative, $F_{(i)}$ ne fournit qu'une mesure globale de l'importance. Ce paramètre ne peut décrire l'apport des variables individuelles à chacune des dimensions de la différence de groupe.

3.6 Coefficients de rapports discriminants

Les CRD ont été proposés à l'origine en guise de mesures de l'importance de la variable par Thomas (1992), qui s'appuyait sur une interprétation géométrique des fonctions discriminantes. Thomas a en outre relevé que les CRD sont analogues aux mesures axiomatiques de Pratt (1987), ce qui contribue à en justifier l'utilisation en guise de mesures de l'importance. Thomas et Zumbo (1995) ont démontré que les CRD ne présentent pas les lacunes attribuées plus haut aux CDR, aux CS et aux valeurs $F_{(i)}$ du test avec exclusion d'une variable. Dans le cas de tests à deux groupes qui ne présentent qu'une seule fonction discriminante, les CRD, désignés par d_i, $i=1,...,p$, peuvent être définis algébriquement par

$$d_i = b_i^{EE} r_i^E = b_i^{TT} r_i^T, \qquad (4)$$

où r_i^T désigne un coefficient de structure défini à l'aide des quantités « totales », au lieu des quantités intra-groupes utilisées pour définir r_i^E. Selon l'équation (4), on peut définir les CRD en utilisant les quantités totales ou les quantités intra-groupes. Ainsi, en préférant les CRD aux CDR pour les mesures de l'importance, ont évite la controverse qui entoure la réduction des CDR. Les valeurs des CRD correspondant à l'exemple des habiletés verbales sont présentées au tableau 3, en regard des valeurs des $F_{(i)}$, des CDR et des CS.

On peut constater à l'examen du tableau 3 que la somme des CRD donne un, ce qui nous permet de juger directement de leur importance relative, c'est-à-dire, de décider de l'importance relative des variables. Thomas (1992) a démontré que chaque valeur de CRD peut être assimilée à une longueur. Par exemple, PRINTEXP représente 31,7 % de la longueur de la fonction discriminante, lorsqu'on assimile cette variable à un vecteur dans l'espace des observations. L'examen du tableau 3 permet de constater qu'en ce qui concerne les données sur les habiletés verbales, les indices du test de *F* avec exclusion d'une variable, les CDR intra-groupes et les CRD donnent tous le même ordre d'importance : la variable qui participe le plus à la discrimination significative du groupe multivarié, mise en évidence par le test T^2 significatif, est PRINTEXP, suivie dans l'ordre par les variables KNTTLSKB, NUMKBKS et finalement NUMKREAD. En règle générale, toutefois, l'ordre d'importance obtenu à l'aide de ces diverses mesures ne sera pas le même.

Tableau 3. Valeurs des CRD (et des autres mesures) pour l'exemple des habiletés verbales

Variable	F(i)	b^{EE}	r^E	CRD (b_i^E x r_i^E)	Rang
PRINTEXP	3,69	0,440	0,719	0,317	1
KNTTLSKB	1,99	0,352	0,753	0,265	2
NUMKBKS	1,97	0,342	0,708	0,242	2
NUMKREAD	1,52	0,287	0,613	0,176	4
				1,000	

3.7 Notes supplémentaires concernant les CRD

Thomas et Zumbo (1995) ont abordé d'autres aspects de l'utilisation des CRD. D'abord, même s'il peut exister des CRD négatifs, la présence de valeurs de CRD négatives importantes indiquerait la présence de variables de réponses fortement corrélées. Ces valeurs pourraient donc être éliminées au moyen d'un ajustement de « crête » (voir Thomas, 1992), ou par l'élimination d'une ou de plusieurs des variables hautement corrélées. Ainsi, les valeurs négatives des CRD n'interdisent pas leur utilisation en guise de mesures de l'importance, ce que Pratt (1987) a lui aussi soutenu à propos des régressions. Deuxièmement, on peut utiliser les CRD pour désigner les variables suppressives, c'est-à-dire les variables dont l'apport à la discrimination significative du groupe découle de leurs liens avec les autres variables de réponses. Une valeur de CRD petite conjuguée à une valeur relativement grande de $|b_i^{TT}|$ est indicative d'une variable suppressive.

4. COMPARAISON MULTIVARIÉE : PLUS DE DEUX GROUPES

4.1 Contexte statistique

Désignons par $g > 2$ le nombre de groupes à partir duquel n_j, $j=1, .., g$ observations ont été tirées indépendamment. Les hypothèses multi-groupes sont un prolongement direct de celles décrites à la section 2 pour deux groupes, c'est-à-dire que les observations provenant de chaque groupe appartiendront à des

distributions normales multivariées à vecteurs moyens μ_j, $j=1, \ldots, g$, et à matrice de covariance commune Σ. L'hypothèse de différence nulle dans ce cas-ci est représentée par H_0: $\mu_1 = \mu_2 = \ldots = \mu_g$, où les éléments des vecteurs moyens sont définis comme nous l'avons fait plus haut, après l'équation (1).

Les tests multivariés de l'hypothèse de différence nulle pour des groupes multiples peuvent être dérivés en utilisant une combinaison linéaire Z des variables de réponse p, avec un vecteur des poids $\boldsymbol{a}$ de p dimensions. Pour un $\boldsymbol{a}$ fixe, une analyse de variance univariée sur Z donne une valeur de F désignée par $F(\boldsymbol{a})$, et, comme dans le cas des deux groupes, les procédures multivariées sont obtenues en maximisant les $F(\boldsymbol{a})$ pour toutes les valeurs possibles de $\boldsymbol{a}$. La maximisation donne l'équation canonique suivante :

$$\boldsymbol{H}\boldsymbol{a} = \lambda \boldsymbol{W}\boldsymbol{a}, \qquad (5)$$

où $\boldsymbol{H}$ et $\boldsymbol{W}$ désignent respectivement l'hypothèse et les matrices inter-groupes de la somme des carrés et des produits croisés (SSCP). L'équation (5) admet des solutions $g^* = min\ (p, g-1)$. Lorsque les valeurs propres $\lambda_j, j=1,\ldots,g^*$ sont inscrites par ordre décroissant, les poids $\boldsymbol{a}_1$ correspondant à λ_1 sont les coefficients discriminants correspondant à la fonction discriminante principale, c'est-à-dire, à la combinaison linéaire des variables qui donnent la valeur F « univariée » maximale.

L'utilisation des guillemets sert ici à rappeler que la valeur maximisée de F n'obéit pas à la distribution de F classique. Le deuxième vecteur des poids $\boldsymbol{a}_2$, correspondant à la valeur propre λ_2, donne la valeur maximale de $F(\boldsymbol{a})$ parmi l'ensemble de la combinaison linéaire des variables de réponses non corrélées avec la première, et ainsi de suite. Les valeurs propres sont proportionnelles aux valeurs maximisées g^* de $F(\boldsymbol{a})$. Pour le cas à deux groupes $g = 2$, l'équation (5) ne donne qu'une seule solution propre, avec $\lambda_1 = 1 = T^2/(N-2)$, et $N = n_1 + n_2$. Pour le cas à plusieurs groupes $g > 2$, l'équation (5) admet deux solutions propres ou plus. Dans ce cas, il existe plusieurs valeurs statistiques multivariées qui sont toutes des fonctions des valeurs propres λ_j (voir Stevens 1992, page 226). On peut également avoir recours à un test séquentiel pour déterminer le nombre des fonctions discriminantes g^* possibles qu'il convient de retenir afin de décrire entièrement les différences de groupes (voir, par exemple, Stevens, 1992, page 275).

Toutes les mesures de l'importance de la variable présentées dans la section 3 pour le cas des deux groupes peuvent s'appliquer aux cas des groupes multiples, avec $g^* > 2$ fonctions discriminantes. Les valeurs correspondantes des CDR, des CS et des CRD correspondants peuvent être définies séparément pour chaque fonction discriminante, la dernière étant fondée sur l'équation (4), comme dans le cas des deux groupes. Les valeurs des CS donnent uniquement des informations univariées et doivent donc être rejetées (Rencher, 1992). Les critiques à l'endroit des CDR s'appliquent toujours, ce qui nous laisse les CRD comme unique choix des mesures de l'importance dans tous les cas où on s'intéresse à l'apport des variables individuelles à chacune des fonctions discriminantes significatives.

4.2 Exemple 2

L'exemple qui suit est tiré d'une étude sur le commerce international réalisée par Papadopoulos, Heslop et Bennett en 1990. Le sous-ensemble de données que nous examinerons à trait aux perceptions de la Russie et de ses habitants exprimées par trois groupes de 150 répondants du Canada, des États-Unis et de l'Australie. Les six variables de réponses examinées sont :

ALIGN	Le répondant considère que son pays est aligné (ou non aligné) avec la Russie
IMMIG	Le répondant serait favorable (ou opposé) à l'immigration russe.
INDUS	Le répondant considère que la Russie est un pays industrialisé (ou non industrialisé)
INVEST	Le répondant est favorable (ou opposé) à la multiplication des investissements en Russie
TIES	Le répondant est favorable (ou opposé) à l'établissement de liens plus étroits avec la Russie
TRUST	Le répondant considère que les Russes sont (ou ne sont pas) des gens en qui on peut avoir confiance.

Les réponses sont notées de un (négative) à sept (positive). Le test MANOVA ordinaire utilisant la procédure Λ de Wilks (Stevens 1992, page 191) laisse constater des différences significatives ($p < 0{,}001$) entre les moyennes des trois groupes de consommateurs, et un test de l'effet résiduel attribuable à la deuxième fonction discriminante montre que les deux dimensions des différences entre les groupes sont importantes ($p = 0{,}008$). En conséquence, le nombre maximal de fonctions discriminantes ($g^* = 2$) sera retenu pour cet exemple.

4.3 « Identification » des deux dimensions des différences inter-groupes significatives

Les CRD peuvent servir à déterminer l'importance relative des variables de réponses individuelles pour chaque dimension de la différence de groupe, c'est-à-dire, pour chaque fonction discriminante. Pour beaucoup d'analystes, l'étape suivante de l'interprétation d'un test multivarié significatif consiste à déterminer les concepts sous-jacents des fonctions discriminantes retenues. Beaucoup de spécialistes, par exemple, Huberty et Wisenbaker (1992), recommandent d'utiliser les coefficients de structure à cette fin. Toutefois, Thomas (1992) pense qu'on devrait plutôt utiliser les CRD. Nous ferons ci-après l'illustration de cette approche en utilisant l'exemple de la perception des consommateurs. Nous présentons au tableau 4 les valeurs des CDR et des CRD pour les deux fonctions discriminantes significatives. On peut constater que les variables ALIGN, IMMIG, INDUS et INVEST jouent une part importante dans la première fonction discriminante, tandis que les variables TRUST, ALIGN et IMMIG sont, dans l'ordre, les variables qui ont le plus d'importance dans la seconde fonction discriminante.

Si on en juge par les signes des valeurs des CDR correspondant aux variables importantes, on peut prévoir, selon le tableau 4, que les répondants obtiendront la note la plus élevée pour la première fonction discriminante si :

ils pensent que leur pays EST aligné avec la Russie;
ils sont FAVORABLES à l'immigration russe;
ils pensent que la Russie N'EST PAS un pays industrialisé;
ils sont FAVORABLES à une multiplication des investissements en Russie.

Les répondants obtiendront par contre la note la plus élevée pour la seconde fonction discriminante si :

ils pensent que leur pays N'EST PAS aligné avec la Russie;
ils sont FAVORABLES à l'immigration russe;
ils pensent que les Russes SONT des gens en qui on peut avoir confiance.

Ainsi, la première fonction discriminante peut être assimilée à un concept « politico-économique », tandis que la seconde fonction discriminante peut être assimilée à un concept « politico-social ».

Toutefois, la présence des variables ALIGN et IMMIG dans les deux fonctions discriminantes signifie qu'il n'existe pas de séparation claire entre ces deux concepts. En analyse factorielle, on peut résoudre le chevauchement entre deux facteurs par rotation de la saturation des facteurs. Thomas (1995) a étudié les avantages d'une rotation des fonctions discriminantes dans l'analyse multivariée de la variance. Il a observé que même si la rotation détruit la propriété de maximisation des fonctions discriminantes, elle préserve néanmoins le $\lambda_j / (1 + \lambda_j)$, c'est-à-dire, la valeur du critère de Pillai-Bartlett, une des valeurs statistiques standard du test MANOVA. Thomas (1995) a conçu un algorithme de rotation afin de maximiser la « simplicité » des CRD, procédure analogue à la maximisation de la simplicité de la saturation des facteurs en analyse factorielle. Les résultats correspondant à l'exemple de la perception des consommateurs sont présentés au tableau 5.

Tableau 4. Valeurs des CDR et des CRD
Exemple de la perception des consommateurs

	1ère fonc. disc.		2e fonc. disc.	
Variable	b^{EE}	**CRD**	b^{EE}	**CRD**
ALIGN	0,410	**0,204**	-0,479	**0,160**
IMMIG	0,556	**0,368**	0,450	**0,278**
INDUS	-0,465	**0,209**	0,211	0,061
INVEST	0,366	**0,233**	-0,096	-0,024
TIES	0,148	0,062	-0,330	-0,040
TRUST	-0,264	-0,076	0,803	**0,564**

Tableau 5. Valeurs des CRD après rotation de la fonction discriminante — Exemple de la perception des consommateurs

	1ère fonc. discr.		2e fonc. discr.	
Variable	**CRD**	**Signe (CRD)**	**CRD**	**Signe (CRD)**
ALIGN	**0,366**	+	-0,002	-
IMMIG	0,047	+	**0,599**	+
INDUS	**0,269**	-	0,011	-
INVEST	**0,145**	+	0,064	+
TIES	0,090	+	-0,068	-
TRUST	0,084	-	**0,405**	+

On a clairement réalisé la structure simple des CRD. Les variables qui ont joué un rôle important dans la première fonction après rotation sont, par ordre décroissant, ALIGN, INDUS et INVEST, et dans la seconde fonction, IMMIG et TRUST. Grâce à la rotation, il devient possible d'assimiler les deux fonctions discriminantes à deux concepts substantivement séparés, c'est-à-dire, qui ne se superposent pas. Le premier concept peut être qualifié de « contact politico-économique » et le second de « contact social ».

4.4 Affichage des dimensions des différences de groupes

Les méthodologistes recommandent habituellement que les moyennes de groupes des valeurs de la fonction discriminante (centres de gravité) soit affichées dans le « plan discriminant » (voir par exemple, Stevens 1992, page 277). Nous présentons à la figure 1 un graphique des centres de gravité de la fonction discriminante avant rotation, où les deux fonctions discriminantes sont représentées par des axes orthogonaux. Cette représentation n'est pas correcte à strictement parler. La figure 1 représente un sous-espace bidimensionnel de l'espace des variables des réponses, ce qui fait que les deux axes discriminants ne sont pas orthogonaux puisque les vecteurs de pondération discriminants $\mathbf{a}_1$ et $\mathbf{a}_2$ ne le sont pas non plus. Néanmoins, on utilise souvent cette méthode de représentation graphique qui nous fournit une interprétation raisonnable des différence de groupes. On peut y constater que les consommateurs canadiens et américains fournissent une interprétation semblable de la dimension « politico-économique » de la Russie, au contraire des Australiens. En ce qui concerne la dimension « politico-sociale », les données canadiennes sont beaucoup plus élevées que celles des Américains, tandis que les Australiens tendent vers la neutralité.

Il est possible de générer une représentation exacte des différences de groupes au moyen d'une interprétation géométrique du test MANOVA dans l'espace N-dimensionnel des observations, tel que le décrit Thomas (1992). Les fonctions discriminantes dans l'espace des observations fournissent des axes orthogonaux en fonction desquels les groupes peuvent être représentés sous forme de vecteurs. Pour l'exemple de la perception des consommateurs, cette approche donne un graphique bidimensionnel qui confirme l'interprétation de la figure 1. Une approche semblable fournit un tableau des différences de groupes en ce qui concerne les fonctions discriminantes après rotation et, ici encore, les conclusions concernant la perception des consommateurs sont semblables à celles susmentionnées. Le manque d'espace nous empêche de fournir des détails supplémentaires.

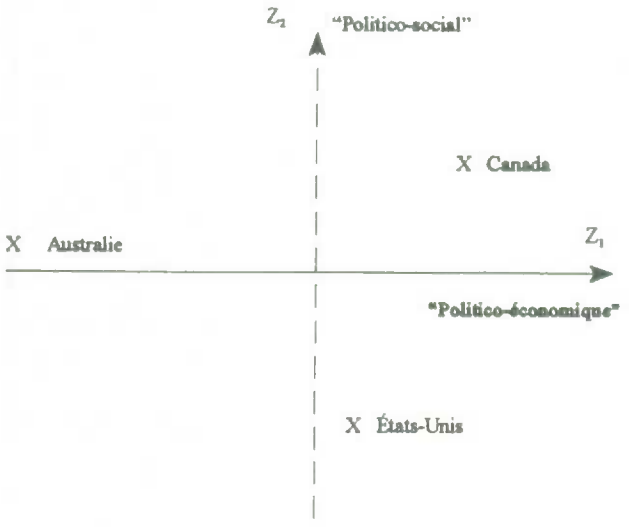


Figure 1. Centres de gravité des groupes pour l'exemple de la perception des consommateurs dans le plan discriminant

5. RÉSUMÉ ET CONCLUSIONS

Dans le présent article, nous avons fourni un aperçu des méthodes d'interprétation des tests T^2 et MANOVA significatifs. Nous avons insisté sur la méthode dont l'utilisation demande uniquement un logiciel MANOVA standard. Après un bref examen des méthodes communément décrites dans la documentation spécialisée, nous avons étudié et illustré les raisons justifiant l'utilisation des coefficients de rapports discriminants (CRD). Nous avons démontré que les CRD donnent la possibilité de : 1) mesurer l'importance des variables de réponses pour chacune des fonctions discriminantes retenues; 2) d'identifier les concepts sous-jacents possibles liés aux fonctions discriminantes; 3) de définir un critère pour la rotation de la fonction discriminante. Nous avons finalement présenté un bref aperçu des stratégies de présentation des différences de groupes en rapport avec les fonctions discriminantes. Nous avons fourni suffisamment de détails pour permettre à un analyste d'explorer l'application des CRD à l'interprétation de tout exemple de test T^2 ou MANOVA. Les analystes qui souhaitent faire l'expérience de la rotation des fonctions discriminantes trouveront les détails techniques requis dans Thomas (1995).

6. BIBLIOGRAPHIE

Huberty, C.J., et Wisenbaker, J.M. (1992). Variable importance in multivariate group comparisons, *Journal of Educational Statistics*, 17, 75-91.

Kruskall, W., et Majors, R. (1989). Concepts of relative importance in recent scientific literature, *The American Statistician*, 43, 2-6.

Papadopoulos, N., Heslop, L.A., et Bennett, D. (1993). National image correlates of product stereotypes: A study of attitudes towards East European countries, dans F. van Raaj et G. Bamossy (éds.) *European Advances in Consumer Behaviour*, 206-213. Amsterdam, Pays-Bas : Association for Consumer Research.

Pratt, J.W. (1987). Dividing the indivisible: using simple symmetry to partition variance explained, dans T. Pukkila et S. Puntanen (éds), *Proceedings of the Second International Conference in Statistics* ,245-260, Tampere, Finlande: Université de Tampere.

Rao, C.R. (1973). *Linear Statistical Inference, 2e éd.* New York : Wiley.

Rencher, A.C. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components, *The American Statistician*, 46, 217- 225.

Rencher, A.C., et Scott, D.T. (1990). Assessing the contribution of individual variables following rejection of a multivariate hypothesis, *Communications in Statistics, Part B-Simulation and Computation*, 19, 535-553.

Sénéchal, M., LeFevre, J.-A., Hudson, E., et Lawson, E.P. (1995). Knowledge of picture books as a predictor of young children's vocabulary, Rapport inédit, Department of Psychology, Carleton University, Ottawa, Canada.

Stevens, J. (1992). *Applied Multivariate Statistics for the Social Sciences*, New Jersey: Lawrence Erlbaum Associates.

Thomas, D.R. (1992). Interpreting discriminant functions: A data analytic approach, *Multivariate Behavioural Research*, 27, 335-362.

Thomas, D.R. (1995). Interpreting significant MANOVA tests: Discriminant ratio coefficients and discriminant function rotation, Working Paper, WPS 95-07, School of Business, Carleton University, Ottawa, Canada.

Thomas, D.R., et Zumbo, B.D. (1995). Using a measure of variable importance to investigate the standardization of discriminant coefficients, *Journal of Educational and Behavioural Statistics* (sous presse).

SESSION 3

Accès et contrôle des données

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

PROTECTION DES RENSEIGNEMENTS PERSONNELS

D.C.G. Brown[1]

RÉSUMÉ

Un nombre croissant de renseignements sur les individus pouvant désormais être traités très rapidement, les Canadiens s'interrogent de plus en plus sur la protection des renseignements personnels qu'ils fournissent. Le gouvernement fédéral est le plus important organe d'archivage de renseignements personnels sur les Canadiens. Or, la *Loi sur la protection des renseignements personnels* définit le droit des individus à contrôler les renseignements que des institutions fédérales détiennent à leur sujet. Même si l'application des principes de confidentialité semble parfois diminuer l'efficacité des opérations, les renseignements personnels qu'exigent nombre de programmes gouvernementaux ne seront obtenus que dans un climat de confiance.

MOTS CLÉS: Protection des renseignements personnels; collecte de renseignements; principes de confidentialité; *Loi sur la protection des renseignements personnels.*

Il est évident, aujourd'hui, que les progrès récents dans le domaine de l'informatique ont eu un effet profond sur le traitement et sur la gestion des renseignements, en particulier sur la quantité de renseignements traités et sur la vitesse à laquelle ils le sont.

Des sondages récents indiquent que plus les Canadiens prennent conscience de l'évolution de la technologie et des possibilités en matière de collecte, de stockage et de manipulation des renseignements personnels, plus ils s'inquiètent des répercussions éventuelles de ces progrès sur la protection des renseignements qui les concernent.

Par "protection des renseignements personnels", j'entends le principe suivant lequel l'individu contrôle, ou détermine lui-même, quels renseignements le concernant seront communiqués. Plus précisément, selon ce principe, les individus devraient savoir quels renseignements sont collectés à leur sujet, par qui et à quelles fins, et, sauf dans des cas exceptionnels, consentir à leur collecte et à leur utilisation.

L'administration fédérale est non seulement un important collecteur, utilisateur et gestionnaire de renseignements en général, mais aussi le plus important organe d'archivage de renseignements personnels sur les Canadiens. Ces derniers fournissent une foule de données personnelles au gouvernement fédéral à diverses fins, notamment pour demander des avantages ou des prestations auxquelles ils ont droit, pour obtenir un numéro d'assurance sociale ou un passeport, pour justifier leur versement d'impôt sur le revenu, pour obtenir des permis et, bien sûr, lorsqu'ils répondent au questionnaire de recensement.

Le gouvernement du Canada a d'abord reconnu officiellement les préoccupations des membres de la population quant à la manipulation des renseignements personnels qui les concernent en ajoutant à la *Loi sur les droits de la personne* un article sur le traitement des renseignements personnels par l'administration fédérale. Puis, en 1983, les dispositions de la *Loi sur la protection des renseignements personnels* ont élargi et précisé les principes généraux de la protection des renseignements personnels sous le contrôle de l'administration fédérale. En outre, d'autres textes législatifs (tels que la *Loi de l'impôt sur le revenu*, la *Loi sur la statistique* et la *Loi sur l'assurance-chômage*) contiennent des dispositions qui régissent de façon précise le traitement des renseignements personnels dans le cadre de certains programmes ou au sein d'institutions particulières.

Selon des sondages récents, en ce qui a trait à la collecte et à l'utilisation de renseignements personnels, les Canadiens craignent surtout que les renseignement recueillis soient inexacts ou utilisés à mauvais escient

[1] David C.G. Brown, Directeur exécutif, Division des pratiques de l'information, des communications et de la sécurité, Secrétariat du Conseil du Trésor, Ottawa (Ontario), Canada, K1A 0R5.

quand des décisions sont prises au sujet de la personne qui les a fournis. La collecte et l'utilisation de renseignements personnels à des fins strictement statistiques n'est, quant à elle, pas contrôlée aussi minutieusement, car elle n'est généralement pas perçue comme posant un risque pour la protection de la vie privée.

Les principes enchâssés dans la *Loi sur la protection des renseignements personnels* en rapport le plus direct avec la collecte de données statistiques sont les suivants :

- l'obligation d'informer les individus de la raison pour laquelle les renseignements sont recueillis;
- l'interdiction d'utiliser des renseignements personnels à d'autres fins que celles pour lesquelles ils sont recueillis;
- l'obligation de donner aux individus, sur demande, accès aux renseignements permettant de les identifier;
- l'obligation d'éliminer les renseignements personnels d'une manière conforme aux normes relatives à la classification de sécurité;
- l'obligation de protéger les renseignements personnels contre la divulgation non autorisée.

Examinons maintenant ces principes l'un après l'autre.

L'obligation d'informer les individus de la raison pour laquelle les renseignements personnels sont recueillis. Les membres du public craignent de ne pouvoir contrôler les renseignements qui les concernent s'ils ne savent pas comment ces derniers seront utilisés et quelles seront les conséquences de cette utilisation. Cela signifie, évidemment, que les personnes à la recherche de renseignements doivent se préparer et, avant d'entamer la collecte, définir comment les renseignements recueillis seront utilisés. Une personne à qui on demande de fournir des renseignements dans le cadre d'un programme gouvernemental doit non seulement être informée de l'objectif de la collecte de données, mais aussi des mesures législatives autorisant cette dernière, du caractère obligatoire ou volontaire de la fourniture des renseignements et des conséquences éventuelles du refus d'obtempérer.

L'interdiction d'utiliser des renseignements personnels à d'autres fins que celles pour lesquelles ils sont recueillis signifie qu'on ne peut dire aux sujets d'une enquête que les renseignements qu'ils fournissent seront utilisés uniquement à des fins statistiques (par exemple, pour étudier la proportion de pilotes titulaires d'un permis qui détiennent également un permis d'opérateur radio), puis décider d'utiliser les renseignements à d'autres fins (par exemple, des poursuites). Les procédures concernant l'utilisation des renseignements à d'autres fins compatibles avec l'objectif original incluent la notification du Commissaire à la protection de la vie privée, qui détermine s'il convient d'avertir les personnes concernées de la nouvelle utilisation des renseignements.

Afin de satisfaire **l'obligation de donner aux individus, sur demande, accès aux renseignements qui les concernent**, il faut disposer d'une méthode permettant de retracer la diffusion des renseignements identifiables et d'établir les responsabilités en ce qui concerne ces renseignements, de sorte qu'il n'y ait aucun doute sur l'identité de la personne chargée de répondre aux demandes ou d'enregistrer les plaintes.

L'obligation d'éliminer les renseignements personnels d'une manière conforme aux normes relatives à la classification de sécurité s'inscrit dans le cadre élargi du principe général de bonne gestion de l'information selon lequel les institutions doivent disposer comme il convient de tous les fonds de renseignements administratifs publics, principe conforme à la dernière notion susmentionnée, à savoir la sécurité des renseignements personnels.

La nécessité de protéger les renseignements personnels contre l'altération ou contre la divulgation non autorisée signifie qu'il faut prendre les mesures de sécurité appropriées pour s'assurer qu'aucune personne non autorisée ne puisse avoir accès aux renseignements recueillis et que ces derniers ne soient pas indûment divulgués, altérés ou détruits.

Malgré l'intégration de mesures de contrôle et de mise en application des règlements, l'exécution de la plupart des programmes fédéraux exigeant la collecte et le traitement de renseignements personnels dépend en grande partie de la volonté des membres du public de fournir des renseignements personnels complets et exacts. À l'heure actuelle, les Canadiens sont, en général, prêts à confier des renseignements personnels à l'administration publique, à condition que ces renseignements ne soient pas utilisés abusivement ou à des fins non divulguées. Si, pour une quelconque raison, ils en venaient à perdre confiance en l'administration, on assisterait à une baisse de la qualité des renseignements fournis volontairement. L'exécution des programmes fédéraux exigerait alors qu'on déploie beaucoup plus de

ressources pour vérifier les renseignements et faire respecter les règlements, et la valeur prédictive des statistiques tirées de ces programme diminuerait considérablement.

Il convient de noter que, grâce à certains progrès techniques, il est désormais possible de recueillir des renseignements personnels à l'insu de la personne concernée. Par exemple, l'utilisation des fonctions «affichage du numéro» ou «affichage du nom» offertes par les compagnies de téléphone parallèlement à un «annuaire inversé» permet aux entreprises (ou aux institutions gouvernementales) de savoir qui (ou quel ménage) a téléphoné. Ce type de renseignements pourrait, par exemple, aider une entreprise à élaborer sa stratégie de commercialisation.

Le nombre croissant de services "en direct" donne aussi la possibilité d'identifier les utilisateurs à leur insu ou de retracer l'usage qu'ils font du système ou des services. Ces «données transactionnelles» permettent aux collecteurs de données d'établir des profils d'utilisateurs vraisemblablement précieux dans diverses situations. À l'échelon fédéral, la *Loi sur la protection des renseignements personnels* interdit aux institutions publiques d'établir de tels profils à l'insu des sujets. Le Manuel du Conseil du Trésor sur la Protection des renseignements personnels contient un chapitre qui traite spécifiquement de l'appariement des données par les institutions publiques, et décrit les critères d'évaluation et d'établissement des programmes d'appariement.

À mesure que les institutions ont augmenté leur capacité de collecte et de manipulation des données, la tentation s'est accrue de fusionner les bases de données afin d'améliorer davantage le rendement. Il convient, toutefois, de maintenir un équilibre délicat entre les gains d'efficacité éventuels et le respect des principes de confidentialité exposés précédemment. La façon dont les membres du public perçoivent la situation est un autre facteur dont il faut tenir compte. Si d'aucuns défendent l'idée d'une base de données centralisée de renseignements personnels (ou même d'un numéro d'identification national), de telles propositions paraissent inacceptables aux Canadiens en général et semblent susciter la crainte de l'avénement du «dictateur»

Au fil des ans, le personnel de Statistique Canada s'est montré sensible aux préoccupations des Canadiens en ce qui concerne la protection des renseignements personnels. Cette attitude a rassuré les membres du public et soutenu le flot de renseignements personnels dont a besoin le Bureau, exploit remarquable à une époque où la défiance du public à l'égard des gouvernants et de la technologie ne fait que croître, et je félicite les employés du Bureau de leur succès.

Les membres de la Division des pratiques de l'information, des communication et de la sécurité du Conseil du Trésor seraient heureux d'aider, dans la mesure du possible, tout ministère obligé de trouver la juste mesure entre l'efficacité et la protection des renseignements personnels, puisque, selon eux, ces deux concepts ne sont pas nécessairement contradictoires. Un simple exercice de planification supplémentaire durant la conception des systèmes informatiques pourrait suffire à mettre l'un en pratique sans sacrifier l'autre.

BIBLIOGRAPHIE

Loi sur la protection des renseignements personnels, L.R., 1985, c. P-21.

Règlements sur la protection des renseignements personnels, DORS/83-508.

Manuel du Conseil du Trésor, Protection des renseignements personnels, 1993.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

FICHIERS DE POPULATION ET PROTECTION DE LA VIE PRIVÉE. L'EXPÉRIENCE DU FICHIER BALSAC DEPUIS 1972

G. Bouchard[1]

RÉSUMÉ

Le fichier BALSAC est une banque de données informatisées constituée à partir du jumelage des actes de naissance, mariage et sépulture. Couvrant les 19e et 20e siècles, ce fichier est complété pour les régions du saguenay et de Charlevoix et il est en cours de réalisation pour les autres régions du Québec (bien qu'à cette échelle, la collecte des données porte principalement sur les actes de mariage seulement).

Géré par l'Institut interuniversitaire de recherches sur les populations (IREP) et possédé par un consortium d'Universités, le fichier est utilisé exclusivement à des fins de recherche scientifique dans le cadre de trois programmes, portant respectivement sur les structures démographiques et sociales, la génétique des populations et les dynamiques culturelles.

Comme on le devine, l'exploitation de la banque est assortie d'un important système de contrôles et de restrictions de nature à protéger adéquatement la confidentialité pour tout ce qui touche à la collecte, la conservation et l'utilisation des données (tels que: surveillance par la Commission d'accès à l'information du Québec, responsabilité publique, permissions d'accès accordées par des instances indépendantes, assermentation des usagers, engagements contractuels, restrictions physiques et informatiques, etc).

MOTS CLÉS: Banque de données; protection de la confidentialité; études de population; IREP.

1. DROITS DES PERSONNES ET BANQUES DE DONNÉES NOMINATIVES

Le développement et l'exploitation des banques de données nominatives informatisées peuvent engendrer, à des degrés divers, des problèmes relatifs à la protection de la confidentialité des informations et à la protection de la vie privée des personnes. Par confidentialité, on désigne ici le caractère de certaines données nominatives dont la divulgation ou la diffusion non contrôlée (par exemple sans le consentement des individus concernés) peut entraîner un préjudice et une violation des droits individuels. Par vie privée, on entend l'univers des informations personnelles ou familiales, sur lesquelles chaque individu détient un contrôle prioritaire. Ainsi, une information médicale est confidentielle dans la mesure où sa divulgation inopportune peut retarder une promotion professionnelle, ternir une réputation, compromettre une vie familiale, etc. Apparemment plus inoffensives, des informations personnelles à caractère économique et culturel peuvent entraîner des effets analogues lorsque diffusées dans un contexte particulier. En outre, des opérations de jumelage peuvent changer la nature des données, en rendant confidentielles des informations qui au départ ne l'étaient pas. La reconstitution des familles à partir des actes de l'état civil en fournit un bon exemple: une fois jumelés, les actes de mariage et de naissance permettent de calculer les intervalles protogénésiques et d'identifier les cas de conception prénuptiale. En certaines circonstances, ils peuvent aussi conduire à détecter des naissances dites illégitimes, identifier les personnes apparentées à des sujets atteints de maladies héréditaires[2], etc.

En conséquence, la gestion d'une banque de données nominatives informatisées doit normalement être assortie d'un système ou protocole destiné à assurer la protection de la vie privée. Ces protocoles consistent

[1] Gérard Bouchard, directeur, Institut interuniversitaire de recherches sur les populations (IREP), 555 boul. de l'Université, Chicoutimi, (Québec), Canada, G7H 2B1.

[2] Ces exemples peuvent sembler anodins; mais notre expérience nous a montré que, pour plusieurs individus concernés, ils ne le sont pas.

dans un ensemble de directives et de procédures régissant les conditions de collecte, de conservation, d'accès, d'utilisation et de diffusion de données. Dans leur constitution, ils doivent tenir compte d'abord de l'environnement juridique ou législatif propre à chaque collectivité, et ensuite de ce que prescrit la morale ou l'éthique collective pour tout ce qui concerne les questions ou matières sur lesquelles le législateur ne s'est pas encore prononcé.

A l'échelle canadienne, l'environnement juridique est déterminé en partie par la législation fédérale et en partie par les législations provinciales, d'où une diversité considérable. Quant aux sensibilités collectives en matière d'éthique, elles sont également très variées, comme on s'en doute. En conséquence, dans le cadre de ce texte, il est tout à fait impossible de proposer un modèle qui serait valable à l'échelle canadienne, la variété des contextes et des situations appelant une variété de solutions. Aussi, dans le but d'alimenter la réflexion et de soumettre une sorte de repère parmi d'autres possibles, il a paru utile d'exposer une démarche particulière, en l'occurrence celle qui a été suivie par notre Institut depuis le début de ses travaux en 1972[3]. Chacun pourra en tirer tout au moins une idée des problèmes qui se posent et du type de dispositions à prendre pour les surmonter.

2. LE FICHIER DE POPULATION BALSAC

Le fichier BALSAC (qui doit son nom aux lettres initiales de plusieurs régions qu'il recouvre) est une forme particulière de banque de données nominatives informatisées. Par définition, un fichier de population est une banque de données qui présente les caractéristiques suivantes:

2.1 les informations qu'elle contient (sur les professions, les résidences, les décès, etc) doivent être explicitement rattachées à des individus;
2.2 ces données nominatives sont raccordées par le biais de programmes de jumelage de manière à mettre en relation, du moins théoriquement, les données relatives à une même personne;
2.3 à un degré variable selon les fichiers, les données accumulées comportent une dimension historique ou rétrospective;
2.4 par le biais des filiations généalogiques, le contenu du fichier peut être parcouru automatiquement en diachronie, sur plusieurs générations[4];
2.5 le fichier est assorti de logiciels d'accès multiple qui permettent de "naviguer" au sein de ses composantes (tables, champs...).

Du point de vue technique, une telle structure peut être concrétisée grâce à l'utilisation de logiciels SGBD (systèmes de gestion de base de données). Concrètement, le fichier BALSAC consiste dans un fichier central, qui contient le signalement de tous les individus impliqués (modules A, B et C de la Figure 1; aussi: Figure 2) et un nombre indéfini de fichiers dits sectoriels contenant des informations spécialisées, à caractère économique, social, culturel, etc. Implanté à l'aide du système de gestion INGRES, BALSAC contient présentement près de 2 millions d'actes de baptême, mariage et sépulture étalés sur les 19e et 20e siècles. Il recouvre actuellement toute la population du Saguenay et de Charlevoix (saisie, jumelage et validation complétés), et il est en cours d'extension sur l'ensemble des régions de la province de Québec (G. Bouchard, 1992) (Carte 1).

Ce fichier, dont la construction a débuté en 1972, est utilisé dans le cadre de trois programmes de recherches. Le premier relève des sciences sociales en général (incluant la géographie et la démographie), le deuxième relève de la génétique humaine, alors que le troisième porte sur les phénomènes culturels. Dans le premier volet, de nombreuses enquêtes ont été réalisées ou sont en cours sur des thèmes comme les mouvements migratoires, le déclin de la fécondité, les rapports villes-campagnes, la formation de la main-d'oeuvre industrielle, la reproduction familiale, les inégalités socio-économiques, etc. Ces enquêtes visent à reconstituer les dynamiques collectives régionales et à en marquer les disparités et les ruptures. Dans le second volet, les travaux se répartissent entre la génétique des populations (apparentement, effets fondateurs, flux génique...) et l'épidémiologie génétique. Dans cette dernière direction, les travaux portent sur la diffusion et la distribution spatiale des gènes, les filières généalogiques de transmission, les paramètres démographiques, économiques et sociaux du risque de génopathie dans les populations. L'objectif poursuivi

[3] L'IREP (Institut interuniversitaire de recherches sur les populations) relève conjointement de l'Université du Québec à Chicoutimi, l'Université Laval, l'Université McGill, l'Université de Montréal, l'Université Concordia et l'Université de Sherbrooke.

[4] Sur ce qui précède, voir G. Bouchard et coll., 1985, 1989; G. Bouchard, 1988, 1992.

dans ce volet est de déterminer la nature et l'évolution du risque à l'échelle régionale et interrégionale et de contribuer ainsi à la prévention des maladies génétiques ou à composante génétique. Dans le troisième programme de recherche, les travaux portent sur les dynamiques culturelles (croyances et comportements religieux, alphabétisation, modèles de prénomination, etc)[5].

3. LES PROBLÈMES DE DROIT ET D'ÉTHIQUE

Les problèmes posés du point de vue éthique et juridique se présentent sous différentes formes. L'énumération que nous en faisons ici n'est certes pas exhaustive et, encore une fois, elle reflète une démarche particulière qui est celle de l'IREP, dans un contexte spécifique. Toutefois, on pourra constater que, dans leurs coordonnées essentielles, les questions évoquées se posent d'une manière assez générale.

3.1 L'accès aux données et la diffusion des informations

La préoccupation la plus élémentaire concerne l'accès aux informations et leur divulgation. Des dispositifs très stricts doivent assurer la protection physique des données (accès aux terminaux, aux mots de passe, aux sous-fichiers, et le reste). Parallèlement, des règles et procédures doivent être mises en place pour déterminer l'accréditation des usagers et pour préciser les modalités d'accès et d'utilisation des données. La plupart des contrôles physiques sont standardisés. Mais les règles relatives à l'accréditation des usagers peuvent être très variables, en fonction des contextes.

3.2 Le consentement des personnes

En règle générale, toutes les informations nominatives (et donc personnelles) incluses dans la banque doivent au préalable avoir été autorisées par les personnes concernées. Mais ce principe ne s'applique pas lorsque les données sont publiques (par exemple: les actes de baptême, mariage et sépulture, les données foncières des bureaux d'enregistrement, les manuscrits des recensements canadiens du 19e siècle...).

Le problème du consentement se complique lorsque a) on utilise sans consentement pour fins de recherche des données nominatives non publiques, b) on utilise pour fins de recherches des données publiques qui n'ont pas été constituées et autorisées pour ces fins, c) par le jumelage automatique ou autre, on transforme des données publiques en données confidentielles qui sont ensuite utilisées sans le consentement des personnes concernées, d) on jumelle deux fichiers ou banques de données nominatives pour produire un troisième fichier, d'une autre nature.

3.3 Inférence et ingérence

La construction des généalogies permet de reconstituer les circuits de transmission des gènes délétères et d'estimer (par "inférence généalogique") leur probabilité de diffusion parmi les descendances et les familles. L'utilisation de ces connaissances peut conduire à des atteintes graves à la vie privée, dans laquelle le(la) chercheur(se) pourrait être tenté(e) de s'immiscer avec un souci de prévention.

3.4 Utilisations abusives des fichiers

Construit à des fins de recherches, un fichier peut ultérieurement s'avérer susceptible d'utilisations pour d'autres fins, par des entreprises commerciales ou industrielles (ex.: sélection du personnel), aussi bien que par des ministères gouvernementaux. Ces perspectives font craindre des recyclages, sinon des récupérations ou "perversions" non anticipées, éventuellement préjudiciables aux droits de la personne. Pensons ici au jumelage potentiel d'un fichier avec un corpus d'éventuels clients de compagnies d'assurance.

3.5 Le droit de regard du citoyen

Ainsi l'exploitation d'un fichier de population, comme de toute autre banque de données nominatives, se prête à des formes d'utilisations qui n'avaient pas nécessairement été entrevues au départ et qui peuvent enfreindre les principes éthiques d'une société. Opérant dans le vase clos universitaire, les chercheurs(ses) peuvent éprouver la tentation de prendre, sans concertation ou consultation publique, des décisions scientifiques qui sont en fait des choix sociaux au premier chef, pour lesquels ils (elles) ne sont aucunement mandatés(es). Ceci pose le problème de la responsabilité collective dans la définition des orientations de recherche.

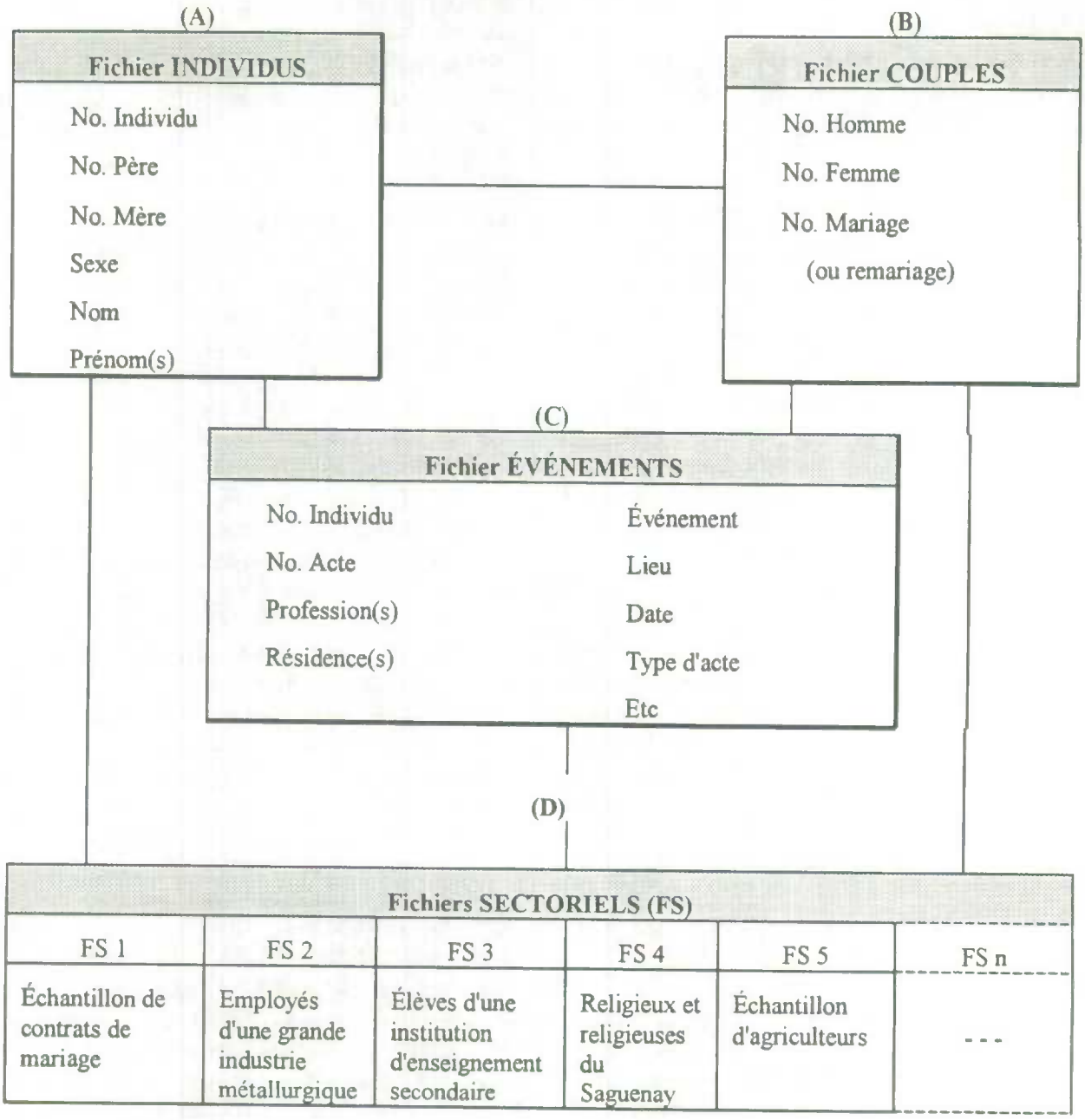
3.6 Les garanties à long terme

La construction d'un fichier de population est l'oeuvre d'une génération de chercheurs(ses) travaillant à la poursuite d'objectifs scientifiques bien définis. Une fois ces objectifs atteints, l'équipe pionnière se dissout mais l'infrastructure scientifique érigée demeure. Alors

---

[5] Sur tout ce qui précède, voir les Documents de l'IREP no. I-C-126, I-C-134, I-C-138. Aussi: les Rapports annuels publiés par l'Institut.

Figure 1
Structure et contenu du fichier de population BALSAC



(IREP)

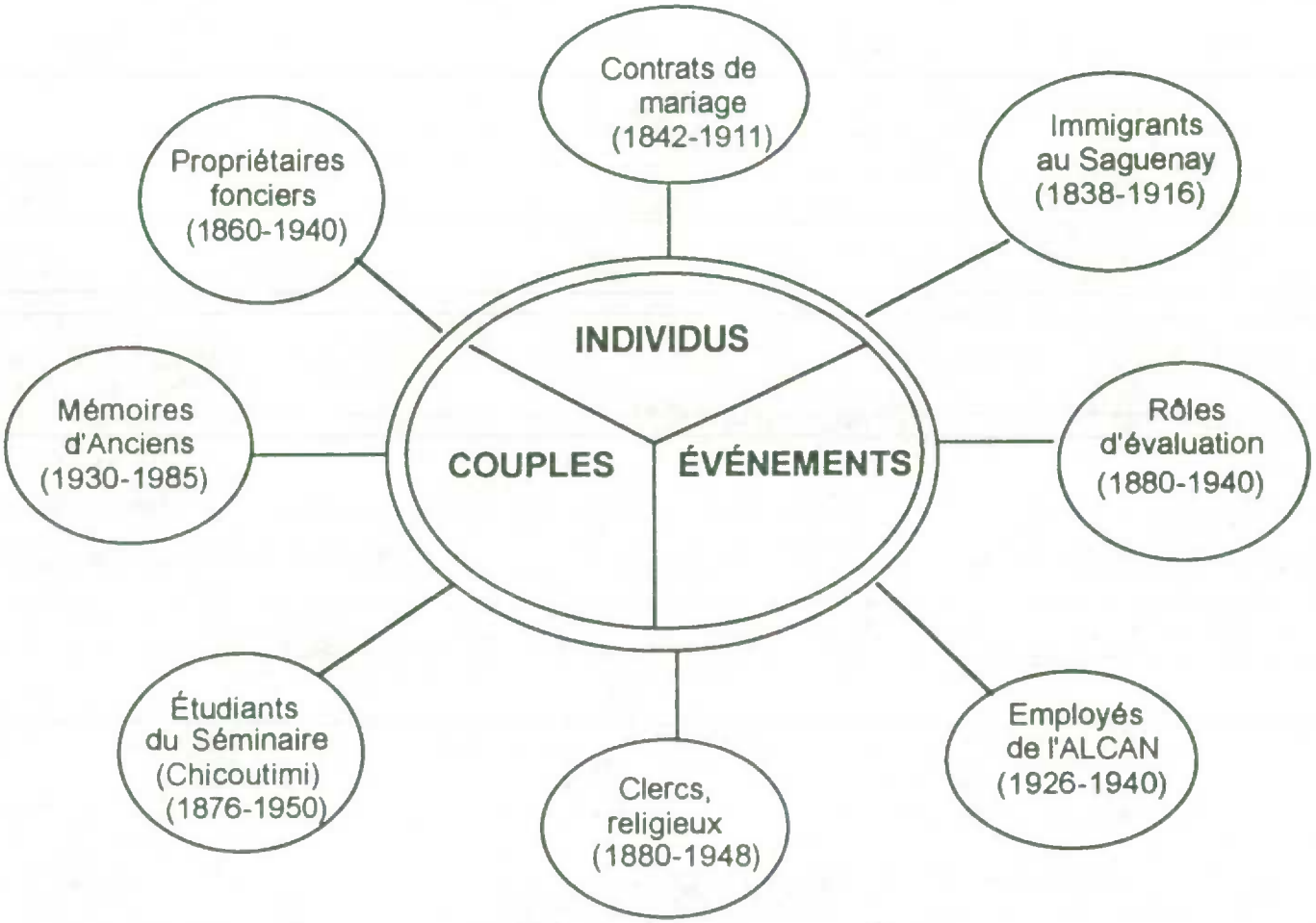
il n'est pas impossible que, les circonstances aidant, celle-ci poursuive sa "carrière" dans un environnement différent, moins sécuritaire, à la merci d'initiatives préjudiciables aux droits des personnes. Toutes les banques de données nominatives sont susceptibles de soulever l'un ou l'autre, sinon chacun des problèmes qui viennent d'être évoqués. Pour cette raison, il est important que la communauté scientifique favorise une prise de conscience générale et contribue à une réflexion qui, du reste, a été amorcée depuis plusieurs années[6].

[6] Rappelons, parmi d'autres, les importantes contributions de notre collègue David Flaherty de l'Université Western (London, Ontario). En ce qui concerne l'IREP, voir en particulier J. Goulet (1992), C. Laberge et B-M Knoppers (1992), B-M Knoppers, C. Laberge et Loïc Cadiet (1992), G. Bouchard (1993).

Figure 2

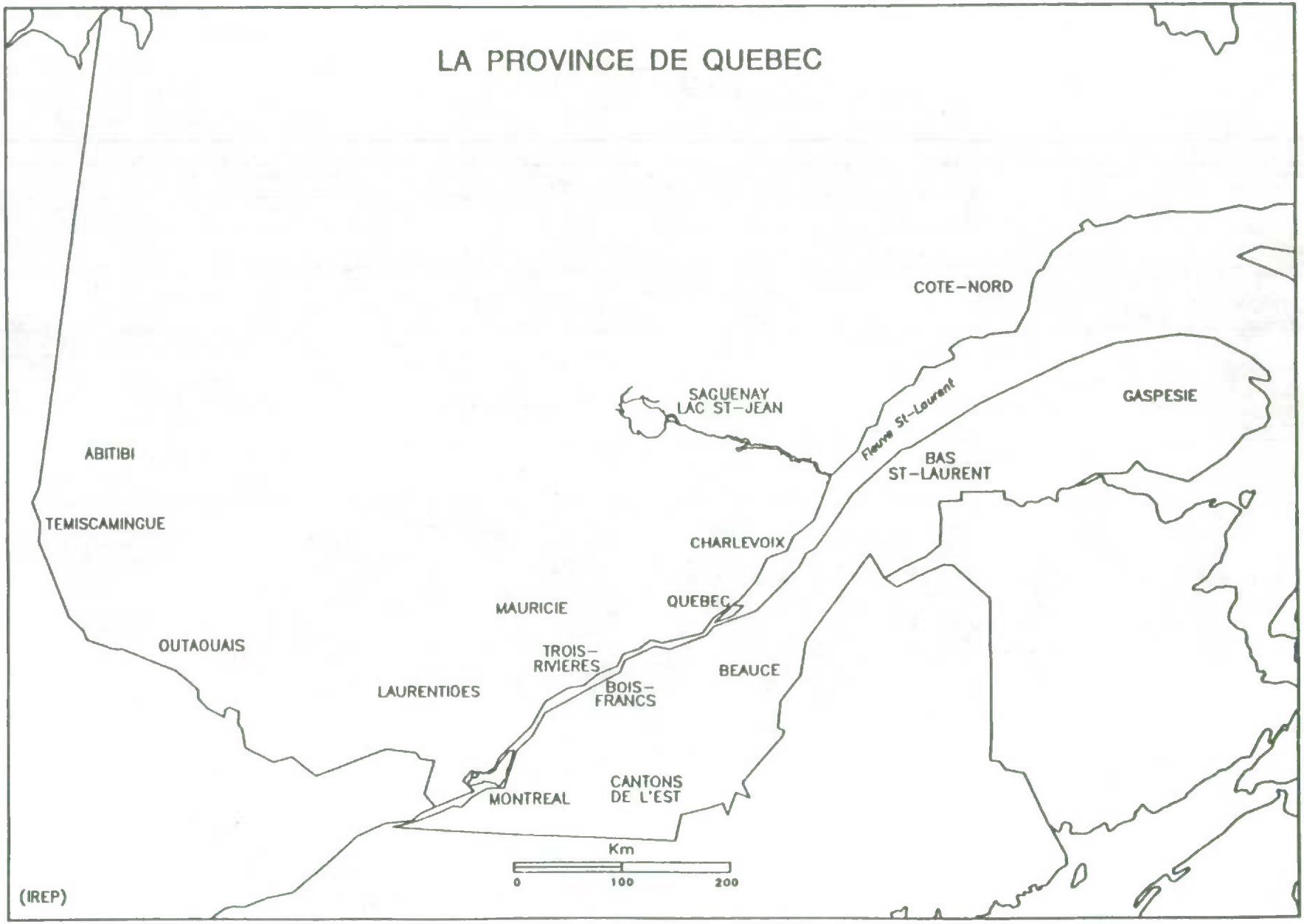
STRUCTURE ET CONTENU DU FICHIER DE POPULATION BALSAC

(Fichier central. Exemples de fichiers sectoriels)



(IREP)

Carte 1


LA PROVINCE DE QUEBEC
COTE-NORD
SAGUENAY
LAC ST-JEAN
Fleuve St-Laurent
GASPESIE
ABITIBI
BAS
ST-LAURENT
TEMISCAMINGUE
CHARLEVOIX
QUEBEC
MAURICIE
OUTAOUAIS
TROIS-
RIVIERES
BEAUCE
LAURENTIDES
BOIS-
FRANCS
CANTONS
DE L'EST
MONTREAL
Km
0
100
200
(IREP)

4. LE PROTOCOLE DE L'IREP

Dans le but d'apporter une solution adéquate à chacune des questions évoquées, l'IREP et l'Université du Québec à Chicoutimi ont mis au point un protocole relativement complexe qui gouverne aujourd'hui la gestion du fichier BALSAC. Nous en rappelons brièvement les grandes lignes. Précisons d'abord que les quatre universités (Université du Québec à Chicoutimi, Universités Laval et McGill, Université de Montréal – donc quatre organismes publics) sont propriétaires et responsables de la gestion du fichier de population. Dès 1977, avec l'aide de juristes, l'IREP[7] et l'Université du Québec à Chicoutimi ont institué un protocole d'accès, de conservation et d'utilisation des données. Ce premier protocole était constitué de directives et de procédures prévoyant aussi bien des mesures à caractère physique ou technique (mots de passe, accès aux locaux; dispositifs d'accès restrictif et de cloisonnement des fichiers informatisés; etc) que des obligations contractuelles faites aux personnes oeuvrant pour l'IREP (ex.: règle de l'anonymat dans les résultats et données publiées ou diffusées)[8]. Sur ce dernier point en particulier, il stipulait l'assermentation du personnel de recherche et des usagers du fichier de même que la signature de divers types de contrats (embauche d'assistants, construction de sous-fichiers, prêt de données...) assortis de pénalités diverses en cas de dérogation aux règles. Il confiait aussi à un comité de dix membres, indépendant de l'IREP, le soin de gérer l'accès aux données et l'application générale du protocole. Enfin, il interdisait l'utilisation du fichier à des fins commerciales.

Durant les années 1980-82, ce cadre d'opération fut réévalué et révisé en profondeur, un mandat en ce sens ayant été donné à une équipe de juristes dirigée par le professeur Jean Goulet, de la Faculté de droit de l'Université Laval. Le nouveau protocole issu de ces travaux reprenait les dispositions précédentes, en les étendant. Il assurait aussi la conformité du nouveau règlement avec les législations toutes récentes du gouvernement canadien et du gouvernement québécois[9]. En vertu de ces amendements, la gestion du fichier en matière d'éthique et de droit relevait de trois instances: un comité universitaire de déontologie opérant à distance de l'IREP, le Conseil d'administration de l'Université du Québec à Chicoutimi (par le biais de son Secrétariat général), la Commission d'accès à l'information du gouvernement du Québec (J. Goulet et coll.., 1983). A partir de ce moment, toute demande d'accès aux données, de création de sous-fichiers et de démarrage de projets de recherche devait être approuvée par ces trois instances[10]. A son tour, ce deuxième protocole a été amendé à quelques reprises durant les années 1980 (IREP, 1989).

En rapport avec la question du consentement, le protocole fait une obligation aux chercheurs(ses) de recueillir ce consentement auprès de toutes les personnes qui fournissent de l'information. Lorsque cette obligation ne peut être respectée (exemple: corpus de données concernant des milliers de personnes dont plusieurs sont décédées), l'IREP s'autorise des dispositions d'exception prévues à cette fin par les législations québécoise et canadienne[11].

Pour ce qui concerne plus spécifiquement l'utilisation du fichier dans le domaine de la génétique humaine, et en particulier aux fins de l'épidémiologie génétique, il a paru justifié d'élaborer un sous-ensemble de règles appropriées au type de problèmes posés sur ce plan. La question de fond peut être posée en ces termes: à partir du moment et dans la mesure où la banque de données, par le biais de l'analyse généalogique ou autrement, peut livrer des connaissances à caractère quasi médical sur des individus (par exemple des connaissances relatives au risque de porter tel ou tel gène délétère), de quelle manière et dans quelles conditions ces connaissances peuvent-elles être utilisées? La politique élaborée est appuyée sur les lignes directrices suivantes. D'abord, il est hors de question de dépister des porteurs de gènes mutants en mettant à contribution le fichier dans le cadre d'opérations dirigées sur des descendances définies comme étant à risque sur la foi de l'inférence

---

7 L'IREP était alors la société de recherches sur les populations (SOREP).

8 Pour le système de protection physique, voir le Document I-C-148 de l'IREP.

9 Par exemple la loi québécoise de 1982 sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels (L.R.Q., C.A-21).

10 Un modus operandi fut établi avec la Commission d'accès dans le but d'alléger le processus d'autorisation.

11 Pour le Québec, voir l'article 19 de la loi sur les services de santé et les services sociaux. Aussi: les articles 59 (paragraphe 5) et 125 de la Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels (L.R.Q., A-21).

généalogique. Le chercheur violerait alors les règles élémentaires du droit et de l'éthique en s'immisçant ainsi dans la vie des individus et des familles sans en avoir été requis expressément par les intéressés eux-mêmes. En deuxième lieu, la diffusion et l'utilisation des connaissances de risque ou de toute information personnelle à caractère médical ou quasi médical tirée du fichier doivent être prises en charge par des instances médicales autorisées, dans le cadre de consultations de conseil génétique entre des professionnels de la santé et des individus qui en font la demande.

Enfin, une demande d'accès au fichier en provenance d'un individu pour autre fin que de recherche scientifique est admissible seulement si les informations visées ne concernent que l'individu demandeur et à la condition que ces informations n'aient pas de connotation médicale ou génétique. En vertu du protocole en vigueur, une telle requête doit alors être acheminée au Secrétariat général de l'Université du Québec à Chicoutimi pour être approuvée. Cependant, il n'est pas donné suite à une demande de ce genre si elle implique des informations à caractère médical ou génétique. L'individu demandeur est alors référé à un service médical compétent.

Théoriquement, le fichier BALSAC pourrait permettre de regrouper toutes les informations relatives à une même personne. Mais en pratique, cette notion de "dossier individuel" n'existe pas. Informatiquement, ce dossier n'est que virtuel car les informations relatives à un même individu sont réparties entre divers sous-fichiers (ou « tables ») qui ont délibérément été structurées de telle manière qu'il est impossible à l'usager de les mettre toutes en relation. Cette opération, qui ne pourrait être réalisée que par les gestionnaires du fichier, nécessiterait des réaménagements informatiques complexes et devrait d'abord avoir recueilli des autorisations externes.

Depuis deux ans, l'IREP est engagé dans une révision en profondeur du protocole de confidentialité, qui subit ainsi sa troisième refonte. L'opération, réalisée encore une fois sous la supervision du juriste Jean Goulet, doit être complétée au cours de l'année 1996. Le nouveau protocole reprend les principes et les principales règles de l'ancien, en les adaptant et en y apportant divers ajouts et précisions. Les lignes directrices qui sous-tendent le protocole sont énoncées au Tableau 1. La structure du système est illustrée à la Figure 3.

Un dernier point mérite attention, celui des droits collectifs. Nous nous référons ici à l'obligation des chercheurs à protéger l'image des collectivités sur lesquelles ils travaillent. Cette question est particulièrement délicate lorsqu'il s'agit de recherches portant sur des maladies génétiques. En l'absence de précautions appropriées entourant les modalités de diffusion des résultats, il est dangereux de donner naissance à des stéréotypes péjoratifs qui restent ensuite, et pour longtemps, associés aux populations en cause. La prise de conscience de ce problème a amené l'IREP à prendre l'initiative d'une sorte de code d'éthique qui a été proposé aux membres de la communauté scientifique et aux divers groupes d'intervenants en matière de maladies héréditaires; il a aussi été porté à l'attention des professionnels des médias[12].

5. CONCLUSION

Comme règle générale, le protocole de l'IREP favorise la transparence des opérations et des décisions de recherche. Régulièrement, l'IREP intervient dans les médias pour faire connaître les développements importants qui sont projetés. Régulièrement aussi, la politique adoptée en matière de confidentialité fait l'objet de présentations et de discussions dans des colloques ou séminaires ouverts au public et réunissant des spécialistes du droit, de l'éthique et de diverses disciplines. Ainsi, la structure multidimensionnelle de décision et de concertation mise en place à partir de l'institution universitaire offre une protection contre les possibilités de commercialisation et de "perversion" du fichier. Appuyé sur des institutions à caractère public, le fichier est raisonnablement à l'abri des accidents de parcours qui pourraient compromettre son avenir à long terme. Nous pensons aussi que le protocole offre une protection adéquate sur chacun des six points énoncés plus haut (partie III). A l'appui de cet énoncé, on peut mentionner le fait qu'en une vingtaine d'années, l'exploitation du fichier n'a donné lieu à aucun incident entraînant préjudice à des personnes ou violation de la vie privée, et elle n'a suscité aucune plainte de la part d'individus se sentant lésés. Il importe de rappeler cependant que les situations créées par les banques de données nominatives sont en constante évolution à cause des développements techniques continus, à cause des transformations de l'environnement juridique et législatif et, enfin, à cause d'une évolution dans les sensibilités collectives en faveur du respect des droits de la personne. Il va sans dire que la communauté scientifique doit appuyer sans réserve ce dernier courant et se montrer disposée à amender continuellement ses

[12] Voir à ce sujet G. Bouchard (1994) et Documents de l'IREP no. III-C-94.

pratiques scientifiques pour les rendre conformes aux principes fondamentaux de la morale collective. Pour toutes ces raisons, il faut bien reconnaître que les protocoles de protection de la vie privée sont toujours provisoires et doivent être régulièrement soumis à des révisions, dans un esprit de réflexion aussi ouverte que possible.

Tableau 1

SYSTÈME DE PROTECTION DE LA CONFIDENTIALITÉ RÉGISSANT L'EXPLOITATION DU FICHIER BALSAC (CARACTÉRISTIQUES PRINCIPALES)

1 - CONTENU DU FICHIER : DONNÉES À CARACTÈRE PUBLIC

2 - PROPRIÉTÉ DU FICHIER : QUATRE UNIVERSITÉS TITULAIRES

3 - PAS D'INGÉRENCE DANS LA VIE PRIVÉE

4 - AUCUNE UTILISATION DU FICHIER À DES FINS DE PROFIT POUR SES PROPRIÉTAIRES OU POUR SES GESTIONNAIRES (RECHERCHE SCIENTIFIQUE SEULEMENT, APPROUVÉE PAR UN COMITÉ D'ÉTHIQUE)

5 - MÉCANISMES EXTERNES D'AUTORISATION D'ACCÈS

- COMITÉ UNIVERSITAIRE DE DÉONTOLOGIE
- DIRECTION DE L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI
- COMMISSION D'ACCÈS À L'INFORMATION DU QUÉBEC

6 - LE FICHIER NE CONSERVE PAS DE DONNÉES MÉDICALES (a)

7 - LOGICIELS D'ACCÈS RESTRICTIF (MÉCANISME DE CLOISONNEMENT DES DONNÉES)

8 - LA NOTION DE "FICHIER INDIVIDUEL" N'EXISTE PAS CONCRÈTEMENT (FRAGMENTATION DES DONNÉES)

9 - DIFFUSION DES RÉSULTATS DE RECHERCHE: RÈGLE DE L'ANONYMAT

10 - CONTRATS D'ASSERMENTATION DES USAGERS ET EMPLOYÉS, ASSORTIS DE PÉNALITÉS

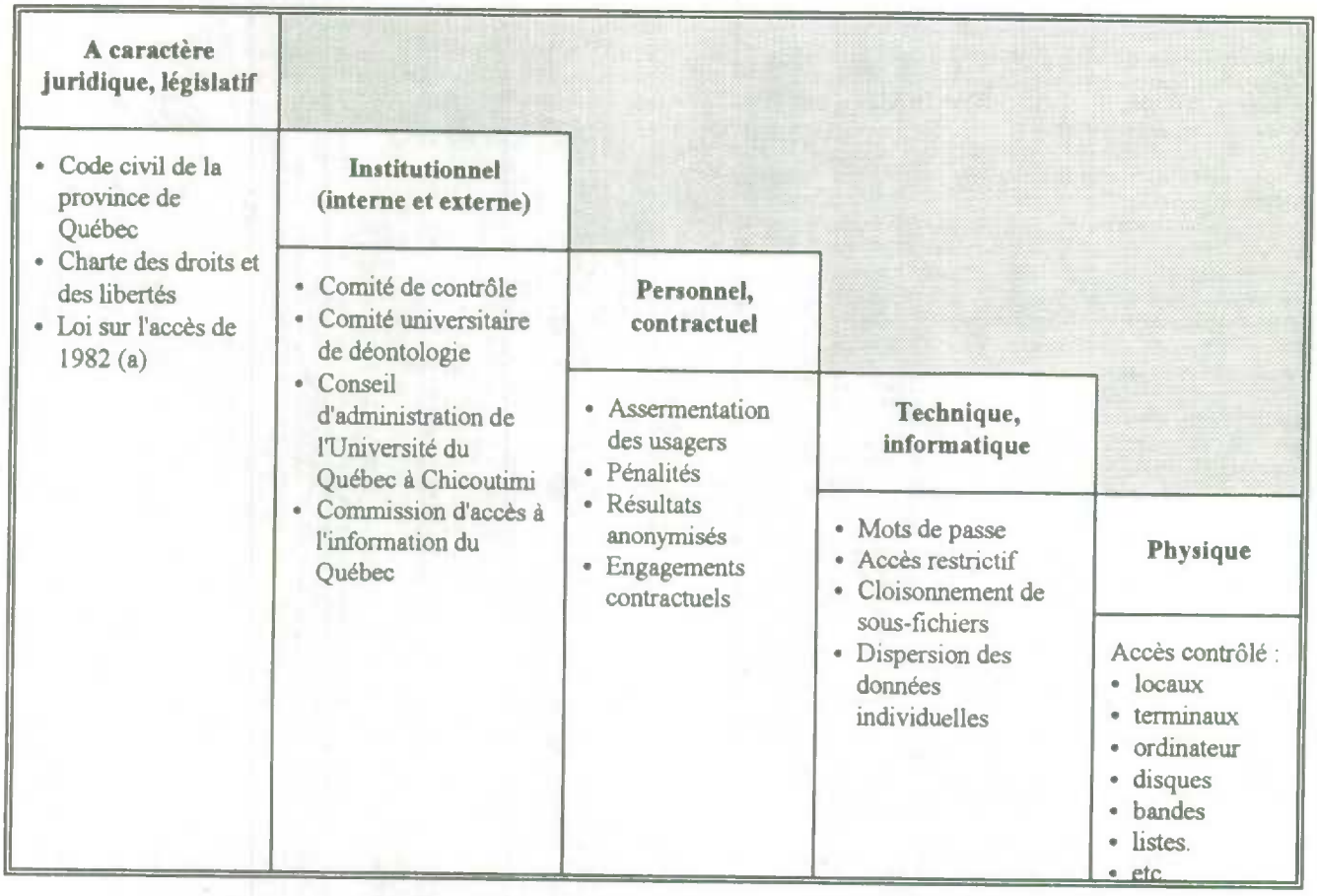
11 - SURVEILLANCE CONSTANTE DES OPÉRATIONS PAR UN COMITÉ DE CONTRÔLE (25 - 30 RÉUNIONS PAR AN)

12 - SOUCI DE PROTÉGER L'IMAGE DES COLLECTIVITÉS VISÉES, EN PLUS DES RÉPUTATIONS INDIVIDUELLES

(a) Voir le Document no. I-C-153 de l'IREP

Figure 3

Structure du protocole de confidentialité.
Types de protections et de contrôles régissant l'accès et l'exploitation du fichier-réseau BALSAC



(a) Loi sur l'accès aux documents des organismes publics et sur la protection des renseignements personnels (L.R.Q., c. A-2.1). Ayant le statut d'organismes publics, quatre Universités sont co-propriétaires du fichier, dont la gestion relève de cette loi.

BIBLIOGRAPHIE

Bouchard G. (1988). Les fichiers-réseaux de population: Un retour à l'individualité, *Histoire sociale/Social History*, XXI, 42, 287-294.

Bouchard G. (1992). Current issues and new prospects for computerized record linkage in the province of Québec, *Historical Methods*, 25, 2, 67-73.

Bouchard G. (1992). *Le Centre interuniversitaire SOREP et le fichier BALSAC, État présent et planification des travaux.*

Bouchard G. (1993). Retracer les gènes dans la population: une infrastructure de recherche pour le 21e siècle, *Mémoires de la Société royale du Canada*, Sixième série, tome IV, 13-24.

Bouchard G. (1994). Les problèmes de droit et d'éthique reliés à l'exploitation d'un fichier de population à des fins génétiques, Marcel J. Mélançon (dir.), *Bioéthique et génétique, Une réflexion collective*. Chicoutimi, Éditions JCL, 33-42.

Bouchard G., Roy R., et Casgrain B. (1985). *Reconstitution automatique des familles.*

Le système SOREP. Dossier no. 2, Université du Québec à Chicoutimi, 2, 745.

Bouchard G., Roy R., Casgrain B., et Hubert M. (1989). Fichier de population et structures de gestion de base de données: le fichier-réseau BALSAC et le système INGRES/INGRID, *Histoire & Mesure*, IV, 1/2, 39-57.

Goulet J. (1992). La législation sur la protection de la vie privée: les principes fondamentaux des lois de première génération, *Les archives non textuelles: réflexions théoriques et expériences pratiques*, Actes du Colloque organisé conjointement par la Division des archives et Programme d'archivistique de l'Université Laval, 20 novembre 1991, Québec: Université Laval, 103-119.

Goulet J., Gagné M., et Girard D. (1983). *Règles de droit et confidentialité*, Dossier no. 1, 175.

IREP (1989). *Règlement concernant la confidentialité des données contenues dans le fichier BALSAC*, Mars, 40.

Knoppers B.-M., Laberge C., et Cadiet L. (dirs.) (1992). *La génétique de l'information à l'informatisation*, Actes du colloque ayant eu lieu à la faculté de droit de l'Université de Montréal, organisé par le C.R.D.P. de l'Université de Montréal et le C.R.J.O. de l'Université de Rennes, Paris: Litec, 387.

Laberge C., et Knoppers B.-M. (dirs.) (1992). *Registres et fichiers génétiques: enjeux scientifiques et normatifs*. Montréal: Association canadienne-française pour l'avancement des sciences (ACFAS), Collection Les cahiers scientifiques, 77, 178.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

ASPECTS JURIDIQUES ET POLITIQUES DE LA CONFIDENTIALITÉ ET DE LA PROTECTION DES RENSEIGNEMENTS PERSONNELS

L. Desramaux[1]

RÉSUMÉ

L'auteur décrit le cadre légal et politique qui régit les activités de Statistique Canada en ce qui a trait à la collecte, à l'utilisation et à la divulgation des renseignements, en mettant l'accent de façon particulière sur les renseignements personnels. Elle met l'accent sur les fondements juridiques de ce cadre, qui sont énoncés dans la *Loi sur la statistique*, la *Loi sur la protection des renseignements personnels* et la *Loi sur l'accès à l'information*. Enfin, elle décrit les politiques, comme la Politique relative au couplage d'enregistrements, la Politique d'information des répondants aux enquêtes et la Politique régissant la diffusion des microdonnées, mises au point par le Bureau pour satisfaire aux exigences des dispositions légales.

MOTS CLÉS : *Loi sur la statistique; Loi sur la protection des renseignements personnels ; Loi sur l'accès à l'information*; Politique d'information des répondants aux enquêtes; Politique régissant la diffusion des microdonnées.

1. INTRODUCTION

La présentation qui suit a pour objet de fournir un aperçu du cadre légal et politique que Statistique Canada a adopté pour s'assurer que le Bureau s'acquitte de façon appropriée de ses responsabilités en ce qui a trait à la collecte, à la protection et à la divulgation des données.

Statistique Canada a deux responsabilités fondamentales. L'une, d'informer, au meilleur de ses capacités, la population et l'autre, d'assurer le secret des renseignements personnels qui lui sont fournis.

Ces deux responsabilités sont intimement liées. Pour être en mesure de répondre aux demandes d'information, il faut absolument disposer d'une source stable de renseignements, et celle-ci ne peut être assurée que si l'engagement aux répondants de protéger les renseignements qu'ils ont fournis est respecté.

Non seulement ces responsabilités sont-elles inextricables, mais elles peuvent parfois être en conflit. La société a besoin de plus de renseignements pour comprendre des problèmes de plus en plus complexes et, pour ce faire, elle a besoin de renseignements de plus en plus détaillés. Par ailleurs, il est à noter qu'il y a un changement d'attitude de la part du public en ce qui concerne la collecte d'information. Les individus semblent plus préoccupés par l'accumulation d'information recueillie à leur sujet et l'utilisation qui en est faite. Ils sont aussi inquiets du pouvoir que possède la technologie qui permet entre autres, la possibilité de jumeler des informations provenant de plusieurs sources et de créer ainsi des banques massives de données sur des personnes. En outre, le public est de plus en plus au fait des droits que lui accordent les lois fédérale et provinciales portant sur la protection des renseignements personnels.

Pour répondre de façon responsable et constante aux préoccupations du public en ce qui a trait à l'accumulation croissante de données, particulièrement celles de nature personnelle, ainsi que pour justifier les pressions en vue d'obtenir des renseignements plus détaillés, Statistique Canada a mis en place un cadre légal et politique qui met l'accent exclusivement sur les questions de confidentialité, de protection des renseignements personnels et de sécurité.

Ce cadre a pour fondement juridique la *Loi sur la statistique*, la *Loi sur la protection des renseignements*

[1] Louise Desramaux, Services d'accès aux données et de contrôle, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6.

personnels et la *Loi sur l'accès à l'information*, auxquelles viennent s'ajouter un certain nombre de politiques et de procédures bien définies, qui favorisent le respect des exigences légales.

2. LOI SUR LA STATISTIQUE

La pierre angulaire du cadre légal et politique de Statistique Canada est la *Loi sur la statistique*. Cette loi a été adoptée pour la première fois en 1918. Elle a été modifiée considérablement au fil des ans, la dernière révision importante ayant été effectuée en 1971. Elle énonce le mandat de Statistique Canada qui consiste à :

- recueillir, compiler et publier des renseignements;
- recenser la population et faire le recensement agricole;
- collaborer avec les ministères;
- favoriser et mettre au point des statistiques sociales et économiques;
- veiller à prévenir le double emploi avec les ministères.

Pour mener à bien un tel mandat, trois exigences sont nécessaires au niveau des dispositions légales : le pouvoir de recueillir des données, l'obligation de la part du répondant de fournir des données, et la protection de la confidentialité des données fournies.

2.1 Pouvoir de recueillir des données

La *Loi sur la statistique* donne au statisticien en chef des pouvoirs étendus de collecte de données, dans une gamme variée de domaines. Elle prévoit la collecte des données, directement auprès des répondants, ainsi que l'accès aux dossiers administratifs détenus par les ministères, les administrations fédérales, provinciales et municipales, et les sociétés.

2.2 Obligation de répondre

Depuis sa proclamation en 1918, la *Loi sur la statistique* stipule que la réponse aux enquêtes de Statistique Canada est obligatoire. Au cours des années 70, les données requises par le gouvernement et les chercheurs pour donner suite aux questions sociales qui se posaient à ce moment-là étaient perçues par les répondants comme une intrusion dans leur vie privée. Certaines des enquêtes d'alors sont devenues plus complexes et, dans certains cas, nécessitaient de longues réponses. Le Bureau était pleinement conscient que, du fait de la nature plus délicate de certaines de ses enquêtes statistiques, il devait permettre la participation volontaire aux enquêtes. Par conséquent, en 1981, la *Loi sur la statistique* a été modifiée pour permettre au Ministre d'autoriser par décret la tenue d'une enquête avec réponse volontaire. Toutefois, en vertu de ces dispositions, le Ministre n'est pas habilité à faire du recensement de la population et du recensement agricole des enquêtes volontaires.

2.3 Protection du secret

Pour contrebalancer les vastes pouvoirs en matière de collecte et l'obligation, dans certains cas, pour les répondants de fournir des renseignements, la *Loi sur la statistique* comprend des dispositions très strictes concernant la confidentialité. L'article de la Loi qui porte sur le secret comprend deux paragraphes. Le premier traite de l'accès à des relevés identifiables faits en vertu de la *Loi sur la statistique* et prévoit que nul ne peut être autorisé à prendre connaissance de renseignements personnels, sauf les personnes qui ont été assermentées en vertu de la loi.

Le deuxième paragraphe porte sur les renseignements qui peuvent être divulgués. Le libellé est plutôt maladroit, mais il importe de retenir que les dispositions sont très contraignantes. En bref, la loi prévoit que les renseignements divulgués ne doivent pas pouvoir être rattachés à un répondant en particulier. Cela a donné lieu à la mise au point d'un certain nombre de méthodes de contrôle de la divulgation pour tenir compte des aspects pratiques de la garantie de confidentialité. La divulgation de renseignements personnels n'est permise que depuis la modification de la *Loi sur la statistique* en 1971. Auparavant, Statistique Canada ne pouvait diffuser que des données agrégées. On a toutefois fait exception à cette règle, même en 1918. Par exemple, il existait un certain niveau de discrétion quant à la divulgation des renseignements sur le transport des personnes. Au fil des ans, les exceptions ont été élargies pour inclure les renseignements ayant trait à un transporteur ou à une entreprise d'utilité publique, aux renseignements ayant trait à des établissements comme les hôpitaux, les établissements pour malades mentaux, les bibliothèques et les établissements d'enseignement, à condition que les malades, pensionnaires ou autres personnes dont s'occupent ces établissements ne puissent pas être identifiés. Il existait en outre un certain niveau de discrétion en ce qui a trait à la divulgation des renseignements fournis par des répondants, qu'il s'agisse d'entreprises ou de personnes, avec l'autorisation écrite de ceux-ci, et pour la divulgation des listes d'entreprises qui pourraient contenir des noms et adresses d'entreprises, des genres d'entreprises ou la taille de celles-ci exprimées en nombre d'employés. Ce pouvoir discrétionnaire peut être exercé uniquement par le

statisticien en chef, exige la publication d'un arrêté par ce dernier et est assujetti à une politique interne.

Même si ces dispositions donnent une certaine souplesse quant à la divulgation des renseignements, laquelle serait inexistante autrement, les dispositions relatives à la confidentialité sont, comme il a été mentionné précédemment, très rigides. On note de plus en plus d'insatisfaction, particulièrement en ce qui a trait aux données des entreprises qui, dans nombre de cas, sont protégées et qui, selon des normes raisonnables, ne semblent pas devoir être protégées de la divulgation, du fait qu'elles sont considérées comme d'intérêt public ou qu'elles ne sont pas de nature délicate.

La règle de protection du secret comporte une autre exception en ce qui a trait au partage des données. En 1971, le mandat de Statistique Canada a été élargi pour inclure la prévention du double emploi dans la collecte des renseignements par les ministères. À cette fin, on applique les dispositions de la loi concernant les accords sur le partage des données avec les ministères, les municipalités et d'autres sociétés. Lorsqu'un accord sur le partage des données est conclu, les dispositions légales prévoient que les répondants soient informés de l'accord et qu'ils puissent s'objecter au partage des renseignements les concernant.

3. LOI SUR L'ACCÈS À L'INFORMATION

La *Loi sur l'accès à l'information* donne au public le droit d'accès aux documents que détiennent les institutions du gouvernement fédéral tout en tenant compte de certaines exemptions, qui sont toutefois limitées et précises. Une de ces exemptions empêche de façon expresse, l'accès par des tiers à l'information recueillie en vertu de la *Loi sur la statistique* qui pourrait permettre d'identifier un répondant. De plus, tous les renseignements que Statistique Canada met ou peut mettre à la disposition du public sont exclus de la loi.

4. LOI SUR LA PROTECTION DES RENSEIGNEMENTS PERSONNELS

Enfin, une troisième loi, non moins importante, régit les activités de Statistique Canada, soit la *Loi sur la protection des renseignements personnels*. Elle comprend des exigences explicites quant à la collecte des renseignements personnels. De façon plus particulière,

- les institutions gouvernementales ne peuvent recueillir des renseignements personnels que lorsqu'ils ont un lien direct avec leurs programmes ou leurs activités;
- les personnes auprès de qui on recueille des renseignements personnels doivent être informées des fins auxquelles ils sont destinés;
- les institutions gouvernementales peuvent utiliser les renseignements personnels uniquement aux fins auxquelles ils ont été recueillis ou préparés par les institutions ou pour les usages qui sont compatibles avec ces fins, à moins que l'individu ne donne son consentement pour d'autres usages.

La *Loi sur la protection des renseignements personnels* autorise en outre les institutions gouvernementales qui y sont assujetties à divulguer des renseignements personnels sans le consentement des personnes, si une autre loi, comme la *Loi sur la statistique*, l'autorise. Cela nous permet un accès continu aux dossiers administratifs contenant des renseignements personnels. Ces dispositions figurent aussi dans les lois provinciales sur la protection des renseignements personnels.

5. CORRÉLATION

Du point de vue de Statistique Canada, les trois lois sont bien agencées. En ce qui a trait à la collecte, étant donné que les données sont recueillies à des fins statistiques uniquement, et que les renseignements personnels et d'autres renseignements permettant d'identifier des personnes sont protégés de l'accès et de la divulgation, le respect de la *Loi sur la protection des renseignements personnels* n'a pas nécessité d'ajustements majeurs.

Dans le cadre de ses responsabilités en matière de divulgation, l'Agence respecte pleinement les objectifs de base de la *Loi sur l'accès à l'information* et de la *Loi sur la protection des renseignements personnels*, c'est-à-dire qu'il assure le maximum de disponibilité des données statistiques, à des niveaux de détail qui ne permettent pas d'identifier les répondants à partir des renseignements qu'ils ont fournis.

Des deux lois, la *Loi sur la protection des renseignements personnels* est celle qui a le plus de répercussions sur la façon dont nous fonctionnons. En fait, deux des politiques internes qui font partie de notre cadre légal et politique ont été élaborées pour donner suite de façon particulière aux problèmes de protection des renseignements personnels.

6. POLITIQUE RELATIVE AU COUPLAGE D'ENREGISTREMENTS

Une politique ministérielle visant le couplage de données a été élaborée, il y a presque 10 ans maintenant, en réponse aux soucis exprimés par le public et le Commissaire fédéral à la vie privée concernant la possibilité, devenue réelle grâce à l'évolution de la technologie, du couplage de renseignements personnels qui pourraient provenir d'une multitude de sources, et ceci sans que les individus concernés en soient conscients.

Bien que les soucis exprimés portaient surtout sur l'utilisation des renseignements jumelés à des fins administratives ou de réglementation, Statistique Canada, fort consciente de l'importance du couplage de données pour un bureau statistique, a voulu assurer que son utilisation pour des fins statistiques et de recherche ne soient pas indûment restreinte ou pire, interdite.

La politique de SC en matière de couplage d'enregistrements ne permet cette activité que si elle satisfait à de nombreuses conditions, notamment:

- le couplage doit servir à des fins de statistique ou de recherche et doit être conforme au mandat de SC;
- la diffusion des produits du couplage doit satisfaire aux dispositions relatives au secret de la *Loi sur la statistique*;
- le couplage ne doit pas se faire au détriment des répondants concernés, et il doit être évident que les avantages qui en découlent servent l'intérêt public;
- le couplage est conforme à un processus prescrit de révision et d'approbation.

Le processus d'examen comporte des paliers multiples. Le premier est un comité de gestion supérieur qui analyse toutes les propositions de couplage afin de formuler des recommandations à l'intention du Comité des politiques, lequel est composé du statisticien en chef et des statisticiens en chef adjoints. Si le Comité des politiques appuie la recommandation du comité subalterne, le statisticien en chef soumet la proposition au ministre responsable de Statistique Canada. Soit dit en passant, selon la sensibilité du couplage, il est possible que le statisticien en chef décide d'élargir le processus d'examen pour y inclure une consultation avec des organismes externes.

7. POLITIQUE D'INFORMATION DES RÉPONDANTS AUX ENQUÊTES

Une autre politique, à savoir la *Politique sur l'information des répondants aux enquêtes*, a pour objectif principal d'informer les répondants du pourquoi d'une enquête particulière. Ainsi, Statistique Canada se conforme aux exigences de la *Loi sur la protection des renseignements personnels* lorsque l' Agence recueille des renseignements de ce type, et favorise en outre la collaboration des répondants, étant donné que lorsque ceux-ci sont bien informés, ils sont plus enclins à collaborer, notamment lorsqu'ils comprennent l'objectif visé par la collecte de certaines données. Ils sont en outre plus susceptibles de fournir des données de qualité.

Tous les instruments d'enquête utilisés par SC dans ses activités de collecte sont passés en revue au niveau central. Dans le cadre du processus d'examen, on évalue l'énoncé des utilisations et des objectifs des enquêtes, ainsi que les déclarations garantissant aux répondants la protection de la confidentialité. L'examen permet en outre de s'assurer que toutes les exigences juridiques ont été respectées, que les autorisations voulues pour la tenue des enquêtes ont été obtenues et que, dans le cas des enquêtes volontaires, un règlement d'exécution a été obtenu.

Les enquêtes qui sont soumises à un examen sont également évaluées dans la perspective de la protection des renseignements personnels, c'est-à-dire que si des identificateurs sont conservés après la collecte, une banque de renseignements personnels doit être créée et enregistrée. Cette mesure est conforme aux dispositions de la *Loi sur la protection des renseignements personnels*, qui exige l'identification des banques de données personnelles et la description de ces banques dans un répertoire des données personnelles publié par le Conseil du Trésor.

Dans les cas où la collecte de données soulève des questions de protection des renseignements personnels, une réunion est prévue avec les représentants du bureau du Commissaire à la protection de la vie privée. Cette réunion ne vise pas à obtenir l'approbation des personnes compétentes, étant donné qu'il est entendu qu'elles doivent conserver leur impartialité, mais plutôt à les informer et à obtenir leur point de vue quant au respect par SC de tous les aspects des dispositions de la *Loi sur la protection des renseignements personnels*.

8. POLITIQUE RÉGISSANT LA DIFFUSION DES MICRODONNÉES

La mise en oeuvre de cette politique sur la diffusion des microdonnées a été confiée au Comité de la diffusion des microdonnées, qui est probablement le comité le plus ancien de SC en matière de confidentialité des données. Ce comité a été créé au début des années 70,

au moment de la révision de la *Loi sur la statistique*. L'un des changements adoptés permettaient la divulgation de données personnelles anonymes.

En vertu de la politique actuelle, la diffusion des microdonnées est autorisée uniquement si elle améliore considérablement la valeur analytique des données recueillies, et lorsque le l' Agence est convaincue que toutes les mesures raisonnables ont été prises pour prévenir l'identification d'unités d'enquête particulières. Afin de réduire au minimum les risques de divulgation, la politique prévoit l'examen par un groupe d'experts, de tous les fichiers de microdonnées que l'on entend divulguer, à partir de critères établis et en faisant preuve de jugement pour l'évaluation des fichiers.

Des fichiers épurés de microdonnées sur les ménages sont publiés depuis quelque temps déjà. Des pressions de plus en plus grandes s'exercent pour produire ce genre d'extrants, et pour y inclure davantage de détails. Parallèlement, les besoins des usagers deviennent de plus en plus complexes, et l'accès à du matériel et des logiciels plus puissants facilite le couplage des fichiers avec d'autres fichiers d'enquête et fichiers administratifs. Ce problème est encore aggravé par l'avènement de nouvelles enquêtes longitudinales à Statistique Canada. En principe, le risque d'identification des répondants à partir des données longitudinales accroît le risque potentiel de divulgation. Toutefois, les microdonnées sont la seule forme d'extrants qui peut pleinement tirer parti du potentiel des enquêtes longitudinales, enquêtes qui ont été justifiées du fait que leur potentiel analytique pouvait être utilisé par une vaste gamme d'usagers.

Cette question suscite de nombreuses préoccupations et fait actuellement l'objet de discussions par les membres du Comité de la diffusion des microdonnées et du Comité de la confidentialité et des mesures législatives. Elle fait en outre l'objet de recherches méthodologiques actives.

9. POLITIQUES RELATIVES À LA SÉCURITÉ

L'objectif principal des politiques de SC en matière de sécurité concerne la protection de toutes les données de nature délicate, y compris les renseignements informatisés, pour éviter que des personnes non autorisées y aient accès.

Il existe une préoccupation de plus en plus grande en ce qui a trait à la capacité réelle de protéger les renseignements informatisés. Les reportages dans les médias au sujet de personnes qui ont utilisé à mauvais escient des systèmes informatiques supposément bien protégés ont rendu le grand public plus sceptique quant aux garanties de sécurité qui lui sont fournies.

Nos politiques informatiques prévoient essentiellement que le traitement, l'entreposage et la transmission des données statistiques de nature délicate doivent se faire uniquement au moyen d'un réseau auquel le grand public n'a pas accès.

Nos politiques en matière de sécurité prévoient en outre que les données statistiques de nature délicate soient contrôlées en tout temps. Seuls les employés autorisés ont accès à ces données, une piste de vérification de la circulation de celles-ci est obligatoire à SC, et les microdonnées ne peuvent sortir des lieux de travail à moins d'une dispense spéciale du statisticien en chef.

10. CONCLUSION

Il semble que le cadre légal et politique ait été un outil efficace. Il existe un rapport très net entre la capacité de SC d'assurer la confidentialité des données recueillies, de respecter les droits concernant la protection des renseignements personnels des répondants, ainsi que de fournir des mesures de sécurité appropriées et le fait que les membres du grand public fournissent volontiers des renseignements à l'Agence.

Ce fait a été prouvé, au fil des ans, par les taux de réponses élevés à nos enquêtes. Cette réalité ressort en outre davantage des résultats d'une enquête sur la protection des renseignements personnels entreprise il y a trois ans. Même si 92 % des répondants ont indiqué être à tout le moins modérément préoccupés par la protection des renseignements personnels, seulement 14 % ont indiqué avoir des réticences à fournir ce type de renseignements à Statistique Canada.

Nous savons que les préoccupations concernant la protection des renseignements continueront de croître au fur et à mesure que s'élargiront les activités de collecte, d'entreposage et d'utilisation des renseignements personnels. Afin de continuer à pouvoir disposer des renseignements dont la société a besoin pour résoudre les problèmes économiques et sociaux, le climat de confiance qui existe actuellement entre SC et ses répondants doit être maintenu. Cela constitue non seulement un défi, mais aussi une responsabilité.

SESSION 4

Qualité des données statistiques

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

L'INCERTITUDE ET L'ERREUR DANS LES RECENSEMENTS ET LES ENQUÊTES : UNE QUESTION SÉRIEUSE

S.E. Fienberg[1]

RÉSUMÉ

Les bureaux de statistique gouvernementaux diffusent une somme abondante de microdonnées venant des recensements et des enquêtes, mais les utilisateurs de ces données s'interrogent souvent sur la manière d'utiliser les «poids d'échantillonnage» et les autres renseignements sur l'incertitude et l'erreur fournis par ces organismes. Les poids d'échantillonnage ne constituent qu'un aspect de l'incertitude et de l'erreur, et leur rôle dans l'analyse varie avec la perspective de l'utilisateur. En règle générale, les autres erreurs que l'erreur d'échantillonnage sont plus préoccupantes. Pourtant, elles occupent une place restreinte dans les rapports des organismes et dans les modèles des utilisateurs. Enfin, les erreurs introduites par les bureaux de statistique afin de protéger la confidentialité ajoutent à l'incertitude. Une façon d'amener organismes et utilisateurs à prendre ces différentes sources d'incertitude et d'erreur au sérieux consiste à examiner d'une façon novatrice la divulgation des microdonnées. La présente communication propose une nouvelle approche intégrée à la divulgation des microdonnées et à la déclaration de l'incertitude et de l'erreur, conforme à la pratique contemporaine de la méthodologie en statistique. Cette approche amènera les organismes et les utilisateurs à prendre au sérieux les erreurs et l'incertitude dans les données du recensement et des enquêtes.

MOTS CLÉS : Bootstrap; confidentialité; tableaux de contingence; fonction de distribution cumulative; protection du secret statistique; modèles logarithmiques linéaires; imputation multiple; modèles de régression.

1. INTRODUCTION

1.1 Buts

Lorsqu'on passe des données brutes du recensement et des enquêtes à l'information, thème de ce symposium, il est souvent question de la nécessité de séparer le signal et le bruit. Comprendre les implications de l'incertitude et de l'erreur joue un rôle essentiel dans un tel exercice, et les statisticiens qui s'occupent des enquêtes ont rassemblé un jeu d'outils qui leur permet d'étudier l'erreur et la variabilité, et d'utiliser les renseignements recueillis pour établir les limites de l'inférence sur divers phénomènes sociaux sous-jacents. Cet exercice gagne cependant considérablement en complexité quand le bureau de statistique s'efforce de partager l'information par la diffusion d'un ou de plusieurs produits de données susceptibles d'aider un large éventail d'utilisateurs à analyser les données qui sont rendues publiques et à en tirer un enseignement quelconque. Un des principaux arguments soulevés ici est qu'un usage généralisé mais intégré de la méthodologie statistique peut faciliter l'acquisition de la sagesse désirée.

Les organismes de statistique, qui produisent les données du recensement et des enquêtes, déplorent souvent le fait que l'utilisateur ne prête pas attention à l'information sur l'incertitude et l'erreur qu'ils lui communiquent régulièrement. De son côté, l'utilisateur soutient que les bureaux de statistique ne tiennent aucun compte de ses objectifs et de ses préoccupations en matière d'analyse. Amener l'utilisateur à prendre la variabilité au sérieux, d'où le titre de cet article, revient à combler ce fossé. Le principal argument avancé ici est que pour convaincre l'utilisateur à envisager sérieusement l'incertitude et l'erreur, il est essentiel d'adopter une nouvelle approche intégrée à la question de l'incertitude et de l'erreur dans les enquêtes, de la collecte des données aux ajustements nécessaires à la protection du secret statistique, en passant par la correction des données au moyen d'une méthode d'estimation articulée sur un modèle. Quoique l'adoption d'une telle approche ne garantisse pas la résolution de

[1] Stephen E. Fienberg, Maurice Falk Professor of Statistics and Social Science, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890.

toutes les difficultés, on pourrait éliminer certains des aspects les plus problématiques de la pondération et de l'analyse des enquêtes. L'approche préconisée s'écarte radicalement de la pratique courante des organismes de statistique, mais elle aurait l'effet salutaire d'intégrer les approches contemporaines à la méthodologie classique de la statistique, comme on le fait déjà dans d'autres autres secteurs que celui des enquêtes.

On se propose ici d'examiner les implications de l'incertitude et de l'erreur dans le contexte des recensements et des enquêtes sous trois angles différents : celui du producteur (à savoir, les bureaux de statistique), celui de l'utilisateur et celui du méthodologiste. Ce faisant, on formulera des remarques sur les pratiques actuelles et on proposera des stratégies méthodologiques en vue de l'élaboration d'un cadre unifié qui permettra de s'attaquer aux problèmes de l'incertitude et de l'erreur.

Débutons par deux assertions :

- Les organismes de statistique gouvernementaux effectuent un excellent travail pour ce qui est de concevoir les recensements et les enquêtes, de recueillir les données et de prendre des mesures, ce qui inclut la préparation de la documentation sur les diverses sources d'erreur.
- Ceux qui utilisent les données statistiques gouvernementales, que ce soit pour la recherche ou pour élaborer des politiques, ont tendance à avoir une très bonne idée de leurs objectifs analytiques et sont plus que prêts à tenir compte de l'incertitude associée aux données pour les atteindre.

1.1 Énoncé du problème

Si les bureaux de statistique effectuent un si bon travail et si les utilisateurs se montrent si pleins de bonne volonté, où est la difficulté? À notre avis, une des difficultés est que les statisticiens d'enquête sont souvent les seuls à définir la portée et la nature des problèmes substantiels et à établir les mesures importantes à cet égard. Comme Fred Gault et Martin Wilk (1995) l'ont récemment suggéré au sujet des statistiques scientifiques et techniques canadiennes :

> «Qu'on le veuille ou non, à l'instar les comptables, les statisticiens spécifient des mesures et des indicateurs qui anticipent souvent sur les politiques auxquelles on s'intéresse. Bref, les programmes de mesure des statisticiens pourraient exercer une profonde influence sur les politiques. Il s'agit d'un aspect beaucoup trop important pour qu'on le confie uniquement aux statisticiens.» [TRADUCTION]

Les statisticiens d'enquête ont toujours pensé en termes d'objectifs généraux pour la collecte de données et l'utilisation descriptive de ces dernières. Ils ont conçu des méthodes statistiques dans une telle perspective. De son côté, l'utilisateur pense plus souvent qu'autrement en fonction de buts très précis au niveau de l'analyse et des politiques, et il a typiquement à l'esprit un cadre de référence qui correspond aux modèles statistiques ordinaires du genre régression. Ces modèles se caractérisent par leurs propres variabilité et erreurs, mais ne reflètent pas nécessairement les aspects de l'enquête par échantillonnage qui revêtaient de l'importance au moment de la collecte des données. Par conséquent, il y a discordance ou, du moins, il se crée un fossé entre l'information sur l'incertitude et l'erreur que les organismes fournissent avec les données publiques et l'usage que l'utilisateur aimerait faire de ces données, y compris sa capacité à exploiter l'information sur l'incertitude et l'erreur.

Pour remédier au problème, nous préconisons que chaque partie en cause entreprenne les choses différemment! Les bureaux de statistique doivent présenter les données et l'information d'une autre manière, tandis que l'utilisateur doit recourir à une approche méthodique pour tenir pleinement compte de l'incertitude et des erreurs inhérentes lors de son analyse des données. Cette recommandation s'appuie sur les progrès récents de la méthode statistique et a pour buts, d'une part, d'intégrer les sources d'erreur au niveau de l'organisme et d'autre part, de proposer à l'utilisateur les données sous un format qui l'aidera à mieux tenir compte de l'erreur lors de la modélisation et des inférences. Le cadre proposé n'est encore qu'une esquisse et beaucoup de détails techniques restent à régler. Quoi qu'il en soit, nous pensons que le moment est venu d'examiner ces questions et que la version actuelle du cadre nous permet de faire les premiers pas vers une approche statistique unifiée à la collecte, à la divulgation et à l'analyse des données.

1.2 Structure de l'article

Le reste du document est structuré comme suit. La partie 2 décrit la position des producteurs et des utilisateurs de données d'une manière un peu plus détaillée. La partie 3 présente les éléments du cadre unifié, notamment une nouvelle approche à la protection du secret statistique et à la divulgation de fichiers de microdonnées à grande diffusion. La façon dont cette nouvelle approche permet à l'utilisateur de prendre

l'erreur et l'incertitude plus au sérieux est expliquée à la partie 4, où sont aussi exposées plusieurs questions de recherche capitales dont les réponses présentent une importance cruciale en vue d'une application dans la réalité. Enfin, l'annexe décrit les relations entre certaines méthodes populaires de protection du secret statistique pour divers jeux de variables catégoriques et les modèles logarithmiques linéaires. J'y esquisse aussi quelques éléments d'une méthode d'application de la stratégie proposée ici, articulée sur un modèle.

2. POINTS DE VUE SUR LES SOURCES D'ERREUR ET L'ANALYSE DES DONNÉES D'ENQUÊTE

2.1 Le point de vue de l'organisme de statistique

Pour le caractériser, nous débuterons par l'approche classique à la conception d'une enquête et à la quantification, telle qu'on la retrouve dans les ouvrages et les articles sur l'erreur d'échantillonnage et les autres erreurs. Nous passerons ensuite aux principaux aspects du traitement et de la diffusion des données qui font partie du quotidien des bureaux de statistique. Sous cet angle, les sources d'erreur comprennent typiquement les éléments suivants :

- Erreur de la base de sondage (à savoir, sous-dénombrement différentiel du recensement);
- Erreur d'échantillonnage (plan d'échantillonnage complexe de l'enquête);
- Erreur de non-réponse (biais et variabilité);
- Erreur de révision (à savoir, imputation des valeurs manquantes);
- Erreur d'appariement (pour les fichiers de données combinés);
- Erreur non due à l'échantillonnage (à savoir, type d'interview, conception du questionnaire, etc.);
- Erreur de révision liée à la confidentialité (à savoir, attribuable à l'erreur introduite dans les données consécutivement à un codage supplémentaire, à la suppression de cellules, à l'addition de bruit ou à l'échange de données).

L'approche habituelle à un tel assortiment d'erreurs et d'incertitudes consiste à «diviser pour mieux régner». Les organismes s'attaquent à chaque composante, voire sous-composante, mais presque toujours séparément du reste. L'expression «diviser pour mieux régner» utilisée ici ne doit pas être prise au sens péjoratif. Elle souligne simplement le fait que les bureaux de statistique gouvernementaux doivent mener à bien les tâches courantes tout en prenant au sérieux la multitude de problèmes qui affectent les vraies données d'enquête (lire Patz, 1996; Fienberg, Gaynor et Junker, 1996). Une fois que chaque composante a été maîtrisée, le statisticien doit essayer de voir comment les combiner avec les autres éléments en un tout intégré (lire, par exemple, Groves, 1989; Lessler et Kalsbeek, 1992). Les modèles intégrés pour l'erreur dont on peut vraiment se servir à des fins analytiques sont rares.

2.2 Le point de vue de l'utilisateur

L'utilisateur typique des statistiques gouvernementales s'intéresse aux relations et aux liens causals qui l'aideront à prendre certaines décisions en matière de politiques. Il se sert des modèles statistiques pour décrire ces relations. Il considère souvent l'«erreur» comme l'inclusion d'un élément d'erreur dans un modèle d'analyse (C p. ex., le terme ϵ de l'erreur de régression dans l'équation $Y = b_0 + b_1 X + \epsilon$. Outre cela, l'utilisateur type dispose de moyens limités pour assimiler la foule d'informations sur l'incertitude et l'erreur qui émanent du bureau de statistique d'où viennent les données.

Des décennies durant, les échantillonneurs et les statisticiens d'enquête (à quelques exceptions notables) se sont efforcés de convaincre les chercheurs sociaux et les fonctionnaires qui élaborent les politiques que la plupart des enquêtes par échantillonnage et des recensements entrepris par le gouvernement avaient un but plus descriptif qu'analytique (pour un des premiers débats sur la différence entre ces deux aspects, lire Deming, 1978). La pénétration du mythe «descriptif» a notamment eu pour conséquence de creuser un gouffre entre ce que les enquêteurs et les organismes gouvernementaux produisent, d'une part, et ce que les utilisateurs s'efforcent de faire avec les données publiques, d'autre part. Les poids d'échantillonnage de l'enquête (qui reflètent habituellement la probabilité de sélection et certains ajustements pour les non-réponses) et les instructions relatives à la variance de l'échantillon servent couramment d'interface entre les producteurs et les utilisateurs, parmi lesquels il est des personnes plus naïves qui effectuent des analyses pondérées pour la simple raison que des poids accompagnent les données sur la bande et parce qu'elles sont persuadées (à tort) qu'il s'agit de la bonne façon d'intégrer les excentricités

du plan d'échantillonnage à l'analyse quand elle repose sur un modèle.

2.3 Analyse de l'objectif de l'utilisateur

Comme on l'a indiqué précédemment, l'utilisateur type s'intéresse aux modèles analytiques, en particulier à ceux qui autorisent des implications causales. On peut donc dire que son objectif consiste à associer les variables des réponses Y aux variables explicatives X au moyen d'un modèle statistique qui s'efforce de reproduire un phénomène sous-jacent quelconque, jugé important. Malheureusement, il est rare qu'on puisse observer et mesurer Y et X directement. Un recensement ou le questionnaire d'une enquête ne débouchent que sur une mesure apparentée mais faillible des quantités auxquelles on s'intéresse vraiment, que nous appellerons Y^* et X^*.

L'utilisateur s'intéresse aux modèles illustrant la distribution conditionnelle de Y étant donné X. Par conséquent, son objectif consiste à estimer une fonction de distribution cumulative (FDC) à variables multiples de la forme $F_{X|Y}$ ou $F_{Y|X,\theta}$ pour diverses valeurs de X, ou du moins les caractéristiques d'une telle FDC. Dans ce cas, le paramètre θ pourrait désigner la moyenne ou la variance de la population μ ou σ^2, ou un paramètre, voire plusieurs, d'un modèle statistique comme le coefficient de régression β, vraisemblablement sous une forme multidimensionnelle. Bien que les ouvrages sur les enquêtes se soient intéressés dans une certaine mesure au problème des fonctions de distribution estimative (lire par exemple Rao, 1994, et les auteurs auxquels il renvoie), la documentation porte principalement sur une forme univariée de Y. Dans la discussion qui suit, nous négligeons ces sources d'erreur de la mesure de X qui débouchent des sources incluses dans l'évaluation de l'organisme proprement dit et lors de la préparation des données.

L'estimation d'une FDC à variables multiples est un problème de statistique général qui englobe plusieurs cas spéciaux intéressants. Supposons, par exemple, que toutes les variables du modèle de l'utilisateur et de l'ensemble de données soient de nature catégorique, comme cela arrive souvent avec un recensement ou une enquête. La FDC équivaut alors essentiellement au tableau des probabilités conditionnelles (de Y, étant donné X), qui correspond à la classification croisée des variables du tableau de contingence (lire Bishop, Fienberg et Holland, 1975). Nous y reviendrons dans l'annexe et proposerons une bibliographie et des notes importantes sur ce cas particulier. Fienberg, Makov, and Steele (1996) nous fournissent d'autres détails à ce sujet.

2.4 Aproche actuelle des organismes de statistique

Au risque de simplifier à outrance, on peut approximativement caractériser l'approche habituelle à la collecte, au traitement et à la diffusion des données ainsi :

- recueillir et «épurer» les données brutes, ce qui comprend la révision, le couplage et d'autres opérations;
- protéger les données en recourant à une méthode quelconque qui préservera le secret statistique;
- diffuser les données résultantes d'une des deux façons qui suivent, parfois même des deux :
  – sous forme de tableaux marginaux pour une plus vaste classification croisée quelconque (p. ex., choix de classifications croisées marginales - - voir la discussion en annexe sur les liens entre les tableaux marginaux et les modèles logarithmiques linéaires);
  – sous forme de fichiers de microdonnées pour les variables liées à celles qui intéressent l'utilisateur (Y^*, X^*);
- estimer θ directement au moyen d'une quantité $\bar{\theta}$ dérivant de l'échantillon.

En réalité, l'utilisateur suit la voie tracée par l'organisme et estime la FDC $F_{X|Y}$ ou $F_{Y|X,\theta}$ directement à partir des données publiques grâce à la FDC «empirique» (convenablement pondérée pour tenir compte de l'impact du plan d'échantillonnage de l'enquête), $F_{Y|X}$, voire effectue une estimation paramétrique plus élaborée et plus lisse à partir du paramètre estimé $\bar{\theta}$, c'est-à-dire $F_{Y|X,\bar{\theta}}$.

2.5 Carences de l'approche actuelle

Bien que cette approche puisse avoir beaucoup de sens pour certains problèmes de statistique descriptive, le fait demeure que $\bar{F}_{Y|X}$ et $\bar{F}_{Y|X,\bar{\theta}}$ reflètent rarement en entier les aspects de l'erreur due au plan d'échantillonnage que beaucoup de gens jugent importante, notamment la création de grappes. En outre, ces paramètres ne reflètent presque jamais les autres sources d'erreur précitées auprès desquelles l'erreur d'échantillonnage paraît anodine. D'autre part, étant donné les méthodes statistiques relativement rudimentaires utilisées pour préserver la confidentialité, l'utilisateur pourrait toujours réussir à «identifier» les sujets à partir des données diffusées. Une façon de l'éviter consiste à s'attaquer aux diverses composantes de l'erreur et à améliorer séparément la protection du secret

statistique. Une autre solution serait de revoir le problème de déclaration des données d'une nouvelle manière, plus intégrée.

3. NOUVELLE STRATÉGIE, NOUVEAU CADRE

Dans cette partie, nous proposons une nouvelle approche pour la divulgation des données d'enquête. Nous commencerons avec les buts de l'utilisateur et verrons comment les bureaux de statistique devraient structurer les données intéressantes en vue de diffuser des produits de données adaptés aux objectifs de l'utilisateur.

3.1 Genèse de fichiers de «pseudo» microdonnées à grande diffusion

Notre nouvelle approche dépend de la production d'un fichier de microdonnées à grande diffusion destiné à soutenir les analyses de la distribution conditionnelle de Y, étant donné X. La première étape de notre recommandation est la suivante :

1. Combiner les données du recensement ou de l'enquête que l'organisme choisirait de divulguer en temps normal sous la forme $\bar{F}_{Y\bullet|X\bullet}$ et $F_{Y\bullet|X\bullet,\bar{\theta}}$ avec les informations statistiques officielles sur l'erreur, c'est-à-dire erreur de révision, d'appariement, de non-réponse, etc., puis se servir d'une technique paramétrique ou semiparamétrique pour estimer $F_{Y|X}$ et $F_{Y|X,\theta}$ avec $\hat{F}_{Y|X}$ et $\hat{F}_{Y|X,\hat{\theta}}$, respectivement, où $\hat{\theta}$ représente une nouvelle estimation de θ selon la distribution des variables Y et X qui intéresse véritablement l'utilisateur.

Pour une estimation non paramétrique de F, on pourrait songer à une approche statistique classique reposant soit sur un estimateur de densité du noyau ou une estimation «lisse» connexe (lire Scott, 1992), soit à une approche bayesienne articulée sur un mélange de procédés Dirichlet (lire West, Müller, et Escobar, 1994; Gelfand et Mukhopadhyay, 1995), ou sur les arbres de Polya (Lavine, 1992). Ces instruments servent néanmoins surtout à la résolution de problèmes qui comptent peu de dimensions. Par conséquent, il est nécessaire d'entreprendre d'autres recherches pour déterminer s'ils peuvent être adaptés aux problèmes multidimensionnels du recensement et des enquêtes auxquels on s'intéresse. Même si ces méthodes ne s'avèrent pas d'une grande efficacité pour l'estimation statistique, elles pourraient se prêter aux exigences de protection du secret statistique, capitales dans le cadre de la stratégie décrite dans cet article.

En quoi la nouvelle estimation lissée de $F_{Y|X}$ diffère-t-elle de l'estimation explicite ou implicite de l'approche actuelle? Voici trois exemples. Examinons d'abord la diffusion des données du recensement. Aux États-Unis comme au Canada, on a beaucoup écrit sur l'importance du sous-dénombrement et sa répartition entre différents groupes de la population et régions géographiques. Ne pas corriger le sous-dénombrement lors de la diffusion de données sous la forme $\bar{F}_{Y\bullet\,X\bullet}$, déboucherait sur une estimation biaisée de la véritable valeur à laquelle on s'intéresse, $F_{Y|X}$. Deuxièmement, en lissant les données pour faire ressortir les liens semblables à une régression, on parvient typiquement à de meilleures estimations assorties de variances beaucoup plus faibles, quoi qu'au prix d'un certain biais potentiel. Enfin, en intégrant l'information des organismes sur les composantes de l'erreur (qui ont tendance à relever la variance) à la méthode d'estimation statistique, on obtient un nouvel estimateur lissé de $F_{Y|X}$.

Les étapes subséquentes de l'exercice sont les suivantes:

2. Au lieu de diffuser la FDC estimée à l'étape 1 qui précède, l'organisme procède à un «échantillonnage» pour créer un fichier de «pseudo» microdonnées baptisé $\hat{F}_{Y|X}$ et $\hat{F}_{Y|X,\hat{\theta}}$. (Le trait supérieur indique qu'il s'agit d'un échantillon de la FDC lissée, conformément à la manière dont nous avons déjà noté la FDC empirique, qui correspond à un échantillon).
3. L'organisme reprend l'«échantillonnage» puis diffuse les fichiers de «pseudo» microdonnées réitérés.

3.2 Caractéristiques du fichier de «pseudo» microdonnées

Le fichier de «pseudo» microdonnées créé de la façon indiquée plus haut présente plusieurs caractéristiques intéressantes. Tout d'abord, si on considère $\hat{F}_{Y|X}$ et $\hat{F}_{Y|X,\hat{\theta}}$ comme une série d'enregistrements publics sur des particuliers, les «particuliers» en question ne correspondent pas nécessairement à des sujets de l'échantillon original de l'enquête. La population est donc rassurée au sujet de la protection de la confidentialité des réponses, même si un intrus parvenait malgré tout à effectuer indirectement des inférences sur les sujets de l'échantillon original.

Il s'agit là d'un point particulièrement important du point de vue de la protection du secret statistique. Puisque les sujets du fichier de «pseudo» microdonnées ne correspondent pas nécessairement à ceux de

l'échantillon original, on apaise au moins partiellement les préoccupations concernant la confidentialité. Après tout, on ne semble même plus diffuser de données sur les membres de cet échantillon. Ce débat sur la protection du secret statistique reste néanmoins quelque peu spécieux. En effet, en réalité, les sujets dont les valeurs Y et X sont fort éloignées de celles du reste de l'échantillon pourraient toujours être reconstitués grâce à ce processus d'estimation statistique complexe et réapparaître virtuellement intacts dans le fichier de «pseudo» microdonnées. Des vérifications empiriques de l'efficacité des méthodes utilisées pour protéger le secret statistique restent donc nécessaires. Nous préconisons en particulier l'étude de ce problème sous l'angle d'un intrus (lire Fienberg, Makov et Sanil, 1994).

Deuxièmement, cette méthode présente des liens étroits avec deux méthodes statistiques qui ont récemment vu le jour : (1) la méthode bootstrap (Efron, 1979; Efron et Tibshirani, 1993; Hall, 1992), une méthode classique qui nécessite un échantillonnage réitéré (non exhaustif) à partir d'une fonction de distribution empirique et (2) l'imputation multiple (Rubin, 1987, 1993), une méthode bayesienne qui permet d'obtenir des valeurs par échantillonnage d'une distribution a posteriori. En ce qui concerne l'estimation implicite à l'approche décrite plus haut, notre préférence va au point de vue bayesien. Nous proposons donc aux bureaux de statistique d'estimer d'abord la fonction de distribution empirique pour créer une distribution a posteriori complète de $F_{Y|X}$ ou $F_{X|Y,\theta}$ puis de procéder à un échantillonnage avec la méthode d'imputation multiple de Rubin. De cette façon, la méthode bootstrap peut être considérée comme une technique d'échantillonnage à peu près analogue à la moyenne de la distribution a posteriori.

Troisièmement, le plan d'échantillonnage des enregistrements rendus publics ne doit pas nécessairement être identique à celui de l'enquête par sondage originale. Pour cette raison, l'organisme pourrait recourir, théoriquement du moins, à un simple échantillonnage aléatoire, voire à un échantillonnage non exhaustif à partir de $\hat{F}_{Y|X}$ ou $\hat{F}_{Y|X,\theta}$. Rubin (1993) le souligne sans expliquer exactement de quelle façon on détermine ce qu'on pourrait appeler la taille de l'échantillon «équivalent» des fichiers de données diffusées. Le principe heuristique est que les données ne fournissent qu'une quantité limitée d'informations que le rééchantillonnage ne permet pas d'augmenter. Afin de maintenir le degré d'exactitude approprié dans les données, on a besoin d'un échantillon bootstrap de taille au moins équivalente, sur le plan conceptuel, à la «taille efficace de l'échantillon» pour le plan d'échantillonnage complexe, de façon à refléter les effets du plan. Il s'agit néanmoins d'une notion un peu problématique, car la «taille efficace de l'échantillon» pourrait très bien varier d'une analyse à l'autre!

L'aspect sans doute le plus important de cette approche reste cependant que l'utilisateur peut désormais se servir de fichiers de «pseudo» microdonnées pour estimer les quantités spécifiques qui l'intéressent, p. ex. θ, au moyen de *méthodes statistiques types*. Essentiellement, l'idée consiste à utiliser une méthode type comme l'analyse de régression, ou une méthode un peu plus élaborée, pour obtenir des estimations cohérentes des coefficients auxquels on s'intéresse. On ne peut toutefois utiliser les estimations habituelles de l'erreur-type qu'entraînent les instruments d'analyse standard. Une des leçons qu'on tire des méthodes bootstrap et d'imputation multiple est que si l'on peut estimer θ en appliquant des méthodes de statistique types à l'échantillon obtenu par une des deux méthodes précitées, il est impossible de se faire une idée juste de la variabilité des estimations sans recourir à d'autres versions du fichier de «pseudo» microdonnées. Les répétitions multiples demeurent cependant relativement faciles à effectuer et l'estimation des variances au moyen de versions multiples des paramètres estimatifs devient un exercice assez simple qui ne requiert aucun logiciel spécial.

4. PRENDRE LA VARIABILITÉ AU SÉRIEUX

Nous pensons qu'il est important de faire une distinction entre la production de fichiers de microdonnées à grande diffusion reposant sur des personnes et des données réelles grâce à un processus de simulation statistique comme celui dont nous venons de parler, et le modèle de microsimulation typique, qui peut reposer sur des modèles statistiques apparentés sans faire appel à des données sur des sujets réels. Il existe une différence appréciable entre les «pseudo personnes», qui ressemblent aux sujets d'où viennent véritablement les données, et les personnes «imaginaires» pour lesquelles des données ont été inventées de toutes pièces par un processus de modélisation stochastique ou non. Le présent article s'appuie sur la première méthode, pas la seconde.

4.1 Avantages du cadre envisagé

Le cadre envisagé, décrit plus haut, présente plusieurs avantages. Tout d'abord, nous croyons qu'il obligerait les bureaux de statistique à prendre plus au sérieux leurs données et leurs sources d'erreur, puisque

ces aspects jouent un rôle clé dans la modélisation décrite à la partie 3. En deuxième lieu, nous pensons qu'on résoudrait ainsi en grande partie le problème de protection du secret statistique. Troisièmement, le cadre produirait des fichiers de microdonnées à grande diffusion d'un type qui permettrait à l'utilisateur de recourir aux méthodes statistiques types et aux méthodes de recherche par modèle. Ces avantages amèneraient les deux parties vers une façon plus simple et plus efficace d'estimer la variance, d'où le titre de la présente communication.

4.2 Exemples de recherches à effectuer

Avant qu'un bureau de statistique puisse se servir du cadre proposé, un certain nombre de détails techniques formidables restent à régler. En voici quelques-uns :

- De quelle manière l'organisme devrait-il combiner les nombreuses sources d'erreur et d'incertitude?

- Quelles méthodes de lissage devrait-on utiliser et quelle devrait être l'importance du lissage?

- Comment peut-on établir la taille de l'échantillon «efficace» pour les fichiers de «pseudo» microdonnées? L'application des principes bootstrap repose sur l'expansion de certaines séries (lire Hall, 1992), ce qui exige l'utilisation d'un échantillon bootstrap de la taille identique à celle de l'échantillon original. Quelle serait la notion équivalente dans le cas qui nous intéresse?

- Combien de répétitions sont nécessaires pour estimer la variance? Rubin (1987, 1993) en préconise quatre ou cinq dans le contexte d'une imputation multiple. Efron et Tibshirani (1993) recourent pour leur part à un nombre de répétitions important. Un nombre inférieur de répétitions suffirait-il pour l'une ou l'autre approche?

Par ailleurs, il se pourrait que l'application proprement dite des algorithmes aux situations très multidimensionnelles que supposent les données du recensement et des enquêtes nécessitent de nouvelles méthodes et de nouveaux principes de statistique. Par exemple, ainsi que nous le suggérons en annexe, les distributions des tableaux de contingence multidimensionnels n'ont été simulées sous réserve de contraintes marginales qu'essentiellement pour des tableaux à deux et à trois dimensions. On aura besoin de nouvelles stratégies et de nouveaux algorithmes pour des tableaux comprenant un plus grand nombre de dimensions. Il s'agit d'un domaine de recherche pointu en statistique et en mathématiques.

Enfin, il se pourrait qu'on doive réfléchir aux problèmes d'estimation statistique décrits ici sous un angle différent de celui habituellement présenté dans la littérature sur la méthodologie. En raison de la multiplicité des objectifs qu'on s'efforce d'atteindre, peut-être devra-t-on envisager de fournir à l'utilisateur des données qui lui permettront de se rapprocher des distributions conditionnelles $\hat{F}_{Y|X}$ et $\hat{F}_{Y|X,\,\theta}$ au lieu de les reproduire d'une manière statistique plus précise.

4.3 Résumé

Dans le présent document, nous nous sommes efforcés de montrer que les organismes gouvernementaux, à l'instar des utilisateurs, assument certaines responsabilités au niveau de l'utilisation des données du recensement et des enquêtes. Les bureaux de statistique ne peuvent désormais plus se limiter à la production de fichiers de grande diffusion et de longues séries de tableaux ainsi qu'ils l'ont fait dans le passé. Ils ne peuvent non plus continuer à ignorer les objectifs analytiques de l'utilisateur. Parallèlement, ce dernier doit apprendre de quelle manière diverses sources d'erreur d'enquête affectent ses objectifs et intégrer cette information aux techniques statistiques dont il se sert.

Nous avons soutenu qu'en examinant et en exploitant les plus récents progrès de la méthodologie statistique, on pourrait élaborer une approche intégrée à la diffusion et à l'analyse des données d'enquête qui aiderait chacun de nous à prendre l'incertitude et l'erreur au sérieux. Le cadre proposé ici constitue peut-être un pas dans cette direction.

ANNEXE : NOTES ET RÉFÉRENCES SUR LE CAS DES DONNÉES CATÉGORIQUES

Cette annexe propose une description annotée du processus d'estimation et de simulation mentionné à la partie 3 pour le cas particulier des variables catégoriques et des classifications croisées. Nous nous concentrerons sur l'estimation paramétrique de la FDC qui, comme indiqué précédemment, revient à estimer la probabilité par cellule dans un tableau de contingence. Fienberg, Makov et Steele (1996) donnent plus de précisions sur cette méthode, dont ils offrent également une description.

La catégorie de modèle statistique la plus couramment utilisée avec les données des tableaux de

contingence est le modèle logarithmique linéaire. Pour une série de plans d'échantillonnage fondamentaux (lire Bishop, Fienberg et Holland, 1975), on note un lien direct entre le modèle logarithmique linéaire à hiérarchie spécifique et une série de tableaux marginaux correspondant aux statistiques suffisantes minimales associées au modèle. Si on ne signale que les totaux marginaux relatifs à un modèle logarithmique linéaire bien ajusté aux données, un autre chercheur pourra déduire les probabilités par cellule du tableau de contingence complet (lire Fienberg, 1975). Par ailleurs, ne fournir qu'un jeu spécifique de tableaux marginaux revient à dire qu'il s'agit des seuls totaux dont on a besoin pour effectuer des inférences. On suggère donc implicitement que tel ou tel modèle logarithmique linéaire donnera les résultats appropriés.

Les deux méthodes les plus couramment utilisées pour protéger le secret statistique avec des variables catégoriques sont (i) la suppression des cellules (lire Carvalho, Dellaert et Osório, 1994; Cox, 1980, 1995; Robertson, 1993; et Subcommittee on Disclosure-Avoidance Techniques, 1994) et (ii) l'échange de données (lire Dalenius et Reiss 1982; Griffin, Navarro et Flores-Baez, 1989; et Subcommittee on Disclosure-Avoidance Techniques, 1994). Malheureusement, la documentation sur la protection du secret statistique avec les variables catégoriques semble s'être radicalement écartée de ce qu'on considère désormais comme les ouvrages de référence sur les modèles logarithmiques linéaires applicables aux données catégoriques. La chose est malheureuse car le maintien de la marge est une notion fondamentale à la fois à la suppression des cellules et à l'échange des données. Dans le premier cas, les cellules sont supprimées sous réserve de contraintes marginales alors que dans le second, on échange des sujets respectant un jeu de marges établies d'une cellule à l'autre, ce qui permet de maintenir les autres totaux. Les principales caractéristiques de ces deux méthodes peuvent être intégrés aux modèles logarithmiques linéaires, ce qui suggère d'autres manières de protéger le secret statistique. Il se pourrait que d'autres résultats sur les modèles logarithmiques linéaires mentionnés dans la littérature nous aident à comprendre les propriétés de méthodes comme la suppression de cellules et l'échange des données (lire la discussion dans Fienberg, 1995).

Trouver un tableau de tri croisé comprenant des chiffres qui satisfont une série donnée de contraintes marginales est un problème qui a tenu éveillé de bon nombre de statisticiens, ces dernières années (lire Agresti, 1993; Zelterman, Chan et Mielke, 1995). Plusieurs algorithmes ont été proposés, mais on s'en est principalement servi pour des classifications croisées à deux et à trois dimensions. De nouvelles idées, extraites de la documentation sur les modèles logarithmiques linéaires graphiques, suggèrent qu'on pourrait enfin être sur le point d'appliquer cette méthode à un plus grand nombre de dimensions (lire Diaconis et Sturmfels, 1993 pour un algorithme éventuel et Lauritzen, 1996 ou Whittaker, 1990 pour plus d'explications sur les modèles graphiques). Le cadre décrit à la partie 3 requiert la production d'une FDC lissée qui servirait à l'échantillonnage. Dans le contexte actuel, on pourrait en déduire, au moins de façon heuristique, qu'il faudrait envisager les tirages de la distribution exacte, sous réserve d'un ensemble fixe de totaux marginaux. Néanmoins, on pourrait aussi choisir de n'utiliser qu'un tableau artificiel s'il satisfait au moins à un modèle logarithmique linéaire quelconque de degré supérieur. Lauritzen et Whittaker, Diaconis et Sturmfels parlent de l'importance des modèles graphiques. Une autre solution consisterait à créer une distribution a posteriori complète des probabilités par cellule pour le tableau, bref recourir aux méthodes d'Epstein et de Fienberg (1992), puis à échantillonner cette distribution.

Pour d'autres explications sur les problèmes et les approches décrits dans l'annexe, lire Fienberg, Makov et Steele (1996).

REMERCIEMENTS

La rédaction de cet article a pu être en partie réalisée grâce à un contrat avec WESTAT et le Bureau of the Census américain. Je tiens particulièrement à remercier David Binder pour m'avoir fait part de ses premières impressions sur la stratégie générale, ainsi que Peter Müller, Danny Pfefferman et Steffen Lauritzen pour m'avoir fourni les références qu'on retrouve dans les pages qui précèdent. Aucun d'eux ne doit néanmoins assumer la responsabilité pour la manière dont j'ai utilisé la documentation suggérée, ni pour mes spéculations sur les possibilités d'appliquer telle ou telle méthode statistique.

BIBLIOGRAPHIE

Agresti, A. (1992). A survey of exact inference for contingency tables (with discussion), *Statistical Science*, 7, 131-177.

Bishop, Y.M.M., Fienberg, S.E., et Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Carvalho, F.de, Dellaert, N., et Osório, M.deS. (1994). Statistical disclosure in two-dimensional tables: General tables, *Journal of the American Statistical Association*, 89, 1547-1557.

Cox, L. (1980). Suppression methodology and statistical disclosure control, *Journal of the American Statistical Association*, 75, 377-385.

Cox, L. (1995). Network models for complementary cell suppression, *Journal of the American Statistical Association*, 90, 1453-1462.

Dalenius, T., et Reiss, S.P. (1982). Data-swapping: a technique for disclosure control, *Journal of Statistical Planning and Inference*, 6, 73-85.

Deming, W.E. (1978). Sample surveys: the field, in *International Encyclopedia of Statistics, Vol 2* (sous la direction de W.H. Kruskal et J.M. Tanur), New York: Macmillan and the Free Press, 867-885.

Diaconis, P., et Sturmfels, B. (1993). Algebraic algorithms for sampling from conditional distributions, Manuscrit inédit.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *Annals of Statistics*, 7, 1-26.

Efron, B., et Tibshirani , R. (1993). *An Introduction to the Bootstrap*, New York: Chapman and Hall.

Epstein, A.D., et Fienberg, S.E. (1992). Bayesian estimation in multidimensional contingency tables, dans *Proceedings of Indo-U.S. Workshop on Bayesian Analysis in Statistics and Econometrics* (sous la direction de P.K. Goel et N.S. Iyengar), notes dans Statistics Vol. 75, New York: Springer-Verlag, 27-47.

Fienberg, S.E. (1975). Perspectives Canada as a social report, *Social Indicators Research*, 2, 153-174.

Fienberg, S.E. (1994a). Conflicts between the needs for access to statistical information and demands for confidentiality, *Journal of Official Statistics*, 10, 115-132.

Fienberg, S.E. (1994b). A radical proposal for the Provision of Micro-data Samples and the Preservation of Confidentiality, Technical Report No. 611, Department of Statistics, Carnegie Mellon University, PA.

Fienberg, S.E. (1995). Discussion of presentations on statistical disclosure methodology, dans *Seminar on New Directions in Statistical Methodology, Statistical Policy Working Paper No. 23*, Federal Committee on Statistical Methodology, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC, Part 1, 68-79.

Fienberg, S.E., Gaynor, M., et Junker, B. (1996). Discussion of "Modelling mortality rates for elderly heart attack patients: Profiling hospitals in the Cooperative Cardiovascular Project," dans *Case Studies in Bayesian Statistics III*, New York: Springer-Verlag (sous presse).

Fienberg, S.E., Makov, U.E., et Sanil, A. (1994). A Bayesian Approach to data Disclosure: Optimal Intruder Behavior for Continuous Data, Technical Report No. 608, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Fienberg, S.E., Makov, U.E., et Steele, R. (1996). Statistical notions of data disclosure avoidance and their relationship to traditional statistical methodology: Data swapping and loglinear models. *Proceedings of the Annual Research Conference, U.S. Bureau of the Census*, U.S. Departement of Commerce, Washington, DC, (à paraître).

Gault, F.D., et Wilk, M.B. (1995). Science and technology measurement -- Rhetoric and reality, Document préparé en vue d'une présentation au "Symposium international sur l'évaluation de l'impact de la R et D", Ottawa, Canada, 13-15 septembre 1995.

Gelfand, A.E., et Mukhopadhyay, S. (1995). On nonparametric Bayesian inference for the distribution of a random sample, *Canadian Journal of Statistics*, 23, 411-420.

Griffin, R., Navarro, A., et Flores-Baez, L. (1989). Disclosure avoidance for the 1990 census, *Proceedings of the Section on Survey Research*, American Statistical Association, 516-521.

Groves, R.M. (1989). *Survey Errors and Survey Costs*, New York: John Wiley.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

Lessler, J.L., et Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*, New York: John Wiley.

Lauritzen, S. (1996). *Graphical Association Models*, New York: Oxford University Press.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modelling, *Annals of Statistics*, 20, 1222-1235.

Patz, R. (1996). Hierarchical models for new modes of educational assessment, Dissertation inédite, Department of Statistics, Carnegie Mellon University.

Robertson, D. (1993). Cell suppression at Statistics Canada, *Proceedings of the Annual Research Conference, U. S. Bureau of the Census*, U.S. Department of Commerce, Washington, DC, 107-131.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley.

Rubin, D.B. (1993). Discussion, statistical disclosure limitation, *Journal of Official Statistics*, 9, 461-468.

Rao, J.N.K. (1994). Estimation totals and distribution functions using auxiliary information at the estimation stage, *Journal of Official Statistics*, 10, 153-165.

Scott, D.W. (1992). *Multivariate Density Estimation. Theory, Practice and Visualization*, New York: John Wiley.

Subcommittee on Disclosure-Avoidance Techniques (1994). Report on Statistical Disclosure Methodology,. Statistical Policy Working Paper No. 22, Federal Committee on Statistical Methodology, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington, DC.

West, M., Müller, P., et Escobar, M. (1994). Hierarchical priors and mixture models, with application in regression and density estimation dans *Aspects of Uncertainty* (sous la direction de P.R. Freeman et A.F.M. Smith), New York: John Wiley, 363-386.

Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics,. New York: John Wiley.

Zelterman, D., Chan, I.S.-F., et Mielke, P.W., Jr. (1995). Exact tests of significance in higher dimensional tables, *American Statistician*, 49, 357-361.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

PROBLÉMATIQUE DE L'AFFECTATION DES RESSOURCES

T.M.F. Smith[1]

RÉSUMÉ

L'affectation des ressources est un problème auquel on s'est relativement peu attardé dans la littérature scientifique. Le cadre théorique de la décision ponctuelle proposé par Fellegi et Sunter (1973), tel que l'ont appliqué Linacre et Trewin (1993), est un progrès dans la bonne direction. L'auteur examine ici plusieurs fonctions de perte et préconise l'erreur quadratique moyenne *(EQM)* totale pour mesurer convenablement la qualité totale. On ne peut cependant recourir à cette méthode que s'il est possible d'estimer la variance, le biais et le coût marginal en termes quantitatifs en usant, au besoin, de son jugement professionnel. Le cadre formel fait ressortir les aspects qui profiteraient d'un complément d'information.

MOTS CLÉS : Théorie de la décision; fonctions de perte; erreurs d'enquête; coûts d'enquête.

1. INTRODUCTION

Dans son ouvrage intitulé «Accuracy of Economic Observations», dont la lecture devrait être obligatoire pour tous ceux qui produisent et utilisent des statistiques économiques, Oskar Morgenstern (1963) montre la fausseté de la plupart des étalons de l'exactitude des statistiques officielles. La qualité des statistiques officielles constitue une source de préoccupation constante depuis les débuts de l'échantillonnage, et le rapport de l'Institut International de Statistique (IIS) de 1926 sur la méthode représentative mentionne ce qui suit, qu'on pourrait qualifier d'énoncé de mission de tout bureau de la statistique :

> *«Il faut s'efforcer le plus possible d'engendrer un climat de confiance mutuelle entre l'organe faisant office de service de statistique officiel et la population, non seulement d'où proviennent les données vitales aux statistiques, mais aussi pour qui s'effectue ce travail. Le service de statistique officiel devrait bien sûr veiller jalousement à sa réputation puisque comme chacun sait, il ne suffit pas que la femme de César soit vertueuse, le monde entier doit en être convaincu.» [Traduction]*

Dans les enquêtes par sondage, l'exactitude demeure la vertu cardinale et on ne peut y prétendre qu'en adoptant des méthodes qui minimiseront les erreurs à un coût établi. Avant de minimiser les erreurs toutefois, il convient de les quantifier, tandis que le contrôle des coûts exige l'obtention de renseignements adéquats sur le coût des méthodes sous-jacentes. Fort de telles informations, l'analyste peut ensuite envisager une répartition rationnelle des ressources.

Le rapport de l'IIS se terminait sur l'avertissement que voici, que tous les statisticiens officiels contemporains reconnaîtront sur-le-champ :

> *«Il serait imprudent de croire que la statistique officielle ne traverse pas une période difficile. Alors que les services de statistique font l'objet de demandes croissantes, tant en ce qui concerne l'extension des calculs à de nouveaux domaines que des enquêtes plus poussées, les finances publiques sont mises à rude épreuve et le budget des bureaux de la statistique a plus tendance à régresser qu'à augmenter. Dans une telle situation, il n'y a guère d'autre solution que prendre tous les moyens nécessaires afin d'en obtenir plus avec les ressources financières et les effectifs existants.» [Traduction]*

[1] T.M.F. Smith, Department of Mathematics, University of Southampton, Southampton, Royaume-Uni, S09 5NH.

Plus ça change, plus c'est la même chose! Soixante-dix ans plus tard, les statisticiens se demandent toujours comment répartir les ressources limitées à leur disposition pour parvenir à des résultats de haute qualité, malgré diverses erreurs. Bien que le rapport de l'IIS ait également reconnu l'importance potentielle des autres erreurs que celle de l'échantillonnage sur l'exactitude des estimations issues des sondages, ce n'est que dans les années 40 qu'on a entrepris des études systématiques sur les effets que les erreurs attribuables aux non-réponses, aux réponses et au codage peuvent avoir sur les estimations, en Inde et aux États-Unis. L'élaboration du modèle de l'erreur d'enquête totale pour le recensement par Hansen et coll. (1961) a établi le cadre théorique nécessaire à l'étude de l'exactitude générale. Des travaux subséquents ont perfectionné certains aspects du modèle, permis de jauger diverses sources d'erreurs et proposé des méthodes permettant d'ajuster l'analyse en présence d'erreurs spécifiques autres que l'erreur d'échantillonnage. Néanmoins, dans la pratique, on s'est relativement peu intéressé à la question de l'exactitude globale et au problème connexe de l'affectation des ressources.

Statistique Canada fait sans conteste partie des organismes qui se sont attaqués au problème et, dans un article présenté à la séance de l'IIS à Vienne, Fellegi et Sunter (1973) proposent un cadre utile pour étudier l'affectation des ressources. La rencontre était présidée par Tore Dalenius qui, depuis les années 50, n'a cessé de prôner l'examen du plan d'échantillonnage global en fonction de l'erreur d'échantillonnage totale. Les autres documents présentés à cette occasion, par Jabine et Tepping (1973) et Nathan (1973), abordaient des particularités du même problème sans toutefois avancer de cadre théorique général. Dans son énoncé de politique, Statistique Canada (1987) définit les critères applicables à la présentation d'information sur les erreurs, alors que Groves (1989) explore le chemin peu fréquenté du coût des enquêtes. Je n'ai pour ma part découvert qu'un article, celui de Linacre et Trewin (1993), traitant des aspects pratiques du problème de la répartition des ressources.

2. ERREUR D'ENQUÊTE TOTALE

On connaît désormais très bien les éléments qui composent l'erreur d'enquête totale. Chaque partie du processus de l'enquête, de la formulation du concept à la collecte et à l'analyse des données, en passant par l'élaboration du plan d'échantillonnage, présente des risques d'erreur. Groves (1989, p. 17) a créé un diagramme illustrant les composantes de l'erreur totale, ventilée en biais et en variances. On peut souvent estimer la variance à partir de l'enquête, ou d'études spéciales entreprises dans le cours de l'enquête, mais calculer le biais exige fréquemment des données externes. Par conséquent, il n'est guère surprenant que le gros de la recherche ait porté sur les éléments de la variance, plus faciles à mesurer. Les études sur le biais dignes de ce nom se rapportent essentiellement au recensement, pour lequel il n'existe pas d'erreur d'échantillonnage. Cette répartition des efforts de recherche par les théoriciens des enquêtes n'est pas à la hauteur de l'apport des biais à l'erreur d'enquête totale, apport qui, logiquement, devrait être du même ordre de grandeur que celui des variances.

Si T estime une valeur quelconque de la population, θ par exemple, et qu'on identifie k sources d'erreur, un modèle permettant l'estimation de l'erreur totale pourrait s'exprimer comme suit :

$$T = \theta + \sum_{j=1}^{k} A_j, \qquad (1)$$

où A_j correspond à l'erreur de la source j. Les erreurs ont une structure complexe puisque l'erreur d'échantillonnage dépend du plan d'échantillonnage, celle des non-réponses, de l'échantillon, celle des réponses, des répondants et celle du codage, de la correction des réponses. La formation et la gestion peuvent influer sur les erreurs à tous les échelons de la collecte des données. Malgré cette complexité, ou peut-être à cause d'elle, on formule habituellement les hypothèses qui suivent au sujet des erreurs:

$$\begin{aligned} E(A_j) &= B_j, \\ V(A_j) &= \sigma_j^2, \\ cov(A_j, A_i) &= 0, \quad j \neq i, \end{aligned} \qquad (2)$$

la dernière étant la plus douteuse. Pour une erreur quelconque, il se pourrait que $\sigma_j^2 = 0$, ou $B_j = 0$, mais pas les deux. Par exemple, les erreurs de la base d'échantillonnage créent un biais; les erreurs d'échantillonnage, par contre, n'aboutissent souvent qu'à des variances, car le biais est fréquemment d'un ordre de grandeur négligeable, tandis que les erreurs dues aux réponses et à la correction peuvent entraîner à la fois un biais et une variance. L'ordre de grandeur et l'importance relative des erreurs dépendront du sondage. Diverses études montrent que dans de nombreux cas, la variance des réponses peut avoir autant sinon plus d'importance que la variance due à l'échantillonnage, et que le biais attribuable à la couverture et aux non-réponses pourrait dominer les éléments de la

variance. Quoique la variance diminue habituellement lorsque la taille de l'échantillon augmente, dans la plupart des cas, le biais reste constant, si bien que son importance relative tend à grandir. Pourtant, dans la plupart des enquêtes, on persiste à ne mesurer que les erreurs d'échantillonnage et on se borne à des déclarations qualitatives sur les variances et les biais d'autres origines. Les rapports mentionnent souvent des seuils de confiance qui reposent uniquement sur une estimation de l'erreur d'échantillonnage, ce qui peut s'avérer inutile en présence d'un biais. L'honnêteté exigerait qu'on s'efforce de mesurer l'erreur d'enquête totale avant la rédaction du rapport.

L'incidence des erreurs ne dépend pas seulement du type de sondage, mais aussi de ce qu'on estime. Si le biais peut être minime dans un agrégat, il arrive qu'il soit appréciable dans les estimations d'une sous-population reprenant la même variable. Heureusement, l'inverse est faux. Si le biais est faible au niveau unitaire, il le sera aussi au niveau de l'agrégat. Selon Jabine et Tepping (1973), bien que le biais d'un changement net puisse être faible quand deux enquêtes s'effectuent dans des conditions analogues, les biais des changements bruts correspondants peuvent être importants. Le biais peut varier dans le temps et l'espace et les valeurs ne s'annulent pas nécessairement quand on détermine l'écart, comme avec les mesures du changement. Le fiasco des sondages d'opinion publiques effectués aux élections générales du Royaume-Uni, en 1992, a été attribué à maints facteurs, mais aucun d'eux n'explique le succès des sondages aux élections antérieures. L'explication la plus plausible est que les sondages étaient largement teintés d'erreurs mais qu'en 1992, ces erreurs se sont additionnées. Les résultats de l'élection déterminent la valeur des sondages d'opinion. Combien d'enquêtes officielles jouissent-elles d'une validation aussi rigoureuse? Les statisticiens officiels sont-ils vraiment sûrs que les erreurs ne s'accumulent pas parfois toutes dans le même sens?

Ma thèse est que les statisticiens n'ont pas réussi à instaurer un mécanisme qui permettrait une évaluation machinale des principales sources d'erreur de l'enquête. Sans une telle évaluation, on ne peut espérer améliorer systématiquement les procédés d'enquête. Sans elle et face au coût des enquêtes, il est en outre impossible de répartir efficacement les ressources destinées aux enquêtes.

3. AFFECTATION DES RESSOURCES

Comment affecter les ressources se résume à un problème de décision. La théorie de la décision en statistique débute par une fonction de perte qui jauge les conséquences des autres décisions. La difficulté qu'il y a à définir une fonction de perte dans le cadre d'une enquête complexe à variables et à buts multiples a découragé plus d'un analyste à recourir à des méthodes formelles. Pourtant, peu importe l'enquête, l'affectation des ressources exige qu'on évalue l'importance relative des divers aspects de l'enquête sur la qualité des résultats. N'effectuer aucun changement reste une décision en soi. Selon d'autres documents présentés au symposium, il appert que la façon dont Statistique Canada répartit ses ressources est loin d'être idéale, puisqu'on semble notamment affecter des ressources excessives à la correction, dans certains cas.

3.1 Fonctions de perte

La fonction de perte choisie dépend de l'utilisateur. Nombreux sont ceux qui utilisent les données des enquêtes. On ne peut donc satisfaire chacun d'eux. Dans une telle situation, il est raisonnable que le statisticien de formation suggère une fonction de perte statistique. Dans les enquêtes par sondage, où dominent les moments de second ordre, on recourt habituellement à l'écart-type l'erreur quadratique moyenne (EQM). L'EQM(T) de l'équation (1), est

$$EQM(T) = E(T-\theta)^2 = V(T) + B^2, \quad (3)$$

où,

$$B = \sum_{i=1}^{k} B_i, \quad (4)$$

représente le biais global. L'EQM varie avec chaque variable et domaine d'étude, si bien que le choix d'une ou de plusieurs EQM dépend des visées de l'utilisateur. Une approche envisageable consisterait à établir une fourchette de fonctions de perte et à estimer les répercussions de différentes affectations (décisions) pour les valeurs de la fourchette. La répartition des pertes facilitera la décision finale.

L'EQM présente-t-elle une utilité quelconque comme fonction de perte pour les problèmes d'affectation des ressources? Quoique la majorité des auteurs y recourent sans même y penser, on note des variantes. Ainsi, Nathan (1973) estime que l'EQM total correspond à la somme de l'EQM de l'échantillonnage et de l'EQM du non-échantillonnage. Pour leur part, Fellegi et Sunter (1973) baptisent leur fonction de perte EQM alors qu'en réalité, ils utilisent une équation analogue à

$$L(T) = \sum_{1}^{k} \sigma_j^2 + \sum_{1}^{k} B_j^2 \quad . \qquad (5)$$

Cette équation ressemble plus à l'EQM total qu'à l'EQM d'un total, puisqu'elle est la somme des EQM qui la composent.

En ce qui concerne les décisions relatives à l'affectation des ressources, L(T) paraît présenter maints avantages par rapport à EQM(T). Soit une enquête présentant les erreurs A_1, A_2, les variances σ_1^2 , et σ_2^2 , et les bias B_1 et B_2 . Si $B_1 > 0$ et $B_2 < 0$, et que $B = B_1 + B_2 > 0$, on pourrait réduire B en augmentant l'importance du biais B_2 . Dans un tel cas, EQM(T) diminue tandis que L(T) augmente. Il est difficile de soutenir qu'une méthode qui accroît le biais d'une erreur sans modifier les autres erreurs puisse améliorer la qualité générale d'une enquête. On peut pousser l'argument un rien plus loin en soulignant que les erreurs à un niveau d'agrégation élevé correspondent à la somme des erreurs des niveaux inférieurs. À l'extrême, on ne pourra garantir la qualité globale qu'en minimisant l'EQM des composantes de l'erreur au niveau de l'unité. Pour cela, il est capital de déplacer le problème en amont dans le processus de l'enquête et d'améliorer les méthodes qui servent à recueillir, à quantifier et à traiter les données.

Les statisticiens officiels incitent les utilisateurs à calculer les intervalles de confiance pour tenir compte des erreurs attribuables à l'enquête. Les déclarations du Bureau of the Census américain au sujet des sources et de l'exactitude des données (lire Alexander (1994)) en sont un bon exemple. Il s'ensuit que les intervalles pourraient servir de fonction de perte. Les effets du biais sur les propriétés de couverture des intervalles de confiance selon la théorie courante sont bien connus. Kish (1965) a tracé un graphique de la superficie des queues prises séparément et globalement pour divers seuils de confiance nominaux, d'après le ratio du biais $R = B/\sigma$. Si $R = 1{,}0$, avec un seuil de confiance de 95 p. 100, la queue de gauche donne 0,0015 et celle de droite, 0,1685, pour une couverture totale de 0,1700. Comme il fallait s'y attendre, le biais a une incidence plus prononcée sur les queues prises séparément. Le véritable problème quand on se sert du niveau de couverture comme fonction de perte est qu'on peut améliorer la couverture en réduisant la taille du ratio B/σ, par exemple en diminuant le biais total, tel qu'indiqué précédemment, ou en augmentant la variance par rapport au biais. Ces deux solutions ne représentent néanmoins ni l'une, ni l'autre une amélioration au niveau de la qualité totale. À elle seule, la couverture ne décrit pas adéquatement les propriétés des intervalles de confiance et ne convient pas à l'évaluation de la qualité totale.

J'en conclus qu'une mesure de la qualité totale de l'enquête devrait reposer sur la somme des composantes, comme c'est le cas avec L(T) en (5), et non sur une mesure globale, comme EQM(T) dans (3). La mesure L(T) suppose implicitement que les composantes de l'erreur ne présentent aucune corrélation entre elles. Or, Groves (1989) fournit des exemples qui démentissent cette hypothèse. On pourrait modifier la mesure de façon à inclure ces corrélations en recourant à une variante quelconque de l'écart de Mahalonobis. Au lieu de L(T), qui repose sur l'échelle de mesure originale, on pourrait utiliser :

$$L^*(T) = \sum_{1}^{k} \sigma_j + \sum_{1}^{k} |B_j| \; . \qquad (6)$$

La réduction d'un des éléments de l'erreur diminue la perte aussi bien pour L(T) que L*(T). Il y a donc amélioration de la qualité.

3.2 Interventions réalisables

Une fois la fonction de perte choisie, l'étape suivante du processus décisionnel consiste à déterminer la série d'interventions envisageables parmi lesquelles on en retiendra une, c'est-à- dire à prendre une décision. Par intervention, on entend modifier la répartition des ressources entre les différentes activités d'enquête. Adopter des bases multiples, tout en réduisant la taille de l'échantillon constitue une intervention. Les statisticiens responsables des opérations, qui dirigent le processus d'enquête, peuvent dresser la liste des activités que l'on pourrait modifier et indiquer la façon de le faire. Soit D_j, $j = 1$, ... , M où M correspond aux activités de l'enquête, et soit ΔD_j, les changements réalisables, habituellement ponctuels. Chaque activité entraîne la consommation de ressources, avec les coûts que cela suppose. Groves (1989) décrit en détail les fonctions de coût de diverses activités d'enquête. Soit C_j, le coût de l'activité D_j à son niveau actuel et ΔC_j la variation du coût amenée par la modification ΔD_j. Le coût augmente avec le niveau d'une activité et vice versa.

Une intervention réalisable correspond à une suite quelconque d'actions qui satisfait à une contrainte budgétaire. Pour un budget donné, toute série de changements

$$\Delta D = (\Delta D_1, ..., \Delta D_M) \qquad (7)$$

de sorte à ce que

$$\sum_{1}^{M} \Delta C_j \leq 0, \qquad (8)$$

est une intervention réalisable. Une intervention réalisable ne peut donc accroître le coût de l'enquête, mais bien le réduire. Les praticiens de la statistique devraient connaître le coût total des activités d'enquête. Pour prendre une décision, il leur faudra aussi savoir les coûts marginaux. Sans renseignements de ce genre, il est difficile de voir comment ils pourraient donner des conseils sur les changements à apporter. Dans le meilleur des cas, ils se borneront à défendre le statu quo.

Lorsqu'on voit le nombre des activités d'enquête, énumérer les changements envisageables paraît une tâche colossale. En réalité, quelques grandes activités auront un impact considérable sur les coûts et une incidence potentielle sur L(T). Éliminer un rappel lors d'un sondage, réduire la taille de l'échantillon en diminuant le nombre d'interviews par UPÉ ou changer la méthode de correction en sont des exemples. En se concentrant d'abord sur les principaux aspects et en modifiant les activités selon diverses grandes combinaisons de ΔD, la tâche n'est pas insurmontable, ainsi qu'on le constatera dans l'étude de cas de la partie 4.

Le cadre proposé ici est identique à celui de Fellegi et Sunter (1973), si ce n'est que ces derniers sont allés un peu plus loin en étudiant les conséquences de changements continus. Ils ont ainsi découvert les conditions nécessaires à une répartition optimale des ressources. Malheureusement, ils sont parvenus à la conclusion que ces conditions sont irréalistes et n'ont aucune valeur d'application pratique. Leurs remarques négatives ont ébranlé la simplicité fondamentale de la structure proposée au départ et semblent avoir fait avorter des recherches plus poussées. Le mieux est souvent l'ennemi du bien. À mon avis, le cadre méthodique des changements ponctuels est valable. Il pourrait donner lieu à des améliorations intéressantes à la qualité des enquêtes, si jamais il était retenu.

3.3 La décision opérationnelle

L'étape suivante consiste à évaluer l'effet des interventions réalisables, représentées par les changements ΔD sur les erreurs d'enquête. Ce facteur se rattache à la fonction de perte qu'on peut désormais écrire $L(T, \Delta D)$. On résout le problème d'affectation des ressources en choisissant l'intervention ΔD, qui minimise $L(T, \Delta D)$. Chaque intervention réalisable agit sur un sous- ensemble d'erreurs d'enquête et la difficulté que doit surmonter le praticien consiste à évaluer les effets résultants. En un mot, il doit remplir les cellules de la matrice des erreurs provenant des interventions. Les effets attribuables à une modification de la taille de l'échantillon peuvent être évalués à partir de ce que l'on sait de la structure des erreurs d'échantillonnage; l'effet d'une meilleure couverture attribuable à la modification du plan d'échantillonnage est plus difficile à déterminer, mais on pourrait toujours parvenir à une estimation grossière. Les estimations pourraient être échelonnées au moyen d'une analyse de sensibilité. Il suffirait ensuite d'évaluer les effets des erreurs de réponse et de calcul au moyen d'expériences poursuivies en cours d'enquête. Pour cela, les statisticiens responsables des opérations et les méthodologistes devront coopérer. En l'absence d'étude utilisable, le statisticien devra user de son savoir-faire pour effectuer un jugement subjectif sur l'incidence des erreurs. Une fois de plus, on pourra recourir à une analyse de sensibilité pour produire une fourchette de jugements subjectifs. On néglige parfois les erreurs attribuables à la publication des résultats en temps opportun. Tous les aspects de l'enquête demandent du temps et de l'argent. Pour l'organisme qui produit les données, mesurer le temps en années-personnes, puis convertir la valeur obtenue en une somme d'argent s'avère relativement aisé. Il n'en va pas de même pour l'utilisateur. De quelle manière l'opportunité des données agit-elle sur la fonction de perte $L(T)$? Une façon de le savoir consiste à supposer qu'au temps $t + s$, $s > 0$, l'utilisateur se servira des données recueillies au temps t (inclusivement) pour formuler une prévision $\hat{T}(t + s|t)$, sur la valeur réelle de T au temps $t + s$. L'erreur de prévision estimative peut être ajoutée à $L(T)$. Habituellement, cette erreur augmente avec s et, plus la variable évolue rapidement dans le temps, plus l'erreur est importante. Cette approche devrait donner une idée raisonnable de la pénalité associée à la production tardive des données.

Une fois que les entrées de la matrice interventions-erreurs ont été calculées, il ne reste qu'à évaluer la fonction de perte pour chaque intervention réalisable. L'intervention ou les interventions donnant la valeur la plus faible pour la fonction de perte pourront subséquemment être envisagées comme mesure susceptible d'améliorer la qualité de l'enquête. Si l'intervention entraînant la perte la plus faible doit être rejetée, on en déduit soit qu'on a sélectionné la mauvaise fonction de perte, auquel cas il faudra en choisir une autre, soit que certaines entrées de la matrice sont erronées. Celles-ci devront donc être repérées et modifiées.

La beauté de ce cadre décisionnel est qu'il permet une identification explicite des problèmes; il devient dès lors impossible de s'abriter derrière des déclarations générales sur la difficulté de changer. Par conséquent, on peut s'attaquer directement aux questions de l'importance des corrections ou des effets des

non-réponses, principales préoccupations des praticiens. La quête de la solution idéale débouchera rarement sur des résultats valables et dans la majorité des cas, une réponse approximative donnera une idée adéquate des changements les plus utiles.

4. UNE ÉTUDE DE CAS

Linacre et Trewin (1993) donnent un exemple de l'application des principes qui précèdent. Il s'agit de la seule étude de cas que j'ai retrouvée dans la littérature à présenter l'erreur d'enquête totale dans un cadre cohérent. Ces auteurs examinent comment on pourrait se servir de diverses études d'évaluation effectuées parallèlement à l'enquête australienne de 1984-1985 sur l'industrie du bâtiment pour orienter l'adaptation de cette enquête pour 1988- 1989. Selon eux, les résultats sont au mieux indicatifs et une évaluation subjective s'impose quand on ne dispose que d'informations qualitatives. Lorsqu'ils doutent de la fiabilité de leur erreur estimative, Linacre et Trewin recourent à des analyses de sensibilité. Il vaut la peine de citer l'introduction de l'étude de cas :

> *«On recourt souvent aux études d'évaluation pour trouver une façon efficace d'atténuer les autres erreurs que celle d'échantillonnage. Une question qui revient fréquemment dans la pratique a trait à la manière dont les ressources disponibles en vue d'une enquête sont attribuées aux différents volets de l'exercice. Que devrait-on investir dans l'élaboration d'un plan d'échantillonnage de bonne qualité, dans les essais-pilotes, dans un recensement sur le terrain plutôt que par la poste, dans un suivi serré des non-répondants, etc.? Chacune des tâches ayant pour but de «réduire l'erreur» requiert des ressources et la difficulté consiste à minimiser l'erreur globale d'un sondage, sous réserve d'une ou de plusieurs contraintes fixes applicables aux ressources.» [Traduction]*

L'article aborde ensuite la question du fonctionnement de l'enquête et examine diverses stratégies d'où les auteurs tirent un jeu d'options envisageables pour la nouvelle enquête. Au lieu de se restreindre aux solutions réalisables, ils évaluent le coût et la fonction de perte de chaque option. La fonction de perte retenue est la racine carrée de l'erreur quadratique moyenne ($\sqrt{EQM}$) d'un estimateur clé. Les options sont ensuite reportées sur le graphique de la figure 1, qui oppose le coût à la valeur de $\sqrt{EQM}$.

La distribution des résultats est remarquable. Les options coûtant 300 000 $ présentent la même valeur de $\sqrt{EQM}$ que celles dix fois plus chères. La solution retenue au départ réduit le coût de l'enquête au tiers du coût enregistré en 1984-1985, mais elle en double la $\sqrt{EQM}$. (Soulignons qu'en ramenant la taille de l'échantillon à 25 p. 100 de celui de 1984-1985, on doublerait aussi l'erreur d'échantillonnage.) Une autre solution réduisant le coût de l'enquête du tiers, par contre, n'entraîne qu'une légère hausse de $\sqrt{EQM}$, et est à peu près deux fois plus efficace que celle sélectionnée avant l'étude d'évaluation.

Cet exemple montre qu'une systématisation du processus décisionnel en vue de la reconception d'une enquête peut donner lieu à des changements draconiens dans le choix des options. Sans cette étude, la première solution retenue aurait été loin de la solution idéale. En outre, la gamme d'options se serait considérablement rétrécie si on n'avait tenu compte que des solutions réalisables, selon certaines contraintes économiques.

5. CONCLUSIONS

Le cadre proposé par Fellegi et Sunter et élargi dans le présent article propose un mécanisme autorisant un examen systématique des conséquences qui découlent des différentes possibilités d'affectation des ressources aux diverses activités de l'enquête. Passer par toutes les étapes du processus décisionnel n'est pas aisé, mais la difficulté d'une tâche ne signifie pas qu'on doit abandonner celle-ci. Ce n'est qu'en adoptant une telle approche qu'on parviendra à cerner les domaines qui exigent un surcroît de données ou une information de meilleure qualité. Les statisticiens devraient faire appel à leur formation professionnelle dans la gestion des travaux de statistique.

Il est intéressant de se demander pourquoi si peu de progrès ont été réalisés jusqu'à présent dans la résolution du problème de l'affectation des ressources. L'explication la plus manifeste est qu'il s'agit d'un problème fort complexe et qu'il faut vaincre une incroyable force d'inertie avant d'apporter des modifications à la plupart des régimes existants. Cependant, les enquêtes évoluent périodiquement et on coupe régulièrement les budgets. Des décisions sont donc prises qui entraînent des changements. Pourquoi a-t-on écrit si peu sur la question? Une raison semble être que les méthodologistes jouent souvent le rôle de combattants du feu, en ce sens qu'on leur demande de

corriger les analyses une fois les erreurs commises. Pour paraphraser Groves (1989), les méthodologistes mesurent les erreurs plus souvent qu'ils ne les réduisent. Bien que la chose soit compréhensible pour ceux qui s'occupent de l'analyse secondaire des données d'enquête, apparemment cette remarque s'applique aussi à bon nombre de statisticiens responsables des opérations, et la majeure partie de la documentation sur les erreurs d'enquête publiée par les statisticiens officiels a trait à l'estimation des erreurs pendant l'analyse plutôt qu'à la réduction de ces dernières au niveau théorique. Il semble également exister un écart culturel entre les responsables des opérations, qui dirigent le sondage et se penchent sur les problèmes pratiques courants des enquêtes, et les méthodologistes, qui s'intéressent davantage aux principes de la statistique. Ces deux groupes doivent se rejoindre, non seulement pour repenser les enquêtes, mais aussi pour améliorer le fonctionnement quotidien des sondages en cours. Pour cela, il faudra néanmoins attendre qu'on intègre les expériences sur l'erreur au train-train des enquêtes; alors seulement possèdera-t-on le genre d'information sur les erreurs d'enquête qui permettra d'apporter les améliorations souhaitées à la qualité des sondages.

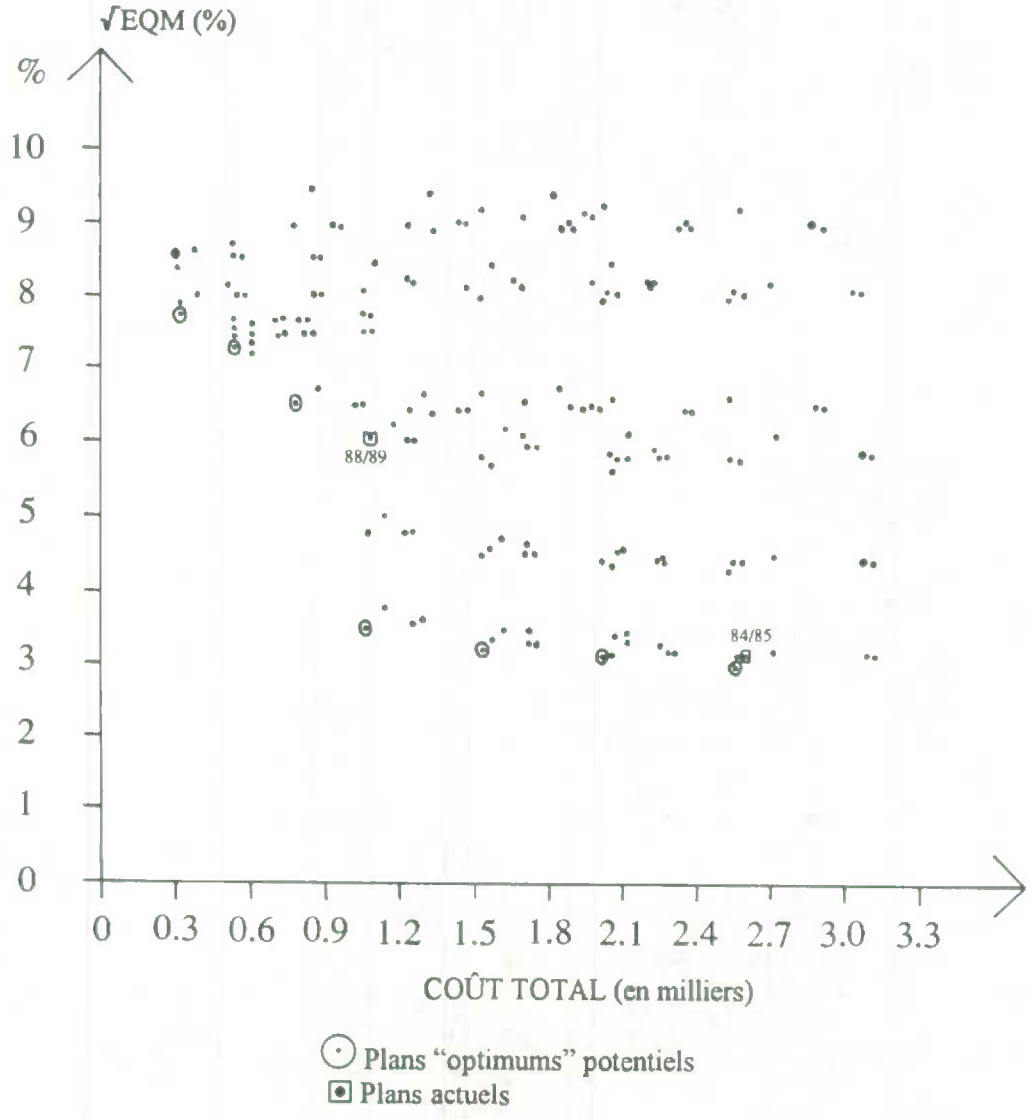


Figure 1. √EQM% v coût total pour plusieurs options d'affectation des ressources

(Reproduit, avec permission, de Linacre et Trewin (1993), Total Survey Design - Application to a Collection of the Construction Industry, *Journal of Official Statistics*, 9, 611-621, Figure 1, p. 618).

L'application de la théorie décisionnelle à l'affectation des ressources devrait s'avérer particulièrement utile pour apporter les changements voulus aux activités d'enquête, consécutivement aux modifications que subissent le budget. La théorie décisionnelle peut aussi avoir son utilité pour introduire des changements majeurs aux méthodes. Les principes d'assurance de la qualité totale nous apprennent qu'on ne peut atténuer les erreurs qu'en apportant des changements au système d'enquête, bref, en remontant le problème plus à l'avant du processus d'enquête. On pourrait soutenir que la plus importante contribution à l'amélioration des enquêtes, la contribution qui présente le meilleur potentiel quant à la réduction de toute une gamme d'erreurs, est l'avènement de l'interview assistée par ordinateur. L'IPAO et l'ITAO ont concouru à diminuer les erreurs de réponse, de correction et de codage, plus toute une série d'erreurs de traitement. Une évaluation qualitative de l'incidence de ces nouvelles méthodes sur les coûts et les pertes devrait accompagner leur mise en pratique. On pourra recourir à une analyse statistique systématique pour confirmer l'incidence des changements apportés au système.

En terminant, il vaut la peine de citer la conclusion à laquelle parviennent Linacre et Trewin :

> *«Pour réaliser d'autres progrès et minimiser l'erreur totale par une affectation rationnelle des ressources, on devra déterminer et vérifier les liens entre l'erreur et les ressources pour la gamme complète des sources d'erreur. On pourrait y arriver d'abord en identifiant les autres grands paramètres des sources d'erreur, puis en créant des modèles qui associeront ces paramètres à l'utilisation des ressources. La présente étude est un pas dans cette direction. Il faut espérer que les organes de statistique qui poursuivent des travaux sur l'affectation optimale des ressources dans le contexte de l'erreur totale signaleront leurs progrès au reste des statisticiens et qu'ainsi s'étendra notre bagage de connaissances dans un domaine relativement peu connu mais d'une importance considérable.» [Traduction]*

REMERCIEMENTS

La présente recherche n'aurait pu être menée à bien sans une subvention offerte par l'Economic and Social Research Council du Royaume-Uni, dans le cadre de son programme de recherche sur l'analyse des grandes bases de données complexes.

BIBLIOGRAPHIE

Alexander, C. H. (1994). Analyse de *Sample surveys 1975-1990: an age or reconciliation? International Statistical Review*, 62, 1, 21-28.

Fellegi, I. P., et Sunter, A. B. (1973). Balance between different sources of survey errors - some Canadian experiences, *Proceedings of the 39th Session of the ISI*, 45, 3, 334-355.

Groves, R. M. (1989). *Survey Errors and Survey Costs*, John Wiley and Sons, New York.

Hansen, M. H., Hurwitz, W. N., et Benshad, M. A. (1961). Measurement errors in censuses and surveys, *Proceedings of the 33rd Session of the ISI, 38*, 359-374.

ISI Report (1926). Report on the representative method in statistics, *Bulletin of the ISI*, 22, 1, 359-438.

Jabine, T. B., et Tepping, B. J. (1973). Controlling the quality of occupation and industry data, *Proceedings of the 39th Session of the ISI*, 45, 3, 360-389.

Kish, L. (1965). *Survey Sampling*, John Wiley & Sons, New York.

Linacre, S. J., et Trewin, D. J. (1993). Total survey design - application to a collection of the construction industry, *Journal of Official Statistics*, 9, 3, 611-621.

Morgenstern, O. (1963). On the Accuracy of Economic Observations, Princeton University Press.

Nathan, G. (1973). Utilization of information on sampling and non-sampling errors for survey design - experiences in Israel, *Proceedings of the 39th Session of the ISI*, 45, 3, 393- 406.

Statistique Canada (1987). Statistics Canada's policy on informing users of data quality and methodology, *Journal of Official Statistics*, 3, 83-92.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

RENSEIGNEMENTS SUR LA QUALITÉ DES DONNÉES FOURNIS AUX UTILISATEURS DE L'ENQUÊTE SOCIALE GÉNÉRALE DE STATISTIQUE CANADA

D. G. Paton[1]

RÉSUMÉ

Pour pouvoir interpréter correctement des données statistiques, les utilisateurs doivent être en mesure d'évaluer la qualité de ces dernières. On peut publier, parallèlement à des estimations calculées à partir des résultats d'une enquête, une mesure de la qualité de ces estimations, telle que le coefficient de variation ou un intervalle de confiance. Toutefois, dans le cas de l'Enquête sociale générale, on diffuse non seulement des estimations, mais aussi un fichier de microdonnées. La plupart des estimations fondées sur les données de l'ESG qui figurent dans les rapports et les publications sont produites à l'extérieur de Statistique Canada, au moyen des microdonnées. Le présent article décrit divers outils que l'équipe de l'Enquête sociale générale fournit aux utilisateurs des données pour préciser la qualité des estimations qu'ils produisent et les lignes directrices à suivre quand ils publient ces estimations.

MOTS CLÉS : Qualité des données; effets du plan de sondage.

1. INTRODUCTION

Pour interpréter correctement les données statistiques, l'utilisateur a besoin de mesures de leur qualité. Le présent article donne un aperçu des moyens utilisés pour informer les utilisateurs des données de l'Enquête sociale générale (ESG) de la qualité de ces dernières. On commence par décrire brièvement l'ESG, puis on examine la politique de Statistique Canada visant à informer les utilisateurs sur la qualité des données. On décrit ensuite les méthodes particulières suivies pour informer les utilisateurs des données de l'ESG, en les comparant à celles d'autres diffuseurs de données d'enquête. Enfin, on envisage certaines modifications qui pourraient être apportées aux pratiques de l'équipe de l'ESG.

Enquête sociale générale

L'Enquête sociale générale (ESG) est une enquête-ménage que Statistique Canada effectue annuellement depuis dix ans. Elle a pour objectif, d'une part, de collecter des données sur un large éventail de caractéristiques sociales, telles que la santé, l'utilisation du temps, les antécédents familiaux, la victimisation et le niveau de scolarité, et, d'autre part, de diffuser les données sous forme agrégée et sous forme de fichier de microdonnées destinées à être analysées par les fonctionnaires, les universitaires, et d'autres organismes et particuliers intéressés.

La population cible de l'ESG englobe la population ne vivant pas en établissement des 10 provinces du Canada. Les données sont collectées par téléphone auprès d'un échantillon sélectionné selon la méthode de composition aléatoire et auprès d'échantillons complémentaires de populations spéciales présentant un intérêt particulier, parfois sélectionnés à partir de listes. D'autres renseignements sur la méthodologie de l'ESG figurent dans Norris et Paton (1991).

Les données de l'ESG sont diffusées de plusieurs façons. Une gamme de résultats particulièrement importants et intéressants sont publiés dans *Le Quotidien*, publication quotidienne de Statistique Canada où sont annoncés les résultats des enquêtes et les dates de diffusion des données. Statistique Canada a publié un rapport détaillé décrivant en détail les résultats de l'ESG pour plusieurs années de référence. Cependant, les fichiers de microdonnées sont le mode le plus important de diffusion des données de l'ESG.

[1] David G. Paton, Division des méthodes d'enquêtes des ménages, Statistique Canada, Ottawa (Ontario), K1A 0T6.

Politique de Statistique Canada

L'ESG étant une enquête de Statistique Canada, les membres de l'équipe de projet diffusent les données et informent les utilisateurs conformément aux politiques de Statistique Canada. Une des sections du Manuel des politiques de Statistique Canada a pour titre «Politique visant à informer les utilisateurs sur la qualité des données et la méthodologie» (Statistique Canada, 1992). Dans l'énoncé de cette politique, le Bureau reconnaît que les utilisateurs de produits statistiques doivent obtenir des renseignements supplémentaires pour pouvoir se servir des données comme il convient et les interpréter correctement, et qu'il lui incombe donc de fournir cette information.

L'énoncé de la politique de Statistique Canada est le suivant :

1. Statistique Canada fournira aux utilisateurs des indicateurs de la qualité des données diffusées et des descriptions des concepts, des définitions et des méthodes utilisés.

2. Il faut joindre aux produits statistiques de la documentation relative à la qualité et à la méthodologie ou prévoir d'y inclure un renvoi à celle-ci.

3. La documentation relative à la qualité et à la méthodologie doit être conforme aux normes et lignes directrices émises, à intervalles irréguliers, aux termes de la présente politique.

4. Dans des circonstances spéciales, il est possible d'obtenir l'autorisation de se soustraire aux exigences de la présente politique en procédant de la façon dont il est fait état au paragraphe ci-dessous traitant des responsabilités.

5. Il faut faire connaître aux promoteurs d'enquêtes effectuées contre remboursement des frais et de projets de consultation statistique à l'issue desquels Statistique Canada ne publiera aucune donnée, les sections pertinentes des normes et lignes directrices émises en vertu de la présente politique et les inciter à les respecter.

Les normes et lignes directrices publiées aux termes de cette politique permettent de dégager trois grandes composantes de l'information supplémentaire jugée essentielle à l'interprétation et à l'utilisation correcte de toute donnée :

1) Que devrait représenter les données? Quels sont les concepts présentant un intérêt ou pertinents en regard des politiques?

2) Comment les données de l'enquête représentent-elles ces concepts? Dans quelle mesure les données s'écartent-elles des concepts présentant un intérêt, en raison du plan de sondage?

3) Quels sont l'erreur et le degré d'incertitude associés aux données? Dans quelle mesure les estimations pourraient-elles s'écarter accidentellement de certaines valeurs réelles?

La première de ces composantes correspond aux concepts et définitions qui, en vertu des normes et des lignes directrices, doivent être fournis en même temps que les données et qui, plus précisément, incluent la description de l'enquête, de la population cible, de la période de référence et du questionnaire utilisé.

La deuxième composante correspond, en grand partie, aux précisions méthodologiques dont le compte rendu est recommandé dans les normes et lignes directrices, y compris la méthode de collecte des données, le plan d'échantillonnage et les méthodes de saisie, de traitement, de vérification, d'imputation, de pondération et d'estimation.

Pour fournir les renseignements correspondant à la troisième composante, à savoir l'erreur et le degré d'incertitude, il faut décrire les diverses sources d'erreur et d'incertitude, et donner des mesures de leur amplitude. À cet égard, il convient de tenir compte de l'erreur de couverture, de l'erreur d'échantillonnage, de l'erreur non due à l'échantillonnage et des erreurs de réponse.

Le cas de l'enquête sociale générale

L'ESG donne lieu à deux diffusions de données importantes. La première, effectuée dans *Le Quotidien*, est assez brève (le communiqué du 6 juin 1995 comptait environ deux pages (Statistique Canada, 1995)), et vise surtout à exposer quelques faits saillants qui se dégagent des données de l'ESG et à annoncer la diffusion des données. L'autre correspond à la publication du fichier de microdonnées à grande diffusion. Ce dernier contient des données sur un grand nombre de variables (400) pour chaque répondant de l'ESG (qui, typiquement, en compte environ 10 000) et est accompagné de 300 à 400 pages de documentation (par exemple, Statistique Canada, 1994). Compte tenu de sa brièveté, l'article sur l'ESG publié dans *Le Quotidien* donne peu de détails sur la qualité et sur le degré d'incertitude des données et des estimations. Un petit encadré d'environ

30 centimètres carrés contient quelques phrases au sujet des concepts, de la population cible, de la période de référence, de la méthode de collecte des données, de la taille de l'échantillon et du taux de réponse de l'enquête. Rien n'y indique que les résultats présentent un certain degré d'incertitude.

Le fichier de microdonnées de l'ESG, qui est le principal produit tiré de cette enquête, contient le fichier des données et de la documentation très détaillée où sont abordées nombre de questions relatives à la qualité des données. On y décrit en détail la population cible et on y précise la période de référence. La documentation contient aussi l'énoncé précis des questions posées, ainsi que des références à d'autres enquêtes fondées sur les mêmes questions. Elle inclut également la description de la méthode de collecte des données et des méthodes de traitement, de vérification et d'imputation. Elle contient la description du plan de sondage, et des précisions sur ce qui distingue la population échantillonnée de la population cible. La méthode de calcul des poids des estimations y est décrite en détail. Enfin, on y mentionne l'erreur de couverture, l'erreur due à la non-réponse et d'autres sources d'erreur.

La documentation ne donne toutefois aucun renseignement précis sur les estimations et sur la variabilité de ces dernières, parce que les détails du calcul des estimations varient selon l'utilisateur des microdonnées et que la variabilité d'une estimation dépend de la méthode suivie pour calculer cette dernière. Pour faciliter la tâche des utilisateurs du fichier, la documentation fournie avec ce dernier contient des lignes directrices concernant le calcul des estimations, ainsi que l'estimation de la variabilité des résultats.

Les lignes directrices concernant le calcul des estimations sont très importantes, puisque le calcul de la variabilité d'une estimation présente peu d'intérêt si l'estimation ne reflète pas le concept qu'on veut étudier. La documentation sur les microdonnées de l'ESG contient des instructions quant à l'utilisation du fichier de données, et des variables et s de pondération inclus dans le fichier, ainsi que des exemples détaillés de leur utilisation. Dans la documentation sur les microdonnées du Cycle 8 (Statistique Canada, 1994), 15 pages sont consacrées à ces lignes directrices et à ces exemples.

Idéalement, les utilisateurs de microdonnées devraient être capables de calculer leurs propres mesures de l'incertitude des estimations qu'ils produisent. Malheureusement, les renseignements précis sur le plan de sondage de l'ESG nécessaires pour calculer ces mesures incluent des données géographiques qui ne peuvent figurer dans le fichier à grande diffusion pour des raisons de confidentialité. Donc, plusieurs lignes directrices sont données aux utilisateurs du fichier de microdonnées pour leur permettre d'appliquer d'autres méthodes de calcul de l'incertitude des estimations qu'ils produisent. Ces lignes directrices englobent les effets du plan de sondage, les tables de variance approximative, les lignes directrices concernant la taille minimale de l'échantillon et les lignes directrices concernant l'ajustement des poids dans certaines situations analytiques. En outre, Statistique Canada offre de calculer, contre remboursement des coûts, les estimations de la variabilité de toute estimation qui intéresse les utilisateurs.

Effets du plan de sondage

Bien que les effets du plan de sondage soient particuliers à chaque estimation, il est parfois utile de déterminer un effet type qui permet de corriger les variances calculées en se fondant sur l'hypothèse d'un échantillonnage aléatoire simple. La variance approximative des estimations calculées de cette façon sera soit une sous-estimation, soit une surestimation de la variance réelle, selon que l'effet réel du plan de sondage sur l'estimation est plus grand ou plus petit que l'effet type utilisé. Donc, la valeur choisie pour l'effet type du plan de sondage doit représenter un compromis entre une valeur trop petite, qui mènerait à considérer bon nombre d'estimations comme plus précises qu'elles ne le sont réellement et une valeur trop grande, qui porterait à juger un trop grand nombre d'estimations comme insuffisamment précises.

Dans le cas de l'ESG, pour choisir l'effet type du plan de sondage, on calcule l'effet réel pour un grand nombre, généralement supérieur à 200, de variables du fichier de microdonnées. L'effet du plan de sondage qui est publié correspond au 75^{e} percentile de la distribution résultante. On choisit cette méthode relativement prudente, car il n'est pas possible de prédire exactement les estimations que calculeront les analystes. Si on choisissait le 50^{e} percentile, les effets du plan de sondage publiés seraient 7 à 20 % plus faibles (voir le tableau 1).

Tables de variances approximatives

Les tables de variances approximatives fournissent des estimations de la variance, tenant compte de l'effet du plan de sondage, de la taille de la population et de la taille de l'échantillon, pour les estimations des totaux et des proportions. L'équipe de l'ESG fournit ces tables avec des instructions d'utilisation pour estimer le coefficient de variation des ratios, des écarts et des écarts entre ratios des estimations, et pour calculer les limites de confiance et faire le test de *t*.

Tableau 1. Effets du plan de sondage pour le 50e percentile et le 75e percentile, ESG-9.

	50%	75%
CANADA	1,42	1,53
T.-N.	1,08	1,27
I.-P.-E.	0,98	1,25
N.-E.	1,05	1,26
N.-B.	1,19	1,38
QC	1,17	1,27
ONT.	1,11	1,25
MAN.	1,06	1,25
SASK.	1,03	1,24
ALB.	1,14	1,27
C.-B.	1,13	1,26

Lignes directrices concernant la taille minimale de l'échantillon

Le coefficient de variation d'une faible proportion, estimé au moyen d'un échantillon aléatoire simple est approximativement une fonction simple, du nombre de répondants qui contribuent au numérateur de la proportion, soit :

$$cv(p)=\sqrt{fpc}\sqrt{\frac{1}{n*}}$$

(où n^* est le nombre de répondants qui contribuent au numérateur de la proportion et *fpc* est le facteur de correction d'échantillonnage pour une population finie).

De cette relation, on tire une règle empirique permettant de généraliser les résultats de l'échantillon aléatoire simple par application d'un effet de plan de sondage qui permet de produire le tableau 2.

On recommande dans la documentation sur les microdonnées de ne pas diffuser ou publier les estimations fondées sur moins de 15 observations. Cette limite est établie pour l'ESG d'après un effet du plan de sondage d'environ 1,5 et l'utilisation d'un cv maximal permis de 33 %.

Facteur d'extrapolation de la variance

On peut concevoir le facteur d'extrapolation de la variance comme le facteur dont est extrapolée la variance de l'estimation du total d'une variable donnée, en raison de la variabilité des poids, quand ces derniers et le plan de sondage sont sans relation avec la variable. Il s'agit du facteur «1+L» de Kish (1992).

Tableau 2. Taille minimale de l'échantillon, selon le coefficient de variation (cv) et l'effet du plan de sondage

C.V. (%)	Effet du plan de sondage		
	1,0	1,5	2,0
10	100	150	200
14	49	74	98
17	36	54	72
20	25	38	50
25	16	24	32
33	9	15	18

Comme l'indique le tableau 3, ce facteur tend à extrapoler les variances calculées pour les données de l'ESG encore plus que celui tenant compte de l'effet du plan de sondage. Son application évite donc de conclure que des résultats sont significatifs alors que les données ne le justifient pas, quand on estime des totaux ou les paramètres d'une régression linéaire, mais non quand on calcule la variable chi carré pour des tableaux de contingence.

Tableau 3. Facteurs d'extrapolation de la variance (FGV) et effets du plan de sondage (EPS), ESG-9.

	EPS 75 %	FGV
CANADA	1,53	1,58
T.-N.	1,27	1,28
Î.P.-É.	1,25	1,28
N.-É.	1,26	1,32
N.-B.	1,38	1,43
QC	1,27	1,28
ONT.	1,25	1,26
MAN.	1,25	1,27
SASK.	1,24	1,37
ALB.	1,27	1,29
C.-B.	1,26	1,28

Ajustement des poids

Cette technique facilite les analyses effectuées au moyen de programmes statistiques ou informatiques mal adaptés à l'utilisation de données d'enquête, particulièrement dans les situations où, pour les algorithmes utilisés, l'application de grands poids implique une précision élevée. La méthode consiste, dans de tels cas, à reproportionner les poids utilisés dans

l'analyse de sorte que leur moyenne soit égale à 1. On effectue cette transformation en divisant chaque coefficient par la moyenne des coefficients. Le recours à cette méthode est utile dans les situations où la transformation des poids ne fait pas varier les estimations, mais fait varier les variances calculées par le programme analytique (si les variances ne changent pas, la transformation des poids n'a aucun effet néfaste.) La situation la plus courante est celle de l'analyse des tableaux de contingence. Quand ils contiennent des proportions plutôt que des dénombrements, ces tableaux ne dépendent pas de l'échelle des poids.

Changements à considérer

En préparant le présent article et en examinant les pratiques actuelles concernant l'ESG, il m'a paru évident que certains changements devraient être apportés ou envisagés. Certaines sections de la documentation devraient fournir plus de détails. Il serait également souhaitable de demander aux utilisateurs de préciser quels renseignements leur seraient utiles et quelles précisions supplémentaires sont nécessaires.

À l'heure actuelle, la documentation décrit comment calculer diverses estimations. Elle pourrait être élargie de façon à offrir des échantillons de programmes dans un ou plusieurs langages de programmation statistique et, même, des échantillons de programmes destinés à exécuter des tâches plus compliquées, telles que des régressions linéaires ou logistiques.

La justification théorique de la règle empirique de la taille minimale de l'échantillon est faible. Il conviendrait donc de déterminer le degré de validité de cette règle et d'essayer de renforcer la justification théorique.

On devrait examiner la possibilité de diffuser un plan de sondage simplifié. Si on pouvait saisir dans un tel plan la plupart de la variabilité d'échantillonnage et diffuser ce plan sans compromettre la confidentialité des données, les utilisateurs pourraient effectuer eux-mêmes des analyses significatives fondées sur le plan de sondage. Pour faciliter ces analyses, Statistique Canada devrait diffuser des programmes échantillons et recommander des outils analytiques.

BIBLIOGRAPHIE

Kish, L. (1992). Weighting for unequal P_i, *Journal of Official Statistics*, 8(2), 183-200.

Norris, D. A., et Paton, D.G. (1991). L'enquête sociale générale canadienne: bilan des cinq premières années, *Techniques d'enquête*, 17, 2, 245-260.

Statistique Canada, (1992). Informing Users of Data Quality and Methodology. Policy (3pp) and Standards and Guidelines (16pp), uncatalogued, part of Statistics Canada Policy Manual.

Statistique Canada (1994). L'enquête sociale générale de 1993 - cycle 8. Les risques auxquels est exposée une personne, Documentation sur le fichier de microdonnées à grande diffusion et Guide de l'utilisateur, (429pp), non-catalogué.

Statistique Canada (1995). Enquête sociale générale: l'ordinateur en milieu de travail. *Le Quotidien*, le 6 juin 1995, 5-6. Statistique Canada, No. 11-001F au catalogue.

SESSION 5

Aspects techniques de la confidentialité

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

RECODAGES GLOBAUX ET SUPPRESSIONS LOCALES DANS LES ENSEMBLES DE MICRODONNÉES[1]

A.G. de Waal et L.C.R.J. Willenborg[1]

RÉSUMÉ

Statistics Netherlands utilise deux méthodes pour assurer la protection du secret statistique des ensembles de microdonnées : la suppression locale et le recodage global. La suppression locale consiste à éliminer certaines des valeurs de certains dossiers en indiquant «données manquantes». Le recodage global est une opération par laquelle on procède au recodage de certaines variables. Idéalement, les suppressions locales et les recodages globaux devraient être déterminés automatiquement et d'une façon optimale; autrement dit, la perte d'information due à ces opérations devrait être réduite au minimum. Dans le présent article, nous nous penchons sur trois problèmes différents : détermination des suppressions locales optimales lorsque les ensembles de microdonnées doivent être protégés uniquement au moyen des suppressions locales; détermination des recodages globaux optimaux lorsque cette méthode est celle retenue pour la protection des ensembles de microdonnées; détermination des suppressions locales et des recodages globaux optimaux lorsqu'il convient, pour protéger les ensembles de microdonnées, d'utiliser une combinaison des deux méthodes. Aucune mesure complexe de l'information n'est nécessaire dans le cas de la suppression locale. Plusieurs formules de programmation par nombres entiers 0-1 sont proposées, selon les objectifs de protection du secret statistique. En ce qui concerne le recodage global et la combinaison de la suppression locale et du recodage global, il convient d'utiliser une mesure élaborée de l'information. Dans le présent article, nous proposons une méthode fondée sur une mesure entropique appropriée. Nous proposons par ailleurs une description détaillée du problème du recodage global et de celui de la combinaison du recodage global et de la suppression locale (RG-SL).

MOTS CLÉS : Protection du secret statistique; microdonnées; suppression locale; recodage global; optimisation; problèmes de programmation par nombres entiers 0-1.

1. INTRODUCTION

La suppression locale et le recodage global sont deux méthodes biens connues de protection des microdonnées contre la divulgation. La suppression locale consiste à éliminer un certain nombre de valeurs de certaines variables contenues dans certains enregistrements de microdonnées en indiquant «données manquantes». Le recodage global est une opération par laquelle on procède au recodage d'un certain nombre de variables. Il s'agit de fusionner un certain nombre de catégories d'une variable donnée pour former de nouvelles catégories. Par exemple, lorsque les catégories «Veuf» et «Divorcé» de la variable «État civil» sont combinées pour former la catégorie «Veuf ou Divorcé», il s'agit d'un recodage de la variable «État civil». Dans le cas du recodage global, tous les enregistrements dans lesquels on trouve ces catégories sont touchés, tandis que dans le cas de la suppression locale, seuls certains enregistrements particuliers sont visés. La suppression locale et le recodage global sont habituellement utilisés en combinaison, mais ils peuvent également s'utiliser séparément. Les deux méthodes visent à produire des fichiers de données contenant moins d'informations détaillées, ce qui réduit d'autant les risques de divulgation.

Idéalement, il serait utile de déterminer les recodages globaux et les suppressions locales d'une façon automatique et optimale, c'est-à-dire, de réduire au minimum la perte d'information due à la mise en oeuvre

---

[1] A.G. de Waal et L.C.R.J. Willenborg, Statistics Netherlands, Prinses Beatrixlaan 428, P.O. Box 959, 2270 AZ Voorburg, Pays-Bas.

Les opinions exprimées dans le présent article sont celles des auteurs et ne reflètent pas nécessairement les politiques de Statistics Netherlands.

de ces méthodes de protection du secret statistique (*statistical disclosure control* ou SDC). Dans le présent article, nous nous penchons sur trois problèmes distincts : le problème de la protection d'un ensemble de microdonnées par un recours à la suppression locale uniquement, celui de la protection d'un ensemble de microdonnées par le recours au recodage global uniquement et celui de la protection d'un ensemble de microdonnées par le recours à une combinaison du recodage global et de la suppression locale (RG-SL).

La suppression locale ne nécessite pas de mesures complexes de la perte d'information. Il est possible de compter le nombre de valeurs supprimées localement. Cette méthode primitive, mais simple, de mesurer la perte d'information ne tient pas comptes dans la suppression locale des valeurs, des variables différentes. Le faire pourrait conduire à des mesures différentes de perte d'information. Grâce à cette mesure directe de la perte d'information due aux suppressions locales, le problème de la suppression locale peut être assimilé à un problème de programmation par nombres entiers 0-1. Évidemment, dans certains cas, la mesure directe de la perte d'information due aux suppressions locales risque de ne pas être appropriée. Dans de tels cas, il faudra peut-être recourir à des mesures plus complexes comme celle que nous aborderons plus loin et qui est fondée sur l'entropie. Dans la section 2 du présent article, nous abordons le problème de la suppression locale. Nous nous penchons en particulier sur un aspect clé du problème de la suppression locale : la combinaison problématique minimale. Cette notion nous permet d'assimiler le problème de la suppression locale à un problème de programmation par nombres entiers 0-1. Cette question fait l'objet de la section 3. Dans la section 4, nous examinons certaines questions connexes liées au problème de la suppression locale.

Le recours à une mesure de la perte d'information semble être un élément inévitable de la détermination du recodage global optimal. Une telle mesure est requise, en particulier, lorsqu'il s'agit de déterminer la combinaison optimale de suppressions locales et de recodages globaux. L'entropie nous paraît être ici un choix naturel. Par contre, le choix d'un modèle de probabilité propice permettant de prendre en compte l'absence (partielle) des informations due aux recodages globaux ou aux suppressions locales paraît moins évident. Nous proposons un modèle simple à utiliser dans de telles situations. Nous présentons, dans la section 5, une méthode de mesure de l'information fondée sur l'entropie.

Le problème du recodage global et celui de la combinaison RG-SL nous paraissent beaucoup plus difficiles à formaliser que le problème de la suppression locale. C'est la raison pour laquelle ces deux problèmes seront abordés plus superficiellement. On n'a pas encore réussi à des déterminer des formules appropriées de programmation par nombres entiers 0-1 pour ces deux problèmes, ni à les résoudre. Dans la section 6 du présent article, nous proposons une entrée en matière portant sur le problème du recodage global et de la combinaison RG-SL. Le problème du recodage global optimal à l'état pur fait l'objet de la section 7. L'objectif consiste à déterminer les recodages globaux qui permettent d'éliminer un ensemble donné de combinaisons rares tout en réduisant au minimum la perte d'information. Dans la section 8, nous présentons une description du problème de la combinaison RG-SL. Finalement, dans la section 9, nous concluons cet examen par une brève discussion.

Le présent article combine l'article de De Waal et Willenborg (1994), consacré entièrement au problème de la suppression locale, à celui de De Waal et Willenborg (1995b), qui porte sur le problème du recodage global et sur celui de la combinaison RG-SL.

2. PRÉSENTATION DU PROBLÈME DE LA SUPPRESSION LOCALE

Les règles de SDC décrivent souvent des combinaisons de catégories de variables repères qui doivent être vérifiées avant la diffusion d'un ensemble de microdonnées. En outre, les règles précisent également la valeur de la fréquence à partir laquelle ces combinaisons sont jugées aptes à être diffusées. Dans les cas où la fréquence d'une combinaison particulière est au moins égale à une valeur limite prescrite, on jugera cette combinaison sûre. Dans les autres cas, la combinaison sera jugée problématique, et des mesures de SDC devront être appliquées. Dans la présente section ainsi que dans les sections 3 et 4, nous présumons que ces précautions consisteront à remplacer certaines valeurs par des «données manquantes», c'est-à-dire à recourir à une suppression locale de ces valeurs.

Le moyen le plus facile de déterminer quelles sont les valeurs des variables à supprimer localement consisterait à procéder ainsi pour chaque combinaison qui doit être vérifiée et pour chaque enregistrement individuel. Ceci peut être fait de deux façons. Premièrement, lorsqu'une valeur est localement supprimée, on la remplace immédiatement par une «donnée manquante». La microdonnée ainsi obtenue est alors utilisée pour déterminer si une combinaison donnée est sûre ou problématique. Deuxièmement, l'ensemble

original de microdonnées peut être utilisé pour déterminer si une combinaison est sûre ou non. Toutefois, les deux méthodes posent des problèmes.

Lorsqu'on utilise la première méthode, certaines combinaisons peuvent sembler à tort ne pas apparaître avec une fréquence suffisante. Par exemple, si nous supprimons la valeur «Boulanger» de la combinaison «Boulanger» x «Étranger» apparaissant dans un enregistrement, cette opération pourrait avoir pour conséquence, lors d'opérations ultérieures, de laisser conclure à une fréquence insuffisante de la combinaison «Boulanger» x «Homme». Cette dernière combinaison risquera donc d'être jugée problématique, même si elle peut apparaître avec une fréquence suffisante dans l'ensemble original de microdonnées. En fait, dans un tel cas, cette combinaison devrait être considérée sûre.

Par contre, lorsqu'on utilise la deuxième méthode pour déterminer la sûreté d'une combinaison et qu'on le fait séparément pour chaque enregistrement, on risque également de rencontrer des difficultés. Supposons, par exemple, que la combinaison «Boulanger» x «Étranger» n'apparaît pas assez fréquemment dans le dossier original et qu'on décide de supprimer localement l'information «Étranger». Supposons en outre que la combinaison «Boulanger» x «Femme» n'apparaît pas non plus assez fréquemment et qu'on décide, dans ce cas, de supprimer localement l'information «Femme». Il n'est pas impossible, dans un tel cas, que la suppression locale ait été exagérée. Au cas où il pourrait exister des personnes qui sont à la fois «Boulangères», «Femmes» et «Étrangères», il aurait mieux valu que l'on choisisse de supprimer localement la catégorie «Boulanger» pour ces personnes, en présumant que les catégories «Étranger» et «Femme» apparaissaient avec une fréquence suffisante. Le nombre de suppressions locales aurait été moindre si on avait supprimé localement l'information «Boulanger» pour ces personnes.

Nous pouvons conclure de ce qui précède qu'il n'est pas possible de décider séparement, pour chaque combinaison problématique et pour chaque enregistrement, quelles sont les valeurs qui devraient être supprimées si nous souhaitons réduire au minimum le nombre de suppressions locales. Il faut décider simultanément quelles sont les valeurs à supprimer localement pour l'ensemble des combinaisons problématiques et des enregistrements.

Pour en arriver à une décision, nous supposons qu'il est nécessaire de vérifier si certaines combinaisons trivariées de catégories de variables repères apparaissent avec une fréquence suffisante[2]. On commence par une vérification univariée de toutes les variables repères. Lorsque la catégorie d'une variable est jugée sûre, on procède ensuite à la vérification des combinaisons bivariées dont elle fait partie. Si une catégorie de variables n'apparaît pas avec une fréquence suffisante (p. ex., «Maire»), il devient alors inutile de vérifier les combinaisons bivariées dans lesquelles cette variable apparaît (p. ex., «Maire» x «Femme»). Il s'agit ensuite de vérifier les combinaisons trivariées dans lesquelles apparaissent uniquement les combinaisons bivariées sûres. Lorsqu'une combinaison bivariée problématique apparaît dans une combinaison trivariée, cette dernière n'a pas à être vérifiée. Par exemple, si la combinaison bivariée «Boulanger» x «Femme» est problématique, il est alors inutile de vérifier la combinaison trivariée «Boulanger» x «Femme» x «Urk»[3]. Après vérification des combinaisons trivariées requises, nous sommes en mesure d'énumérer pour l'ensemble des enregistrements les combinaisons univariées, bivariées et trivariées problématiques. Nous désignerons ci-après les combinaisons de cette liste par l'expression «*combinaisons problématiques minimales*».

La méthode décrite ci-haut a notamment pour conséquence de faire en sorte que lorsqu'on supprime localement une valeur dans une combinaison n-variée problématique minimale, la combinaison (n-1)-variée qui en découlera sera sûre. Cette propriété des combinaisons problématiques minimales facilite la détermination du nombre minimal de suppressions locales.

Pour conclure la présente section, nous voudrions souligner qu'il n'est pas essentiel que la valeur limite utilisée pour déterminer si une combinaison est sûre ou non soit un nombre fixe. En fait, la valeur limite risque de dépendre de la combinaison. Par exemple, la valeur limite d'une combinaison univariée pourrait être plus élevée que la valeur limite d'une combinaison bivariée. Dans ce cas, il pourra arriver qu'une combinaison bivariée «Maire» x «Femme» soit jugée sûre tandis que la variable «Maire» est jugée problématique lorsqu'elle est envisagée isolément. Dans une telle situation, la variable

---

[2] Nous nous intéressons uniquement aux variables repères puisque nos mesures de protection d'un ensemble de microdonnées portent uniquement sur ce type de variables. Lorsque nous parlons d'une catégorie de variables, il s'agit d'une catégorie de variables repères.

[3] Urk est un pittoresque village de pêcheurs des Pays-Bas.

«Maire» sera toujours assimilée à une combinaison problématique minimale. Les formules de programmation par nombres entiers 0-1 des sections 3 et 4 demeurent valides dans un tel cas.

3. RÉDUIRE AU MINIMUM LE NOMBRE DE SUPPRESSIONS LOCALES

Le premier problème que nous examinons ici est celui de la détermination d'un nombre minimal de suppressions locales qui permettront d'assurer la sûreté de l'ensemble de microdonnées qui en découlera. Ce problème peut être formalisé comme suit. Imaginons qu'il faille supprimer certaines catégories de variables dans certains enregistrements. Pour chaque catégorie j d'une combinaison problématique minimale de l'enregistrement i, nous introduisons une variable factice Y_{ij}. Cette variable factice est égale à 0 si la catégorie j de l'enregistrement i n'est pas supprimée ou si la catégorie j n'existe pas dans l'enregistrement i, autrement, elle est égale à 1. Pour chaque combinaison problématique minimale et pour chaque enregistrement, il faut tenir compte de la contrainte suivante : il faut supprimer au moins une catégorie d'une combinaison problématique minimale dans un enregistrement. En d'autres mots, la somme des Y_{ij} des catégories correspondantes j est égale à au moins 1. Comme nous l'avons signalé auparavant, cette contrainte est nécessaire et suffisante pour assurer la sûreté de cette combinaison. La somme pondérée des Y_{ij} nous sert de fonction cible.

En termes mathématiques, le problème de programmation par nombres entiers 0-1 se pose comme suit. Désignons le nombre total d'enregistrements problématiques par I, et le nombre total de catégories des combinaisons problématiques par J. Après la renumérotation des enregistrements et des variables, les variables factices $Y_{ij} (i = 1, ..., I; j = 1, ..., J)$ doivent satisfaire à

$$y_{ij} = \begin{cases} 1 & \text{si la catégorie } j \text{ de l'enregistrement } i \text{ est supprimée,} \\ 0 & \text{si la catégorie } j \text{ de l'enregistrement n'est pas supprimée} \\ & \text{ou si la catégorie } j \text{ n'existe pas dans l'enregistrement } i. \end{cases} \quad (1)$$

Supposons qu'il existe K combinaisons problématiques minimales dans l'ensemble de microdonnées. Désignons par ailleurs un c_{jk} égal à 1 si la catégorie j existe dans la combinaison problématique minimale $k (j = 1, ..., J; k = 1, ..., K)$ et égal à 0 autrement. Les contraintes du problème sont déterminées par la formule suivante

$$\sum_{j=1}^{J} c_{jk} y_{ij} \geq d_{ik}, \text{ pour tous les } i = 1, ..., I; k = 1, ..., K, \quad (2)$$

où d_{ik} est égal à 1 si la combinaison problématique minimale k apparaît dans l'enregistrement I, et égal à 0 dans les autres cas. Les contraintes déterminées par la formule (2) doivent s'appliquer puisque au moins une catégorie de chacune des combinaisons problématiques minimales doit être supprimée.

Imaginons maintenant la fonction cible suivante

$$\sum_{i=1}^{I} \sum_{j=1}^{J} w_{ij} y_{ij}, \quad (3)$$

où w_{ij} désigne le poids non négatif de la catégorie j de l'enregistrement I qui doit être précisé par l'utilisateur. Notre problème consiste à minimiser la fonction cible (3), compte tenu des contraintes déterminées en (2).

À noter que si nous choisissons de faire en sorte que tous les poids w_{ij} soient égaux à un, notre objectif consistera alors à réduire au minimum le nombre de suppressions locales. Comme les poids de la fonction cible (3) peuvent être assimilés arbitrairement à des nombres non négatifs, le problème décrit ci-haut est plus général. Les poids nous permettent de faire une différence entre l'importance relative de catégories spécifiques contenues dans des dossiers spécifiques, pour autant qu'on s'intéresse à la suppression locale.

Le problème décrit ci-haut peut être résolu en utilisant un algorithme standard de résolution des problèmes de programmation par nombres entiers 0-1;IP par exemple, un algorithme de séparation et évaluation (voir Nemhauser et Wolsey, 1988). Par ailleurs, le problème peut être réduit en un certain nombre de problèmes plus petits. Il peut avant tout être décomposé en sous-problèmes correspondant à chaque enregistrement distinct. Pour chaque enregistrement I, la fonction cible (3) doit être remplacée par la fonction cible

$$\sum_{j=1}^{J} w_{ij} y_{ij}, \quad (4)$$

Les contraintes à prendre en compte pour ce problème sont toutes celles indiquées en (2) pour autant qu'elles se rapportent à l'enregistrement I. Même ce sous-problème correspondant à chaque enregistrement peut parfois être subdivisé en un nombre de sous-problèmes plus petits. Imaginons que les combinaisons problématiques minimales d'un enregistrement

particulier correspondent aux sommets d'un graphique. Si deux combinaisons problématiques minimales partagent la même catégorie, elles seront alors jointes en une arête. Un tel graphique peut être déconnecté. Dans un tel cas, il est constitué de plusieurs sous-graphiques connectés qui sont mutuellement déconnectés. Chaque sous-graphique correspond à un sous-problème, c'est-à-dire le problème de la minimisation de (4), compte tenu des contraintes voulant que les combinaisons problématiques minimales correspondant aux sommets du graphique soient rendues sûres. Ainsi, nous serons parfois en mesure de réduire le problème original en un certain nombre de sous-problèmes plus petits. Toutefois, même ces sous-problèmes pourront parfois être réduits en problèmes encore plus petits dont certains pourront être triviaux. Cette réduction supplémentaire découle de l'observation selon laquelle seules les variables factices correspondant aux catégories qui apparaissent dans plus d'une combinaison problématique minimale doivent être considérées. Des combinaisons qui sont toujours problématiques après la suppression (optimale) de certaines de ces catégories pourront être rendues sûres en fonction d'une liste de priorités. Aucun problème d'optimisation n'a à être résolu pour les combinaisons qui demeurent problématiques, puisqu'elles ne partagent pas une catégorie commune.

Nous proposons ci-après certains exemples qui rendront cette observation plus claire. Supposons qu'on souhaite réduire au minimum le nombre de suppressions locales dans un enregistrement particulier. Si aucune des combinaisons problématiques minimales ne possède une catégorie en commun, il n'y aura pas de problème d'optimisation. Pour chaque combinaison problématique minimale d'un tel enregistrement, il s'agit de supprimer une valeur unique que l'on peut choisir arbitrairement. En pratique, les valeurs à supprimer pourraient être déterminées au moyen d'une liste de priorités. Cette liste doit être fournie par l'utilisateur. Voici à quoi elle pourrait ressembler : supprimer d'abord une catégorie de la variable «Résidence», puis une catégorie de la variable «Sexe», suivie d'une autre de la catégorie «Nationalité» et finalement une autre de la catégorie «Profession». Si un enregistrement présente deux combinaisons problématiques, par exemple, «Boulanger» x «Étranger» et «Femme» x «Urk», il conviendra alors de supprimer «Urk» et «Étranger» dans cet enregistrement.

Pour les cas où certaines des combinaisons problématiques minimales partageront une catégorie commune, la situation sera quelque peu plus complexe. Supposons que dans un enregistrement donné, les combinaisons «Boulanger» x «Femme» et «Boulanger» x «Étranger» x «Urk» sont des combinaisons problématiques minimales. Dans un tel cas, nous pouvons réduire au minimum le nombre de suppressions locales en supprimant «Boulanger». Les deux combinaisons qui en découlent, c'est-à-dire, «Femme» et «Étranger» x «Urk», deviennent alors sûres. À noter que nous n'utilisons pas une liste de priorités pour cet enregistrement. Toutefois, cela pourra s'avérer nécessaire pour les enregistrements dans lesquels des combinaisons problématiques minimales possèdent plusieurs catégories communes.

Ainsi, en pratique, on peut s'attendre que le problème général d'optimisation soit réduit en un certain nombre de sous-problèmes plus petits. Ceci signifie qu'il peut même être envisageable de tester l'ensemble des possibilités afin de minimiser la fonction cible. Toutefois, il pourrait être plus rapide d'utiliser un algorithme standard pour résoudre les problèmes de programmation par nombres entiers 0-1.

Dans les sections 3, 4 et 5 du présent article, nous présentons les solutions aux problèmes posés en nous fondant sur un exemple standard. Dans cet exemple, il existe onze enregistrements problématiques et 21 catégories différentes. Ces enregistrements contiennent les combinaisons problématiques minimales suivantes :

enregistrement 1: 'A'x'B' et 'B'x'C'
enregistrement 2: 'A'x'D' et 'A'x'E'
enregistrement 3: 'C'x'F'
enregistrement 4: 'G'x'H' et 'H'x'I'
enregistrement 5: 'J'x'K'
enregistrement 6: 'J'x'L'
enregistrement 7: 'M'x'N' et 'N'x'O'
enregistrement 8: 'M'x'O'
enregistrement 9:'P'x'Q' et 'Q'x'R'
enregistrement 10:'S'x'T'
enregistrement 11: 'S'x'U'

Exemple : Si la fonction cible (4) possède des poids w qui sont tous égaux à un, la solution du problème examiné dans la présente section est donnée par :

suppression dans l'enregistrement 1: 'B'
suppression dans l'enregistrement 2: 'A'
suppression dans l'enregistrement 3: 'F'
suppression dans l'enregistrement 4: 'H'
suppression dans l'enregistrement 5: 'J'
suppression dans l'enregistrement 6: 'J'
suppression dans l'enregistrement 7: 'N'
suppression dans l'enregistrement 8: 'O'
suppression dans l'enregistrement 9: 'Q'
suppression dans l'enregistrement 10: 'T'
suppression dans l'enregistrement 11: 'U'

Ainsi, 11 valeurs sont localement supprimées et 10 catégories différentes sont prises en compte.

4. SUITE AU PROBLÈME DE SUPPRESSION LOCALE

Dans la présente section, nous nous penchons sur un certain nombre de problèmes semblables à celui examiné dans la section 3. Premièrement, au lieu de réduire au minimum le nombre total de suppressions locales, l'utilisateur des données pourrait souhaiter réduire au minimum le nombre de catégories différentes qui sont supprimées. Il pourrait, par exemple, juger qu'une catégorie localement supprimée dans certains enregistrements est peu ou pas appropriée aux fins des analyses statistiques. En d'autres mots, les catégories localement supprimées pourraient présenter, pour lui, une valeur limitée ou nulle.

Réduction au minimum du nombre de catégories différentes localement supprimées :

On peut formuler ce deuxième problème comme suit. D'abord, nous introduisons certaines variables factices nouvelles. Pour chaque catégorie j qui apparaît dans une combinaison problématique minimale, nous introduisons une variable factice z_j définie comme suit

$$z_j = \begin{cases} 1 & \text{si la catégorie } j \text{ est localement supprimée} \\ 0 & \text{si la catégorie } j \text{ n'est pas localement supprimée.} \end{cases} \quad (5)$$

À noter que les valeurs de z_j sont indépendantes des enregistrements. Il convient ensuite de satisfaire aux contraintes suivantes :

$$\sum_{j=1}^{J} c_{jk} z_j \geq 1, \text{ pour l'ensemble des } k=1,\ldots,k. \quad (6)$$

Examinons la fonction cible suivante :

$$\sum_{j=1}^{J} z_j. \quad (7)$$

La fonction cible (7) doit être minimisée compte tenu des contraintes déterminées par (6). Le problème d'optimisation qui se pose alors est un problème de recouvrement d'ensemble (cardinalité minimale).

Ce problème est plus difficile à résoudre que le problème de la section 3 puisque dans ce cas-ci, les enregistrements ne peuvent pas être considérés indépendamment. Toutefois, on peut dans beaucoup de cas le réduire à des sous-problèmes plus petits puisque les remarques faites à la fin de la section 3 s'appliquent également ici. Le problème peut parfois être réduit encore en sous-problèmes correspondant à des sous-graphiques connectés. Dans un tel cas, les combinaisons problématiques minimales correspondent aux sommets d'un graphique. Deux sommets sont joints par une arête si les combinaisons problématiques minimales correspondantes ont une catégorie en commun et si les deux combinaisons apparaissent au moins une fois simultanément dans un enregistrement. En outre, seules les variables factices correspondant aux catégories qui apparaissent dans plus d'une combinaison problématique minimale doivent être optimisées. Certains des algorithmes utilisés pour résoudre le problème ci-dessus ont été examinés par Van Gelderen (1995). Il semble qu'on puisse parvenir en peu de temps à une solution presque optimale.

Lorsque ce problème est résolu, le nombre de catégories différentes qui sont localement supprimées est réduit au minimum. Toutefois, pour certains enregistrements, il pourrait arriver qu'on ait supprimé plus de catégories que nécessaire. Pour résoudre ce problème, chaque enregistrement doit être vérifié séparément. Si le nombre des catégories de combinaisons problématiques minimales d'un enregistrement qui ont été localement supprimées est trop élevé, (c.-à-d., plus que le nombre de combinaisons problématiques minimales), certaines de ces suppressions locales pourront alors être remplacées par leur valeur originale. On peut déterminer celles des catégories supprimées qui peuvent être remplacées par leurs valeurs originales en s'aidant de la liste de priorités, en prenant garde de maintenir la sûreté de l'enregistrement.

Ce problème peut être élargi en remplaçant (7) par une somme pondérée des valeurs z_j. On pourra ainsi permettre à l'utilisateur d'indiquer l'importance de chacune des catégories. Les catégories importantes devraient recevoir un poids élevé, et les catégories moins importantes un poids plus faible. Le problème ainsi obtenu peut parfois être décomposé en un certain nombre de sous-problèmes, selon un procédé semblable à celui décrit dans la section 3.

Exemple : (suite) Nous revenons maintenant à l'exemple utilisé plus haut. Le problème examiné dans la présente section peut notamment être résolu comme suit :

suppression dans l'enregistrement	1: 'A' et 'C'
suppression dans l'enregistrement	2: 'A'
suppression dans l'enregistrement	3: 'C'
suppression dans l'enregistrement	4: 'H'
suppression dans l'enregistrement	5: 'J'
suppression dans l'enregistrement	6: 'J'
suppression dans l'enregistrement	7: 'M' et 'O'

suppression dans l'enregistrement 8: 'O'
suppression dans l'enregistrement 9: 'Q'
suppression dans l'enregistrement 10: 'S'
suppression dans l'enregistrement 11: 'S'

Ainsi, 13 valeurs sont localement supprimées et 8 catégories différentes sont prises en compte.

Maximisation du nombre de catégories différentes localement supprimées, compte tenu d'une réduction au minimum du nombre de suppressions locales :

On peut encore élaborer quelque peu sur les problèmes examinés jusqu'à maintenant. Supposons que le nombre de suppressions locales a été réduit au minimum par la résolution du problème de programmation par nombres entiers 0-1IP de la section 3. Supposons en outre que parmi les solutions obtenues, nous souhaitons trouver celle qui supprime un nombre maximal de catégories différentes. De cette manière, les suppressions locales s'étendront probablement plus ou moins également à l'ensemble des diverses catégories.

Ce problème peut être formalisé comme suit. Désignons par N_{min} le nombre minimal de suppressions locales. Ce nombre est connu puisque nous présumons que le problème de la section 3 a été résolu. Nous voulons utiliser les variables y_{ij} et les variables z_j en un seul problème. La variable z_j devrait être égale à un si et seulement s'il existe une valeur y_{ij} égale à un pour certains I. On peut réaliser ces conditions en utilisant un grand nombre W et en introduisant les relations suivantes :

$$Wz_j \geq \sum_{i=1}^{I} y_{ij}, \text{ pour tous les } j=1,\ldots, J \qquad (8)$$

et

$$y_{ij} \geq z_j, \text{ pour tous les } i=1,\ldots,I;\ j=1,\ldots, J. \qquad (9)$$

Comme nous souhaitons obtenir le nombre minimal de suppressions locales, nous devons ajouter la contrainte suivante :

$$\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} = N_{min}. \qquad (10)$$

La fonction cible envisagée est donnée par (7). Cette fonction cible doit être maximisée compte tenu des contraintes données par (2), (8), (9) et (10).

Exemple: (suite) Revenons encore une fois à notre exemple. Le problème décrit ci-dessus peut être résolu comme suit :

suppression dans l'enregistrement 1: 'B'
suppression dans l'enregistrement 2: 'A'
suppression dans l'enregistrement 3: 'F'
suppression dans l'enregistrement 4: 'H'
suppression dans l'enregistrement 5: 'K'
suppression dans l'enregistrement 6: 'L'
suppression dans l'enregistrement 7: 'N'
suppression dans l'enregistrement 8: 'M'
suppression dans l'enregistrement 9: 'Q'
suppression dans l'enregistrement 10: 'T'
suppression dans l'enregistrement 11: 'U'

Ainsi, 11 valeurs sont localement supprimées et 11 catégories différentes sont prises en compte.

Réduction au minimum du nombre de catégories différentes supprimées localement, lorsque le nombre de suppressions locales a été réduit au minimum :

Comme dans le problème examiné plus haut, l'utilisateur des données pourrait souhaiter supprimer le moins possible de catégories différentes tout en supprimant le moins de valeurs possibles. Dans ce cas, la fonction cible (7) doit être minimisée compte tenu des contraintes données par (2), (8), (9) et (10).

Exemple : (suite) Si nous revenons à notre exemple, on peut donner au problème ci-dessus la solution suivante :

suppression dans l'enregistrement 1: 'B'
suppression dans l'enregistrement 2: 'A'
suppression dans l'enregistrement 3: 'F'
suppression dans l'enregistrement 4: 'H'
suppression dans l'enregistrement 5: 'J'
suppression dans l'enregistrement 6: 'J'
suppression dans l'enregistrement 7: 'N'
suppression dans l'enregistrement 8: 'M'
suppression dans l'enregistrement 9: 'Q'
suppression dans l'enregistrement 10: 'S'
suppression dans l'enregistrement 11: 'S'

Ainsi, 11 valeurs sont localement supprimées et 9 catégories différentes sont prises en compte.

Réduction au minimum du nombre de suppressions locales, compte tenu de la réduction au minimum préalable du nombre de catégories différentes localement supprimées :

Le dernier problème que nous examinons est le suivant. Imaginons que le nombre de catégories différentes localement supprimées a été réduit au minimum par la résolution du premier problème de programmation par nombres entiers 0-1 IP examiné dans la présente section. Supposons par ailleurs que parmi ces solutions, nous voulons trouver celle qui mène à la suppression d'un nombre minimal de valeurs.

Désignons par M_{min} le nombre minimal de catégories différentes supprimées localement. Ce nombre est connu puisque nous présumons que le problème correspondant a été résolu (approximation acceptable). Introduisons la contrainte suivante :

$$\sum_{j=1}^{J} z_j = M_{min} . \qquad (11)$$

Dans ce cas, nous devons minimiser (3), toutes les valeurs de w_{ij} étant égales à un, compte tenu des contraintes indiquées par (2), (8), (9) et (10).

Exemple : (suite) Pour la dernière fois, revenons à notre exemple. La solution du problème ci-dessus est donnée par :

suppression dans l'enregistrement	1: 'A' et 'C'
suppression dans l'enregistrement	2: 'A'
suppression dans l'enregistrement	3: 'C'
suppression dans l'enregistrement	4: 'H'
suppression dans l'enregistrement	5: 'J'
suppression dans l'enregistrement	6: 'J'
suppression dans l'enregistrement	7: 'N'
suppression dans l'enregistrement	8: 'M'
suppression dans l'enregistrement	9: 'Q'
suppression dans l'enregistrement	10: 'S'
suppression dans l'enregistrement	11: 'S'

Ainsi, 12 valeurs sont supprimées localement et 8 catégories différentes sont prises en compte.

Les problèmes de suppression locale examinés dans le présent article ne tiennent pas compte de la dépendance entre les variables. Par exemple, il est inutile de supprimer la valeur «Femme» de la variable «Sexe» lorsque la valeur de la variable «Quand avez-vous donné naissance à votre premier enfant?» est «1976». La valeur de la seconde variable signifie nécessairement que celle de la première est «Femme». On peut étendre l'application des problèmes de suppression locale examinés dans le présent article pour prendre en compte de telles dépendances (voir De Waal et Willenborg, 1995a). Nous ne nous attardons pas sur cette question ici, préférant porter notre attention sur le problème du recodage global et sur celui de la combinaison RG-SL.

5. REMARQUES PRÉLIMINAIRES CONCERNANT LE PROBLÈME DU RECODAGE GLOBAL

Dans l'analyse qui suit du problème du recodage global et du problème de la combinaison RG-SL, nous présumerons, comme dans les sections précédentes, que la protection d'un ensemble de microdonnées consiste à s'assurer que certaines combinaisons de catégories des variables repères ont une fréquence suffisante, c'est-à-dire une fréquence qui dépasse une certaine valeur limite. Lorsqu'une certaine combinaison de catégories de variables repères n'apparaît pas avec une fréquence suffisante, il s'agit soit de supprimer localement une ou plusieurs des valeurs de cette combinaison de catégories ou de recoder globalement une ou plusieurs des variables de manière que la combinaison correspondante des catégories globalement recodées présente une fréquence suffisante.

Le recodage global et la suppression locale sont des opérations qui mènent toutes deux à une perte d'information. Si nous souhaitons offrir aux utilisateurs des données le plus d'information possible, nous devons déterminer les recodages globaux et les suppressions locales de manière à transformer un ensemble de microdonnées problématiques en un ensemble sûr tout en conservant la plus grande proportion possible d'informations. Il faut, pour y arriver, résoudre un certain nombre de problèmes : déterminer une méthode propice de mesure de l'information, élaborer la formule mathématique du problème de recodage global et du problème de la combinaison RG-SL; élaborer des algorithmes qui permettent de résoudre ces problèmes (en donnant une approximation acceptable).

Mais avant de décrire le problème du recodage global et celui de la combinaison RG-SL, il convient de clarifier la nature précise des recodages globaux. Pour réaliser un tel recodage global pour une variable, il faut définir une structure de proximité sur le domaine de cette variable. Supposons par exemple que sur le domaine D d'une variable (clé) donnée, une métrique, ou fonction de distance d a été définie. Cette métrique définit la distance qui sépare deux catégories données, C_1 et C_2 de D, et est exprimée par $d(C_1 , C_2)$. Nous présumerons que d peut également prendre la valeur '∞' (infini), afin de décrire une situation où les deux catégories seraient trop éloignées pour qu'on puisse imaginer leur combinaison en une seule catégorie. Imaginons maintenant une catégorie C_1 qui doit être combinée avec une ou plusieurs des autres catégories de D dans le cadre d'une opération de recodage global. Il devient alors clair qu'il faut pour cela chercher des catégories «dans le voisinage» de C_1, ce qui, dans notre cas, devrait signifier que $d(C_1 , C_2) < \tau$ pour une certaine valeur critique de τ. Supposons maintenant qu'il existe une autre catégorie C_3 qui se trouve suffisamment près de C_1, c'est-à-dire, $d(C_1 , C_3) < \tau$. Pour pouvoir accepter $C_1 + C_2 + C_3$ à titre de recodage global valide, il faudrait également que $d(C_2 , C_3) < \tau$.

Nous pouvons poursuivre ce processus de détermination des recodages globaux valides jusqu'à ce que tous les recodages globaux valides intéressant C_1 aient été déterminés. À toutes les étapes, il convient par ailleurs de s'assurer que les distances mutuelles qui séparent les catégories à recoder en une seule sont inférieures à τ.

Il convient de se pencher plus avant sur la façon de définir une structure de proximité sur le domaine d'une variable qui pourra servir aux fins de recodage global. Les possibilités décrites ci-après nous paraissent utiles.

1. En dérivant une métrique sur D à l'aide des méthodes statistiques (multivariées) appropriées, comme dans le cas du regroupement. Nous présentons ci-après un exemple d'une telle méthode. Désignons par V la variable en cause. Désignons par ailleurs par C_1 et C_2 deux catégories de V. Sélectionnons une variable V qui est étroitement corrélée avec V. Déterminons maintenant les fréquences du tableau de contingence $V \times V$. Lorsque les fréquences de la ligne définie par $V=C_1$ sont données par $(n_1, ..., n_s)$ et que les fréquences correspondantes de la ligne définie par $V=C_2$ sont données par $(m_1, ..., m_s)$, la distance qui sépare C_1 de C_2 peut alors être définie par

$$d(C_1, C_2) = \sum_{i=1}^{s} \frac{(n_i - m_i)^2}{(n_i + m_i)}, \qquad (12)$$

où $(n_i - m_i)^2 / (n_i + m_i)$ est égal à zéro par définition lorsque les valeurs n_i et m_i sont égales à zéro. Lorsque la distance entre chaque paire est inférieure à τ, on peut regrouper un certain nombre de catégories en une seule.

2. En utilisant un graphique élémentaire sur D qui définit directement les points voisins. La structure de proximité à utiliser pour le recodage global peut être dérivée de ce graphique, compte tenu d'une valeur limite τ, d'une façon semblable à celle utilisée dans l'exemple précédent, en présumant que chaque bord a une longueur de 1. Pour certains types de variables, on peut imaginer un graphique de «voisinage direct». Par exemple, on peut utiliser l'ordonnancement des catégories pour les variables ordinales, la hiérarchisation pour les variables hiérarchiques (il s'agit d'une structure arborescente en présentation graphique), et la contiguïté pour les aires géographiques.

Ainsi, il existe diverses façons de déterminer des structures de proximité définies pour une variable clé. Le processus peut être passablement laborieux au départ, mais lorsqu'il a été effectué pour une variable clé, le résultat peut également servir pour d'autres ensembles de microdonnées qui contiennent la même variable.

Au lieu de déterminer une structure de proximité pour chaque variable clé, le responsable de la protection du secret statistique pourrait également préciser lui-même les recodages globaux valides. Le problème consiste alors à choisir le recodage global valide (et les suppressions locales) qui minimisent la perte d'information tout en protégeant l'ensemble de microdonnées. Comme le nombre de recodages globaux valides indiqué par le responsable de la protection du secret statistique sera généralement moindre que lorsque les structures de proximité sont utilisées, la complexité du recodage global et de la combinaison RG-SL en sera réduite.

6. MESURES DE LA PERTE D'INFORMATION

Pour pouvoir comparer les informations contenues dans un ensemble de microdonnées avant et après les recodages globaux et les suppressions locales, il faut élaborer une méthode adéquate de mesure de l'information. Nous proposons d'utiliser l'entropie pour mesurer la perte d'information due aux recodages globaux et aux suppressions locales. L'entropie est une mesure bien connue de l'incertitude entourant la valeur d'une variable. Plus l'incertitude concernant la valeur réelle d'une variable est élevée, plus sera élevée également la perte d'information. Nous commencerons par expliquer comment la perte d'information due aux recodages globaux et aux suppressions locales peut être mesurée séparément pour chaque variable et pour chaque enregistrement.

Supposons, par exemple, que la variable «État civil» peut prendre quatre valeurs : «Veuf», «Divorcé», «Marié» et «Célibataire». Supposons en outre que deux catégories de cette variable, «Veuf» et «Divorcé», ont été combinées en une seule : «Veuf ou Divorcé». Dans un tel cas, l'utilisateur ne connaît pas la valeur originale de la variable lorsqu'il trouve cette catégorie combinée dans un enregistrement. Il doit deviner si la valeur originale est «Veuf» ou «Divorcé». Ainsi, pour cet utilisateur, la valeur originale de la variable «État civil» est entourée, dans ce cas précis, d'une certaine dose d'incertitude. Lorsque, par contre, la valeur de l'«État civil» est localement supprimée, l'utilisateur doit deviner laquelle des quatre catégories possibles est la bonne. Ici encore, la valeur originale de l'«État civil» s'entoure d'une certaine dose d'incertitude. Dans les deux cas, celui de recodage global et de la suppression locale, on peut mesurer la valeur originale de l'«État civil» au moyen de l'entropie.

Dans les cas où les catégories «Veuf» et «Divorcé» ont été combinées en une seule catégorie «Veuf ou

Divorcé», l'entropie se définit de la façon suivante. Désignons par p_w la probabilité que la valeur originale de l'«État civil» soit «Veuf» et par p_D la probabilité que cette valeur soit «Divorcé». L'entropie H, correspondant à la perte d'information due au recodage global, nous est donnée par la formule suivante :

$$H = -p_w' \log(p_w') - p_D' \log(p_D') , \qquad (13)$$

où p_W' et p_D' désignent les probabilités conditionnelles que la valeur originale de l'«État civil» soit égale à «Veuf» et «Divorcé» respectivement, compte tenu d'une valeur recodée de l'«État civil» égale à «Veuf ou divorcé». Cette situation s'exprime mathématiquement comme suit :

$$p_W' = \frac{p_W}{p_W + p_D} \text{ et } p_D' = \frac{p_D}{p_W + p_D} . \qquad (14)$$

Dans les cas de suppression locale de l'«État civil», l'entropie sera définie comme suit. Désignons par p_W, la probabilité que la valeur originale de l'«État civil» soit «Veuf», par p_D, la probabilité que cette valeur soit «Divorcé», par p_M la probabilité que cette valeur soir «Marié», et par $p_U = 1 - p_W - p_D - p_M$ la probabilité qu'elles soit égale à «Célibataire». La valeur de l'entropie H, c'est-à-dire la perte d'information due à la suppression locale, s'obtient par la formule

$$H = -p_w \log(p_w) - p_D \log(p_D) - p_M \log(p_M) - p_U \log(p_U) . \qquad (15)$$

À noter que la suppression locale constitue en fait un cas extrême de recodage global; il s'agit du recodage de l'ensemble des catégories en une seule. Cette propriété est prise en compte dans la mesure de l'entropie ci-dessus. Le recodage global diffère cependant de la suppression locale en ce qu'il conduit à une perte d'information dans tous les enregistrements qui contiennent au moins une des valeurs recodées, tandis que la suppression locale n'entraîne une perte d'information que dans le dossier correspondant à la variable supprimée.

En général, lorsque les catégories $C_1\ C_2\ ,...,C_n$ d'une variable V sont combinées en une seule, $C_1 + C_2 + ... + C_n$, la perte d'information due à ce recodage global est donnée par

$$H = -\sum_{i=1}^{m} p_i' \log(p_i') , \qquad (16)$$

où p_i' correspond à la probabilité conditionnelle que la valeur originale de V soit égale à C_i, compte tenu d'une valeur recodée égale à $C_1 + C_2 + ... + C_n$, et d'un nombre m de catégories de V. À noter que $p_j' = 0$ lorsque C ne fait pas partie de la catégorie recodée $C_1 + C_2 + ... + C_n$. Lorsque $p_j' = 0$; $p_j' \log(p_j')$ est égal à zéro par définition. Lorsque la valeur de V est supprimée localement, la formule (16) permet également de calculer la perte d'information.

On peut évaluer grâce à la même formule (16) la perte d'information subie pour une variable particulière et un enregistrement particulier par suite du recodage global et de la suppression locale. Il convient toutefois, avant d'utiliser cette formule, de résoudre un problème fondamental : les probabilités p_i ne peuvent être calculées sans qu'on détermine au préalable un modèle particulier pour l'utilisateur des données. Plusieurs de ces modèles peuvent être imaginés. Par exemple, l'utilisateur des données peut exploiter la totalité des informations contenues dans les dossiers pour déterminer les probabilités p_i. Pour y parvenir, on peut avoir recours à diverses méthodes d'analyse multivariée parfois assez élaborées. On peut également présumer qu'un utilisateur tiendra compte du fait que certaines des «données manquantes» sont une conséquence directe de la SDC. Lorsqu'une «donnée manquante» est due à la SDC, la valeur originale se présente rarement en combinaison avec d'autres catégories. L'utilisateur peut utiliser cette information pour évaluer les probabilités p_i. Toutefois, nous présumerons ici que les probabilités p_i sont calculées d'une manière plutôt directe que nous expliquerons ci-après.

Supposons qu'une variable V peut prendre m valeurs possibles. La fréquence de la $I^{ième}$ catégorie, C_i, de la variable V dans la population est égale à N_i. Désignons par N_V le nombre de sujets compris dans la population cible de la variable V. Notez que ce N_V peut être différent d'une variable à l'autre. Par exemple, la question «Quel est votre âge?» et la réponse correspondante ont une population cible passablement différente de la question «Quand avez-vous donné naissance à votre premier enfant?». La première question est posée à l'ensemble de la population tandis que la seconde ne s'adresse qu'aux femmes qui ont accouché. La distribution de probabilité (inconditionnelle) de V sert à évaluer les probabilités p_i, ainsi p_i est défini par

$$p_i = \frac{N_i}{N_V} . \qquad (17)$$

Il arrive souvent en pratique qu'on ne connaisse pas les valeurs de N_V et de N_i, tandis que dans la plupart des cas, seules les notes d'un échantillon de la population

sont connues. Nous pouvons résoudre ce problème de la façon suivante. Désignons par n_i la fréquence de la catégorie C_i de la variable V de l'échantillon, et par n_V le nombre de sujets de la population cible V dans l'échantillon. En ce qui concerne les «données manquantes», nous présumons ce qui suit. Premièrement, une «donnée manquante» peut être causée par un sujet qui ne fait pas partie de la population cible de la variable V. Pour un tel sujet, la valeur de la variable V est «manquante» par définition. Deuxièmement, nous présumons que la probabilité qu'une «donnée manquante» soit causée par un sujet de la population cible est égale pour tous les sujets de la population cible. Compte tenu de ces suppositions, nous pouvons calculer la probabilité estimative p_i de la formule (17) par

$$\hat{p}_i = \frac{n_i}{n_V} . \qquad (18)$$

Comme il arrive souvent que la formule (17) ne soit pas applicable, nous utilisons la formule (18) pour estimer l'entropie d'un recodage global ou d'une suppression locale.

La valeur estimative calculée par (18) de la probabilité donnée en (17) est plutôt simpliste. Lorsque certains des nombres n_i sont petits, la qualité des estimations des probabilités correspondantes peut être assez mauvaise. On peut cependant proposer des méthodes plus perfectionnées de calculer les probabilités données par la formule (17).

Les probabilités p_i doivent être évaluées pour nous permettre de calculer la perte d'information due aux recodages globaux et aux suppressions locales. Au lieu de recourir au modèle plutôt simple décrit ci-dessus pour évaluer ces probabilités p_i, on peut élaborer des modèles plus perfectionnés. Nous tenons cependant à souligner qu'il n'est pas nécessaire d'utiliser un modèle très élaboré puisque aucune méthode officielle de mesure de l'information ne permet de quantifier pleinement la valeur «véritable» de la perte d'information d'un ensemble de microdonnées due au recodage global et à la suppression locale. En fait, cette perte d'information d'un ensemble de microdonnées est déterminée en très grande partie par les utilisateurs de ces données qui font appel à des considérations subjectives. En outre, la méthode de mesure de l'information est un outil qui doit servir uniquement à trouver les bons recodages globaux et les bonnes suppressions locales. Il n'est toutefois pas possible de trouver les «meilleurs» recodages globaux et les «meilleures» suppressions locales, puisque «meilleur» est un terme subjectif.

Jusqu'à maintenant, nous avons surtout porté notre attention sur la perte d'information due au recodage global d'une variable ou à la suppression locale d'une valeur de la variable dans un enregistrement. Toutefois, il nous faut également mesurer la perte d'information due aux recodages globaux et aux suppressions locales effectués dans l'ensemble complet de microdonnées. La méthode de mesure de la perte d'information dans un ensemble complet de microdonnées que nous nous proposons d'utiliser est une somme pondérée des pertes d'information des variables contenues dans les enregistrements. Lorsque la perte d'information de la $I^{ième}$ variable du $j^{ième}$ enregistrement est désignée par H_{ij}, la perte d'information pour l'ensemble complet de microdonnées est donnée par

$$\Sigma_i \Sigma_j w_{ij} H_{ij}, \qquad (19)$$

où w_{ij} représente un poids non négatif et où la somme concerne l'ensemble des variables I et des dossiers j. Plus le poids w_{ij} sera grand, plus la valeur de la $I^{ième}$ variable du $j^{ième}$ dossier sera importante. Les poids peuvent être choisis par un utilisateur à partir de considérations subjectives. La perte d'information H_{ij} de la $I^{ième}$ variable du $j^{ième}$ enregistrement due au recodage global ou à la suppression locale est mesurée à l'aide de la méthode décrite dans la présente section.

Pour ne pas compliquer inutilement les choses, nous proposons d'utiliser des poids égaux pour des enregistrements différents, c'est-à-dire, de remplacer (19) par

$$\Sigma_i \Sigma_j w_i H_{ij}. \qquad (20)$$

Autrement dit, les poids ne pourront être différents que d'une variable à l'autre, et non d'un enregistrement à l'autre. La valeur de H_{ij} est mesurée à l'aide de la formule (16). Cependant, nous suggérons de faire en sorte que la plupart des w_i soient égaux à un, et de n'utiliser un poids plus élevé que pour les variables qui sont nettement plus importantes que d'autres. De la même façon, seules les variables qui sont de toute évidence moins importantes que les autres devraient être assorties d'un poids inférieur à un.

Il est également possible d'utiliser des poids égaux pour des variables différentes, c'est-à-dire de remplacer (19) par

$$\Sigma_i \Sigma_j w_j H_{ij}. \qquad (21)$$

Dans ce cas, les poids d'échantillonnage des enregistrements constituent des choix naturels pour les poids w_j.

7. LE PROBLÈME DU RECODAGE GLOBAL

On peut énoncer en termes généraux le problème du recodage global optimal. Pour un ensemble donné de combinaisons problématiques minimales, l'objectif consiste à procéder aux recodage globaux (d'une partie) des variables contenues dans ces combinaisons afin de les éliminer tout en réduisant au minimum la perte d'information de l'ensemble de microdonnées. Même si elle peut paraître simple à première vue, une telle opération est passablement complexe. Contrairement au problème de la suppression locale où le remplacement d'au moins une valeur d'une combinaison problématique minimale par une «donnée manquante» produirait automatiquement une combinaison sûre, le recodage global ne se prête pas à une solution aussi facile : le recours à un recodage global valide portant sur deux catégories dont l'une semble faire partie d'une combinaison problématique ne garantit pas l'«élimination» de cette combinaison problématique, c'est-à-dire son absorption dans une nouvelle catégorie dont la fréquence est suffisante.

Pour chaque opération de recodage global valide, il faut déterminer dans l'ensemble de données les enregistrements qui seraient touchés si on procédait au recodage, et déterminer si l'opération permettrait d'éliminer efficacement une combinaison rare. Si tel est le cas, le recodage global pourra présenter un certain intérêt et on pourra déterminer la perte d'information qui en découlera.

Le problème du recodage global s'apparente étroitement au problème qui se pose lorsqu'il s'agit de protéger un ensemble de tableaux liés, c'est-à-dire, de tableaux comportant des variables communes obtenues à partir du même fichier de base. Supposons qu'on veuille utiliser la même catégorisation pour chacune des variables de chacun des tableaux où cette variable apparaît. Supposons en outre que nous souhaitons protéger les tableaux contre la divulgation au moyen d'un recodage portant uniquement sur les variables. Dans un tel cas, les tableaux peuvent être comparés aux combinaisons problématiques de microdonnées. Ils s'en distinguent uniquement par le critère sous-jacent qui détermine le caractère problématique d'un tableau ou d'une combinaison.

8. LE PROBLÈME DE LA COMBINAISON RG-SL

Pour un ensemble donné de combinaisons problématiques minimales, le problème de la combinaison RG-SL consiste à recoder globalement (une partie) des variables et à supprimer localement certaines des valeurs de manière que l'ensemble de microdonnées soit sûr et que la perte des informations due aux opérations de recodage global et de suppression locale soit minimale. La résolution de ce problème de RG-SL ou du problème de recodage global pour donner un résultat optimal, c'est-à-dire pour limiter autant que possible la perte d'information, constitue un difficile problème d'optimisation.

Dans les cas où on peut résoudre le problème du recodage global, celui du RG-SL peut également être résolu, en théorie, en prenant en compte toutes les partitions de l'ensemble de combinaisons problématiques. Étant donné un ensemble U de combinaisons problématiques, examinons toutes les partitions de U en deux ensembles U_1 et U_2. Il s'agit, d'une manière générale, d'éliminer les combinaisons de U_1 par recodage global optimal, et d'éliminer celles de U_2 par suppression locale optimale. Nous décrivons ci-après en détail la façon de procéder. Premièrement, nous éliminons toutes les combinaisons en U_1 par recodage global optimal. Ensuite, nous déterminons pour les nouvelles variables clé l'ensemble des combinaisons problématiques qui restent (dans les cas extrêmes, cet ensemble peut être vide). Désignons cet ensemble par U_2'. Éliminons maintenant les combinaisons de cet ensemble par le biais des suppressions locales optimales. On présume que la perte d'information pour chaque valeur supprimée sera mesurée par entropie.

On peut calculer la perte d'information due au recodage global ou à la suppression locale. La perte totale d'information correspond à la somme de ces pertes d'information partielles. Pour en terminer avec cette partition $\{U_1, U_2\}$, inverser les rôles de U_1 et U_2 dans la procédure décrite ci-haut et répéter pour la nouvelle situation.

Le nombre de bipartitions $\{U_1, U_2\}$ de U à prendre en compte peut être passablement élevé : s'il existe n combinaisons rares, il pourra y avoir 2^n bipartitions. Il s'agirait de mettre au point une méthode utile qui permettrait de chercher efficacement, dans l'ensemble des bipartitions, un ensemble donné de combinaisons problématiques minimales.

9. DISCUSSION

La suppression locale optimale et automatique des catégories d'enregistrements afin d'en protéger le secret statistique ne semble pas trop difficile à réaliser. Elle est particulièrement facile lorsque le nombre total de

catégories supprimées doit être réduit au minimum puisque les problèmes qui en découlent - un problème pour chaque enregistrement problématique - sont tous très petits. Dans les cas où le nombre de catégories localement supprimées différentes doit être réduit au minimum, le problème est un peu plus complexe. Toutefois, dans beaucoup de cas, on peut le décomposer en un certain nombre de problèmes plus petits et donc plus faciles à résoudre. Ainsi, il semble possible de résoudre ce problème d'une façon passablement efficace. Finalement, il est également possible de résoudre certaines versions de ces deux problèmes de base tels que la réduction au minimum du nombre de catégories localement supprimées différentes lorsque le nombre de valeurs localement supprimées est minimal.

La méthode fondamentale de détermination des suppressions locales que nous avons décrite dans le présent article, c'est-à-dire la réduction au minimum d'un nombre de suppressions locales au moyen de la résolution d'un problème de programmation par nombres entiers 0-1 IP, peut également servir dans d'autres cas. Par exemple, l'utilisateur voudra peut-être remplacer certaines des catégories d'une combinaison problématique par d'autres catégories au lieu d'en supprimer localement les valeurs. Ce problème peut être résolu, à tout le moins en ce qui a trait à l'optimisation, d'une façon semblable à celui examiné à la section 3. On commence par supprimer localement les valeurs des combinaisons problématiques minimales d'une façon optimale. On procède ensuite à l'imputation des valeurs qui remplacent les valeurs supprimées. Il faut pour cela déterminer les imputations possibles pour chaque enregistrement. L'imputation est possible dans les cas où l'enregistrement obtenu est sûr, c'est-à-dire, si aucune des combinaisons n'apparaît dans la liste des combinaisons problématiques minimales et si toutes les règles de révision prescrites (le cas échéant) sont respectées. Le problème principal n'en est pas un d'optimisation, mais plutôt d'imputation afin d'assurer l'intégrité des données. La formulation proposée ici montre que le problème de suppression locale est lié au problème de révision et d'imputation, c'est-à-dire au problème de la localisation des erreurs commises par les répondants d'un questionnaire et de leur remplacement par des réponses meilleures. Les révisions, c'est-à-dire la vérification des règles prescrites pour les données d'un enregistrement, permettent de vérifier si des erreurs ont été commises par un répondant. On peut comparer ces révisions aux vérifications de fréquence qui doivent être réalisées dans le cadre de la SDC. En fait, les révisions de fréquence peuvent être assimilées à des macro-révisions.

La possibilité de procéder à la suppression locale optimale automatique des catégories au moyen d'un programme informatique comme ARGUS (voir Pieters et De Waal, 1995; De Waal et Pieters, 1995) accélère énormément le processus de production d'un ensemble de microdonnées sûres. Elle permet en outre aux utilisateurs éventuels des ensembles de microdonnées de participer à la production d'une version sûre de cet ensemble qui répondra au mieux à leurs besoins, compte tenu des contraintes de la SDC. Les bureaux de la statistique et les utilisateurs des ensembles de microdonnées peuvent tirer avantage de cette situation. Le bureau de la statistique profite du fait que les utilisateurs des ensembles de microdonnées se montreront moins critiques vis-à-vis des méthodes de surveillance de la divulgation des données, et les utilisateurs profitent d'un accès assez rapide à la plupart des informations utiles, compte tenu des restrictions en vigueur.

Le recodage global automatique des variables visant à protéger les ensembles de microdonnées contre la divulgation est beaucoup plus difficile à réaliser. Ceci est vrai tant dans le cas des ensembles de microdonnées protégées par recodage global uniquement que dans le cas des ensembles de microdonnées protégées par une combinaison de suppressions locales et de recodages globaux. Un des problèmes qui se pose dans cette situation est la nécessité d'élaborer une méthode de mesure de la perte d'information qui en découle. Dans le présent article, nous avons proposé une mesure fondée sur l'entropie.

Il semble plutôt difficile d'interpréter le problème de recodage global et le problème de RG-SL comme un problème d'optimisation. Dans le présent article, nous nous sommes limités à une description verbale de ces problèmes. Le problème d'élaborer une formule acceptable du recodage global et de la combinaison RG-SL demeure entier. Les problèmes que cela pose pourraient être résolus en partie en donnant au responsable du secret statistique le loisir de préciser lui-même les recodages globaux valides. De cette façon, la complexité du problème de recodage global et du problème de la combinaison RG-SL pourrait être considérablement réduite.

BIBLIOGRAPHIE

De Waal, A.G., et Pieters, A.J. (1995). ARGUS User's Guide, Rapport, Department of Statistical Methods, Statistics Netherlands, Voorburg.

De Waal, A.G., et Willenborg, L.C.R.J. (1994). Minimizing the Number of Local Suppressions, Rapport, Department of Statistical Methods, Statistics Netherlands, Voorburg.

De Waal, A.G., et Willenborg, L.C.R.J. (1995a). Local Suppression in Statistical Disclosure Control and Data Editing, Rapport, Department of Statistical Methods, Statistics Netherlands, Voorburg.

De Waal, A.G., et Willenborg, L.C.R.J. (1995b). Optimum Global Recoding and Local suppression in Microdata Sets. Rapport, Department of Statistical Methods, Statistics Netherlands, Voorburg.

Nemhauser, G.L., et Wolsey, L.A. (1988). *Integer and Combinatorial Optimization,* Wiley, New York.

Pieters, A.J. et De Waal, A.G. (1995). A Demonstration of ARGUS, Rapport, Department of Statistical Methods, Statistics Netherlands, Voorburg.

Van Gelderen, R. (1995). ARGUS: Statistical Disclosure Control of Survey Data, Rapport, Department of Statistical Methods, Statistics Netherlands, Voorburg.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

UTILISATION DU BRUIT POUR RESTREINDRE LA DIVULGATION DES DONNÉES SUR LES ENTREPRISES DANS LES TABLEAUX

B. T. Evans et L. Zayatz[1]

RÉSUMÉ

Le *Bureau of the Census* cherche de nouvelles méthodes pour restreindre la divulgation de l'information qu'il pourrait utiliser dans les tableaux de données sur les entreprises. La méthode courante consiste à supprimer les cellules des tableaux susceptibles de dévoiler des renseignements spécifiques sur un répondant. Dans l'espoir de trouver une autre méthode qui permettrait la publication de données plus nombreuses et ainsi de répondre à la demande croissante de tableaux spéciaux, on envisage maintenant d'ajouter du bruit aux microdonnées sous-jacentes. En effet, en dénaturant les données du répondant, il est possible de protéger ce dernier sans supprimer le total des cellules. Néanmoins, on peut s'interroger sur l'utilité des données après l'addition de bruit. Le présent document examine les avantages et les inconvénients de l'addition de bruit aux microdonnées avant la mise en tableaux, tant sous l'angle des problèmes de divulgation que sous celui du comportement des estimations publiées.

MOTS CLÉS : Confidentialité; divulgation; bruit; suppression de cellules.

1. INTRODUCTION

Le *Census Bureau* utilise l'établissement comme unité répondante dans bon nombre des enquêtes économiques et des recensements qu'il entreprend. On pondère (s'il y a lieu) et regroupe les réponses des établissements, et des estimations sont généralement produites en fonction de variables catégoriques, par exemple le code de la Classification type des industries (CTI) ou la région. Il est raisonnable de supposer que l'utilisateur parvient généralement bien à identifier les établissements qui, combinés, donnent le résultat apparaissant dans une cellule, surtout s'il met à profit ses connaissances générales et les sources accessibles au public, et si les tableaux reposent sur des paramètres géographiques et d'autres caractéristiques.

Le *Census Bureau* recueille les données des répondants en vertu du *Title 13* du *U.S. Code*, qui lui interdit de diffuser une publication quelconque si les données fournies par un établissement ou un particulier aux termes de la loi concernée peuvent être identifiées. La difficulté que pose cette restriction est d'empêcher l'utilisateur de déduire les valeurs signalées par le répondant de celles qui apparaissent dans les tableaux divulgués au public. Le *Census Bureau* doit s'assurer, d'une part, que la valeur inscrite dans une cellule ne s'approche pas trop des données d'un répondant et, d'autre part, qu'un répondant ou un groupe de répondants ne puisse soustraire sa (ses) propre(s) contribution(s) de la valeur de la cellule pour se faire une «idée» de la contribution d'un autre répondant (Cox et Zayatz, 1993).

2. SUPPRESSION DES CELLULES

La technique à laquelle le *Census Bureau* recourt présentement pour limiter la divulgation des données sur les établissements dans les tableaux est la suppression des cellules. Les cellules pour lesquelles existe un risque de divulgation sont identifiées de deux façons --- d'après la règle n-k ou d'après la règle *p %* (lire *Federal Committee on Statistical Methodology*, 1994 pour des explications détaillées sur ces règles). Les cellules qui dérogent à une de ces règles sont dites «sensibles» et sont les premières à être supprimées (suppressions primaires).

La technique de suppression des cellules prévient la divulgation des données en éliminant les cellules sensibles et un nombre suffisant de cellules

---

[1] B. Timothy Evans et Laura Zayatz, Statistical Research Division, Bureau of the Census, Washington, D.C. 20233, É.-U.

additionnelles, ou complémentaires, de telle sorte que la valeur des cellules primaires ne peut être estimée avec précision par manipulation de ses relations d'addition avec le total (Cox et Zayatz, 1993). Quand on supprime une cellule, on en efface la valeur totale et on remplace celle-ci par la marque «D».

Bien que les principes à la base de l'identification des cellules qui peuvent entraîner des risques de divulgation soient relativement simples, le mécanisme servant à déterminer quelles cellules complémentaires seront éliminées pour protéger les premières est très complexe. Une personne sans connaissance de la programmation linéaire éprouvera d'énormes difficultés à comprendre la méthode utilisée et le logiciel apparenté. En outre, étant donné la structure des programmes informatiques, le mécanisme doit être effectué séparément pour chaque tableau de donnés. Par conséquent, l'analyste doit garder la trace des cellules qui ont été supprimées d'un tableau de données à l'autre (puisqu'elles doivent être supprimées et protégées dans tous les tableaux subséquents) et noter quelles cellules ont déjà été publiées (donc ne peuvent servir de cellules complémentaires).

Coordonner les suppressions d'un tableau à l'autre devient presque irréalisable à cause de la multiplication des tableaux spéciaux. Pour vraiment empêcher la divulgation, il faudrait prendre en note *tous* les tableaux spéciaux réclamés par l'*ensemble* des utilisateurs et identifier non seulement les cellules qui ont déjà été supprimées dans un tableau quelconque, mais aussi celles qui ne l'ont pas été et, avec les cellules existantes d'un autre tableau, pourraient servir à deviner la valeur d'une cellule disparue. On serait donc contraint de consigner les liens entre toutes les cellules, pour tous les tableaux publiés et les tableaux spéciaux. Nous ne disposons tout simplement pas des ressources nécessaires pour cela.

Un autre inconvénient de taille est qu'en supprimant des cellules, on élimine passablement de renseignements dont la divulgation ne poserait aucun risque. En effet, les cellules complémentaires renferment de l'information qui pourrait être publiée si on pouvait protéger d'une autre façon les cellules sensibles, objet de la suppression initiale. La nécessité de supprimer des cellules additionnelles débouche souvent sur la production de tableaux où pullulent les D, surtout aux niveaux de précision les plus élevés. Les utilisateurs se plaignent fréquemment qu'on élimine beaucoup trop de données.

3. ADDITION DE BRUIT AUX MICRODONNÉES AVANT LA MISE EN TABLEAUX

3.1 Généralités

Une autre manière de protéger les données des répondants consiste à les dénaturer. Supposons qu'on fausse légèrement les données de chaque répondant, par exemple de 10 %. Quand une cellule ne se rapporte qu'à un établissement, ou lorsqu'un établissement domine une cellule, la valeur indiquée ne correspondrait pas à une approximation précise de la valeur de l'établissement car elle aurait été altérée. En ajoutant du bruit, on interdirait la divulgation de la véritable valeur se rapportant à l'établissement dominant.

Une dénaturation méticuleuse et méthodique minimiserait aussi les effets du bruit sur les cellules qui, selon la méthode habituelle, ne seraient pas supprimées. De cette façon, il est possible de protéger chaque établissement sans compromettre la qualité des estimations.

Un multiplicateur ou facteur de bruit serait attribué à chaque établissement de l'univers d'échantillonnage. Si l'établissement concerné participe à un sondage ou au recensement, toutes ses valeurs seraient multipliées par le facteur retenu. Dans le cadre d'une enquête ou d'un recensement particuliers, on multiplierait la valeur de tous les établissements par le facteur de bruit correspondant avant la mise en tableaux des données. Notons que l'usage du même multiplicateur pour un établissement, chaque fois où les données se rapportant à ce dernier apparaissent dans un tableau, donnerait des valeurs cohérentes d'un tableau à l'autre. Bref, si une cellule apparaît dans plusieurs tableaux, sa valeur restera la même.

3.2 Les multiplicateurs

Pour fausser les données d'un établissement d'environ 10 %, il faudrait les multiplier par un nombre approchant 1,1 ou 0,9. N'importe quel type de distribution pourrait servir à la sélection du multiplicateur. Par exemple, pour dénaturer les données d'un établissement à la hausse, on pourrait choisir le multiplicateur dans une distribution normale ayant 1,1 pour moyenne et une très faible variance, soit de 0,05 ou de 0,01. Quoi qu'il en soit, on préférera une distribution centrée sur 1,1 (moyenne, mode, etc.) et une variance d'environ 1,1. Pour que tous les multiplicateurs soient égaux à 1,1 ou davantage, ce qui dénaturerait la valeur des cellules ne comptant qu'un établissement d'au moins 10 %, il suffirait de tronquer la distribution à 1,1 et de se débarrasser des valeurs inférieures.

Peu importe la distribution choisie pour obtenir des multiplicateurs voisins de 1,1, il est capital d'utiliser la même distribution, ou plutôt son «image symétrique», pour produire les multiplicateurs voisins de 0,9. Bref, si jamais on venait à examiner les deux distributions, on se rendrait compte qu'elles sont symétriques par rapport à 1. Nous reviendrons à cette condition à la partie 3.3.

Dans la pratique courante, l'unité d'analyse qui nous intéresse sur le plan de la divulgation est la *compagnie*. À savoir, on désire protéger les données des répondants au niveau de la compagnie et des établissements qui la composent. Pour protéger les valeurs de la compagnie, le bruit devra les relever ou les réduire. En d'autres termes, tous les établissements appartenant à la compagnie devraient subir une dénaturation analogue, donc avoir à peu près (mais pas exactement) le même multiplicateur. De cette façon, si la cellule ne comprend que les établissements d'une seule compagnie, l'estimation sera faussée d'environ 10 %. Sinon, l'estimation pourrait s'approcher considérablement de la valeur véritable de la compagnie, car le bruit faussant les établissements positivement (multiplicateur > 1) et celui les faussant négativement (multiplicateur < 1) s'annuleraient pratiquement. En dénaturant tous les établissements d'une compagnie dans le même sens, on protège donc les données au niveau de la compagnie.

Rappelons que les établissements qui appartiennent à la même compagnie ne devraient pas avoir *exactement* le même multiplicateur. Il arrive que ces établissements se livrent concurrence, même s'ils appartiennent à la même société. Supposons que deux petits établissements apparaissent chacun seul dans des cellules distinctes. L'un d'eux pourrait utiliser sa valeur véritable, qu'il connaît, et la valeur inscrite dans la cellule où il est seul pour calculer le multiplicateur à l'origine du bruit. Si on se servait du même multiplicateur pour tous les établissements de la compagnie, cet établissement pourrait découvrir la valeur véritable de l'établissement concurrent en appliquant le multiplicateur qu'il a découvert à la valeur apparaissant dans la cellule qui représente l'établissement concurrent. Ainsi, pour protéger les établissements qui appartiennent à la même compagnie, chacun devrait avoir un multiplicateur légèrement différent.

3.3 Attribution des multiplicateurs et répercussions sur les estimations

On devrait pouvoir se servir des multiplicateurs afin de minimiser l'effet du bruit sur les cellules qui ne présentent aucun risque de divulgation. Les estimations des niveaux d'agrégation supérieurs en posent rarement. Lors des enquêtes économiques et des recensements, les estimations infranationales les plus courantes reposent sur la CTI, la région ou la taille. L'affectation des multiplicateurs devrait dénaturer ces estimations le moins possible.

Une façon d'y arriver consisterait à faire en sorte que parmi tous les établissements qui concourent à une estimation, à peu près autant soient faussés positivement (multiplicateur > 1) et négativement (multiplicateur < à 1). Plus précisément, puisque la taille varie avec l'établissement, il faudrait s'assurer que le *total absolu* de bruit ajouté et le *total absolu* de bruit soustrait s'annulent quand on additionne les valeurs des établissements qui entrent dans l'estimation.

Pour cela, les établissements devraient être triés par CTI × région × taille avant l'attribution d'un multiplicateur. Le premier établissement recevrait un multiplicateur voisin de 1,1; le deuxième un multiplicateur d'environ 0,9; le troisième, un multiplicateur de 1,1; le quatrième, 0,9 et ainsi de suite. Une telle méthode est préférable à l'attribution de multiplicateurs au hasard car ainsi, pour chaque établissement qui reçoit un multiplicateur supérieur à 1, s'en trouve un autre de même importance, du même code CTI et de la même région, qui reçoit un multiplicateur inférieur à 1. Lors du calcul de l'estimation cumulative, le bruit des deux établissements devrait donc avoir tendance à s'annuler.

Théoriquement, l'attribution aléatoire et l'attribution méthodique d'un facteur de bruit devraient faire en sorte que le bruit prévu dans une estimation quelconque ait une valeur espérée de zéro, grâce à la distribution symétrique des multiplicateurs. Néanmoins, l'attribution méthodique tire parti de la stratification invisible inhérente à l'univers des établissements, ce qui devrait contribuer à réduire la *variance* de la quantité de bruit, comparativement à ce qui se produirait si les multiplicateurs étaient attribués au hasard.

Pour les autres cellules non sensibles, on obtient également, qu'en moyenne, les estimations ne sont que légèrement altérées. Dans le cas des estimations cumulatives qui ne sont pas obtenues selon la méthode CTI × région × taille, les établissements dont la valeur est positivement ou négativement dénaturée devraient normalement se compenser dans les cellules détaillées groupant un grand nombre de répondants et les cellules détaillées qui en comprennent très peu, mais d'une taille approximativement identique. La plupart de ces cellules devraient donc inclure très peu de bruit, même si on ne peut le confirmer avec autant d'efficacité que pour les autres estimations cumulatives. Pour ces dernières, l'attribution méthodique d'un facteur de bruit aura le

même effet qu'une attribution aléatoire, quant à la somme moyenne de bruit existant.

Par contre, une cellule où domine un seul répondant présentera vraisemblablement une forte accumulation de bruit. Si le répondant le plus important dépasse les autres de beaucoup dans la cellule, la probabilité que les établissements faussés positivement et négativement s'annulent quand on calculera le bruit total de l'estimation est beaucoup plus faible. En d'autres termes, plus le répondant principal est important, plus la quantité de bruit affectant l'estimation de la cellule s'approchera du bruit qui le dénature (environ 10 %). Les cellules qui posent les plus grands risques de divulgation seront donc aussi celles qui présentent généralement le plus de bruit.

3.4 Addition de bruit aux données d'enquête

Dans les enquêtes, on pondère habituellement les données de chaque répondant de façon inversement proportionnelle à la probabilité d'inclusion de l'établissement. Un poids élevé assurera une certaine protection contre la divulgation de la valeur réelle signalée par le répondant. Pour préserver la protection conférée par le poids de l'échantillon, on ajouterait du bruit de la façon suivante aux données de l'enquête.

Pour chaque établissement dans une cellule, on calculera

valeur de l'établissement × [multiplicateur + (poids - 1)]

puis on additionnera la valeur dénaturée des établissements pour parvenir au total de la cellule. Remarquons qu'on ajoute le bruit à un seul multiple de la valeur de l'établissement, les autres multiples (poids - 1) n'étant pas modifiés.

Cette méthode a pour avantage de changer considérablement la valeur (pondérée) des établissements représentant une certitude ou une quasi-certitude (ceux de poids égal à 1 ou presque) et de dénaturer très peu la valeur pondérée des établissements assortis d'un poids élevé. Un tel mécanisme est souhaitable, car on s'intéresse plus aux risques de divulgation dans le cas des établissements qui peuvent être identifiés avec certitude, puisque le poids ne permet pas d'en protéger les données.

3.5 Actualisation des multiplicateurs

Bon nombre d'enquêtes publient des statistiques exprimant les tendances, c'est-à-dire des statistiques révélant la fluctuation en pour cent d'une variable, d'un point à un autre dans le temps. Appliquer le même multiplicateur à un établissement lors des itérations successives d'une enquête périodique reviendrait à révéler le changement exact en pour cent que subissent les établissements quand ils sont seuls dans leur cellule, bref à en divulguer les données. (Le multiplicateur commun peut être dérivé des estimations à tous les niveaux, ce qui révélerait la véritable variation en pour cent.) Les multiplicateurs devraient donc changer d'une période à l'autre, si on veut protéger les tendances.

Pour que ce type de statistique garde son utilité, il faudrait actualiser les multiplicateurs afin que les valeurs d'un établissement donné soient toujours faussées dans le même sens, à chaque enquête. Bref, si on choisit le multiplicateur initial au moyen d'une distribution centrée sur 0,9 ou presque, tous les autres multiplicateurs devraient être sélectionnés de la même façon. Si le multiplicateur original approche 1,1, tous les nouveaux multiplicateurs en feraient autant. De cette façon, si l'addition de bruit biaise une estimation, malgré tous les efforts déployés pour l'éviter (voir la partie 3.3), au moins le biais sera-t-il approximativement le même d'une période à l'autre, ce qui préservera la tendance. Sans cela, le bruit pourrait masquer la tendance sous-jacente.

En appliquant un multiplicateur analogue au même établissement d'une période à l'autre, on maintiendra les qualités longitudinales des données. De plus, en modifiant légèrement le facteur de bruit entre deux moments dans le temps, on évitera de divulguer la valeur exacte de la variation en pour cent enregistrée par un établissement.

3.6 Application de multiplicateurs différents à différents éléments d'information

En plus de protéger la valeur des éléments d'information fournis par un établissement, il convient de masquer les liens entre différents éléments d'information. En supposant qu'un établissement signale son revenu total et ses composantes, par exemple les revenus tirés de la publicité, dans le cadre d'une enquête, on devra protéger le ratio entre les recettes publicitaires et le revenu global de l'établissement. Cet aspect ne soulève de difficultés que pour les cellules constituées d'un seul établissement. Tant et aussi longtemps que la cellule comprend deux établissements ou davantage, on ne pourra distinguer la part exacte de l'élément d'information appartenant à chaque établissement. Lorsque la cellule ne compte qu'un seul établissement cependant, on sait pertinemment que l'élément d'information (avec ou sans bruit) vient entièrement du répondant. Par conséquent, si on multiplie les éléments d'information du numérateur et du dénominateur d'un rapport par le même facteur de bruit, la simplification du

rapport fera disparaître ce facteur, dévoilant la véritable valeur pour l'établissement.

Pour préserver les relations entre variables, on appliquerait des multiplicateurs différents à chaque élément d'information. Tout en maintenant le multiplicateur de base de l'établissement, on attribuerait un facteur de correction différent à chaque élément d'information publié. Supposons par exemple que le multiplicateur de l'établissement A soit 1,12 et celui de l'établissement B soit 0,87. Au moment de calculer le revenu total, on pourrait multiplier la valeur de l'établissement A par 1,123 et celle de l'établissement B par 0,867. On procéderait donc à une correction de 0,003 par addition de cette valeur (ou soustraction) au multiplicateur de base, lors du calcul du revenu total. La valeur de l'établissement A pourrait être multipliée par 1,125 et celle de l'établissement B par 0,865 au moment de totaliser le revenu tiré de la publicité. On ajouterait donc 0,005 au multiplicateur de base ou on en soustrairait la même valeur en calculant le rapport entre le revenu tiré de la publicité et le revenu total de l'établissement. Ainsi, les corrections différentes au multiplicateur de base des éléments d'information interdiraient la divulgation du rapport exact.

Un des principaux inconvénients de l'usage de multiplicateurs différents pour les éléments d'information est qu'on ne pourrait plus garantir la concordance entre la somme des éléments d'information et le total de l'établissement. Une solution éventuelle consisterait à établir un élément d'information correspondant à l'écart entre la valeur globale et la somme des éléments d'information. Ainsi, on garantirait l'additivité des éléments d'information, mais l'effet sur l'élément d'information sélectionné serait impossible à prévoir.

3.7 Marquage des cellules très dénaturées

Les cellules affectées par un bruit important, un écart de 7 % de la valeur ou davantage par exemple, seraient marquées pour signaler à l'utilisateur que la valeur indiquée pourrait ne pas s'avérer d'une grande utilité. Ces cellules comprendraient les cellules les plus sensibles, plus quelques autres très dénaturées du fait du hasard. Dans la description de l'indicateur, on expliquerait pourquoi et comment le bruit a été ajouté et laisserait savoir à l'utilisateur que certaines restrictions ont été imposées pour éviter la divulgation. On pourrait aussi se servir du même indicateur pour les cellules sensibles qui n'ont pas été très dénaturées en raison de la répartition aléatoire des multiplicateurs. De cette façon, l'utilisateur *croirait* au moins que la valeur dans la cellule est altérée par un bruit important et hésiterait à y voir une valeur fiable. Le nombre de cellules de ce genre serait sans doute peu élevé.

Les cellules trop affectées par le bruit, et les cellules sensibles insuffisamment protégées seraient marquées et ne contiendraient aucune valeur. La valeur de la cellule pourrait être dérivée, mais le fait qu'elle n'apparaisse pas dans le tableau donnerait l'impression qu'on ne juge pas l'estimation fiable. Cette manière de procéder s'apparente un peu à celle relative aux cellules caractérisées par un coefficient de variation (CV) élevé, dans les publications sur les enquêtes. En ne publiant pas les données réelles, on pourrait aussi atténuer la divulgation *apparente* de cellules ne comptant qu'un seul établissement et des cellules sensibles qui ne sont pas touchées par un bruit important.

4. AVANTAGES DU BRUIT

4.1 Simplicité de la procédure

Ajouter du bruit aux données de l'établissement avant la production des tableaux présente plusieurs avantages par rapport à la méthode classique de suppression des cellules. Tout d'abord, la nouvelle technique est beaucoup plus simple et moins laborieuse. Il suffit de multiplier les éléments d'information de chaque établissement par le facteur de bruit, en procédant s'il y a lieu à des ajustements différents selon l'élément d'information, avant la mise en tableaux. Il faudrait encore indiquer dans chaque tableau les cellules sensibles insuffisamment protégées (qu'on supprimerait normalement dès le départ), pour les marquer, mais on échapperait au processus complexe et fastidieux qui consiste à choisir les cellules complémentaires à éliminer. Le bruit ne serait pas spécifique au tableau alors que les cellules complémentaires doivent être sélectionnées pour chaque tableau. De plus, les logiciels servant à ajouter le bruit seraient beaucoup plus facile à élaborer, à modifier, à exécuter et à comprendre.

4.2 Publication plus facile de nombreux produits de données

Un autre grand avantage de cette méthode est qu'on n'aurait plus besoin de coordonner la suppression des cellules entre les tableaux. Actuellement, il est nécessaire d'effectuer une analyse de divulgation distincte pour chaque produit de données. On doit donc garder la trace des cellules qui ont déjà été publiées ou supprimées d'un produit de données à l'autre. (Sinon, une cellule complémentaire supprimée dans un tableau pourrait réapparaître dans un autre.) Il est difficile de suivre les cellules supprimées. Il s'agit d'un exercice

délicat et on le comprend mal. Par contre, la tâche s'avérera superflue si on protège les estimations par l'addition de bruit. À chaque diffusion, il suffirait d'identifier et de marquer les cellules supprimées au départ ou renfermant une quantité de bruit supérieure au seuil de tolérance.

Ne plus avoir à coordonner les suppressions nous permettrait de répondre facilement et rapidement aux demandes de tableaux spéciaux. La protection des estimations par leur dénaturation nous autoriserait à produire autant de tableaux spéciaux que nécessaire sans avoir à vérifier les estimations publiées antérieurement. Encore une fois, il suffirait d'identifier les cellules primaires qui ont été éliminées ou qui contiennent trop de bruit en les marquant.

4.3 Publication de données plus nombreuses dans les tableaux ordinaires

L'addition de bruit a été principalement conçue pour surmonter les difficultés qu'entraîne la suppression des cellules en la présence de tableaux spéciaux multiples, mais elle pourrait aussi déboucher sur la diffusion de données plus nombreuses dans les publications ordinaires. Avec la suppression des cellules, l'utilisateur perd l'information des cellules primaires et complémentaires éliminées. Avec la nouvelle méthode, les cellules sensibles (habituellement supprimées au départ) seraient altérées par un bruit important et seraient marquées en conséquence. Les cellules non sensibles, par contre, subiraient peu de modifications, y compris les cellules complémentaires qu'on aurait supprimées en d'autres circonstances. La technique du bruit devrait donc procurer aux utilisateurs un plus grand nombre de renseignements utiles dans les publications où de nombreuses cellules complémentaires sont habituellement supprimées.

Notons que l'addition de bruit n'améliorera pas énormément les tableaux où la plupart des cellules manquantes sont des cellules primaires. Dans ce cas, seule une diminution du niveau de détail réduira le nombre de cellules pour lesquelles les données ont été éliminées ou considérablement altérées.

5. INCONVÉNIENTS DU BRUIT

5.1 Protection insuffisante des cellules se rapportant à un seul établissement

Certains répondants pourraient soutenir que l'addition de bruit ne protège pas assez la valeur des cellules ne concernant qu'un seul établissement. En effet, si la valeur d'une cellule vient d'un unique établissement, la méthode de la suppression de la cellule fait disparaître la valeur, qui est remplacée par la lettre «D». Avec la technique du bruit, une marque indiquerait que la valeur a été considérablement modifiée, mais peut-être réussira-t-on tout de même à la découvrir par soustraction. L'indicateur pourrait atténuer l'*impression* de divulgation, puisque aucune valeur n'apparaîtrait dans la cellule. Cependant, il se peut que le répondant craigne que la valeur dérivée approche de la valeur réelle, même si elle est très dénaturée et qu'on précise qu'elle n'est pas fiable. La suppression des cellules peut donner l'impression d'une meilleure protection.

Par ailleurs, pour chaque valeur supprimée selon l'approche classique, on peut déduire une fourchette de valeurs. Si on élimine la valeur 100 par exemple, en examinant les cellules voisines, l'utilisateur peut se douter que la valeur manquante se situe entre 84 et 124, par exemple. C'est souvent ce qui se produit. Dans ce cas, l'utilisateur prendra le point milieu comme estimation de la valeur dans la cellule (104 dans notre exemple). Il arrive que le point milieu se trouve près de la valeur réelle, mais pas toujours.

Quelle méthode assure la meilleure protection? Il s'agit d'une question subjective à laquelle il est difficile de répondre. On pourrait tout aussi bien se demander si la suppression des cellules protège *suffisamment* les cellules sensibles.

5.2 Perception de la qualité des données

Certains répondants pourraient hésiter à consacrer du temps à la formulation des réponses appropriées sachant que le *Census Bureau* va les dénaturer. Il faudrait insister sur le fait que le bruit n'est pas ajouté de façon anarchique. Le bruit serait appliqué sans biais, d'une manière contrôlée, afin de sauvegarder les propriétés statistiques des données et de veiller à ce que l'effet sur les estimations non sensibles soit négligeable. Pour préserver les propriétés des données, il faudrait d'abord déterminer celles-ci. Et pour évaluer l'effet du bruit sur d'importantes estimations cumulatives, on devra connaître la valeur véritable de ces dernières, sans bruit, aussi précisément que possible. Il est donc primordial que le bruit soit ajouté à la valeur véritable des données si on veut que celles-ci soient altérées d'une manière prévisible.

Certains utilisateurs pourraient également avoir des appréhensions au sujet de la qualité des données, après addition du bruit. En recourant à cette technique, nous espérons que des tableaux multiples et spéciaux et un plus grand nombre de cellules (même si leur valeur est dénaturée) intéresseront plus l'utilisateur que les valeurs véritables (au prix des cellules supprimées). Une

étiquette attirerait l'attention de l'utilisateur sur les cellules dont les résultats ont été dénaturés par addition de bruit. L'utilisateur sait aussi que les estimations issues des enquêtes sont déjà faussées par l'erreur d'échantillonnage, signalée par le CV, donc ne présentent pas une valeur exacte. Même les données du recensement intègrent un certain «bruit» en raison de diverses erreurs, outre celle de l'échantillonnage (erreurs de réponse, de saisie, de calcul, etc.). En règle générale, l'utilisateur sait qu'on publie la meilleure estimation possible de la valeur de chaque cellule.

Quoi qu'il en soit, le fait qu'on veuille fausser les données sciemment pourrait amener l'utilisateur à croire que les chiffres des tableaux ne constituent pas la meilleure estimation possible. En effet, alors qu'on s'efforce de contrôler et d'éliminer les autres types d'erreur qui affectent les valeurs publiées, l'addition de bruit reviendrait *à introduire intentionnellement* une erreur dans l'estimation. Il conviendrait d'insister sur le but de l'exercice et de rappeler à l'utilisateur que le bruit a été ajouté d'une façon qui en minimise les effets sur les estimations non sensibles.

6. L'ADDITION DE BRUIT, VARIANTES

6.1 Addition de bruit et suppression de certaines cellules

Comme indiqué à la partie 5.1, l'idée de publier la valeur d'une cellule constituée d'à peine un ou deux établissements, même avec beaucoup de bruit, inquiète beaucoup de gens. Pour y remédier, certaines personnes ont demandé si on pourrait à la fois introduire du bruit *et* supprimer les cellules concernées, plus un nombre suffisant de cellules complémentaires. (Les cellules sensibles comptant trois établissements ou davantage ne seraient pas supprimées mais protégées par le bruit et marquées en conséquence.)

On pourrait éliminer les cellules ne comptant qu'un ou deux établissements, mais cette approche présente plusieurs inconvénients. Ainsi, il faudrait appliquer simultanément deux procédures, ce qui rendrait le processus de protection des données plus laborieux et encore plus difficile à comprendre. On devrait toujours coordonner la suppression de cellules entre les tableaux ordinaires et spéciaux, même si les suppressions sont moins nombreuses. Cette approche semble présenter les inconvénients des deux précédentes sans résoudre le problème que posent les multiples tableaux spéciaux.

Le fait qu'une étiquette, *plutôt* qu'une valeur, apparaisse dans la cellule pourrait renforcer l'idée que la valeur est masquée par le bruit. Néanmoins, il reviendra aux répondants de déterminer si le bruit et l'étiquette constituent une protection *suffisante*.

6.2 Suppression de cellules dans les tableaux ordinaires et addition de bruit dans les tableaux spéciaux

Maintes personnes ont demandé si on pourrait supprimer les cellules dans les publications ou les tableaux ordinaires et ajouter du bruit dans les tableaux spéciaux. Pareille pratique pourrait compromettre la protection qu'offre la suppression des cellules. En effet, certaines cellules d'un tableau ordinaire pourraient apparaître dans un tableau spécial. Si les cellules supprimées dans la publication ordinaire sont des cellules primaires, le problème ne se posera pas, car ces cellules incluront considérablement de bruit dans le tableau spécial et seront marquées. Toutefois, les cellules complémentaires éliminées dans la publication ordinaire ne seront pas autant dénaturées dans le tableau spécial, donc ne seront pas marquées. L'utilisateur pourrait donc retranscrire les valeurs du tableau spécial, relativement peu dénaturées, dans les cellules correspondantes du tableau ordinaire puis, par addition et soustraction, obtenir une approximation suffisante des valeurs primaires supprimées au départ. La protection dont bénéficient les cellules primaires s'en trouverait grandement affaiblie.

Une autre difficulté que soulèverait l'addition de bruit uniquement aux tableaux spéciaux est l'incohérence entre les tableaux. Si un tableau spécial reprend certaines cellules d'un tableau ordinaire, la valeur de ces dernières serait faussée dans le premier tableau mais pas dans le second. Trouver la même cellule à deux endroits distincts avec une valeur différente reviendrait à un manque d'uniformité.

6.3 Addition de bruit et itération en fonction de la valeur véritable

6.3.1 Stratégie générale

Une des principales objections à l'addition de bruit est que *toutes* les estimations seraient touchées, pas seulement celles présentant des risques de divulgation. Une solution à ce problème consisterait à ajouter du bruit et à faire en sorte que la valeur publiée des estimations les plus importantes (qu'on présume se situer au degré d'agrégation le plus élevé) corresponde à la valeur véritable (sans bruit). Il suffirait ensuite d'appliquer la méthode d'itération aux cellules intérieures ou d'ajuster leur valeur de façon proportionnelle pour que leur somme soit égale à l'estimation globale.

Il conviendrait d'abord d'établir le degré d'agrégation à partir duquel les estimations devraient correspondre à

leur valeur véritable, soit à leur valeur dépourvue de bruit. Le degré d'agrégation devrait être assez haut pour que le nombre de cellules sensibles soit très faible, voire nul. Le bruit serait appliqué à tous les établissements et les estimations seraient établies après dénaturation. Puis, au degré d'agrégation le plus bas où les estimations devraient correspondre à leur valeur véritable, on contraindrait les estimations dénaturées à reprendre leur valeur réelle, sans bruit, avant de calculer la valeur des cellules intérieures par itération, de manière à préserver l'additivité et les proportions. Dans les tableaux à dimensions multiples, la valeur des cellules intérieures serait obtenue, par itération, simultanément pour tous les totaux marginaux qui sont maintenus à leur valeur véritable. Aux degrés d'agrégation supérieurs, les estimations correspondraient automatiquement à la valeur véritable, puisqu'elles résulteraient de la somme de composants qui ont déjà retrouvé leur valeur réelle par itération.

De cette façon, bon nombre de cellules non sensibles échapperaient au bruit, c'est-à-dire présenteraient leur valeur réelle. L'itération n'aurait pas non plus d'effet très prononcé sur les autres estimations. Puisque les estimations aux degrés d'agrégation supérieurs, où il y aurait itération, ne devraient pas présenter beaucoup de bruit au départ, le facteur d'itération devrait être relativement faible. Les cellules dont la valeur est très dénaturée avant l'itération incluraient encore beaucoup de bruit par la suite, de sorte que les cellules sensibles seraient toujours protégées.

L'itération peut se faire de deux façons, tel qu'indiqué dans les deux parties qui suivent. La première méthode consiste à appliquer l'itération à chaque tableau, séparément; l'autre, à l'appliquer simultanément aux totaux marginaux fixés, avant la production des tableaux. Chaque méthode présente des avantages et des inconvénients.

6.3.2 Itération de chaque tableau

Une approche consiste à appliquer l'itération à chaque produit de données. Cette méthode a pour avantage d'aider l'analyste à déterminer le degré d'agrégation le plus bas auquel l'estimation ne présente aucun risque de divulgation pour chaque tableau (y compris les tableaux spéciaux). Les cellules intérieures du tableau feraient ensuite l'objet d'une itération, mais seulement pour les totaux marginaux qui apparaissent vraiment dans le tableau, peu importe les autres totaux marginaux fixés dans les autres tableaux. En règle générale, un plus grand nombre de cellules présenteraient ainsi une valeur égale à leur valeur véritable, comparativement à l'itération simultanée des estimations en fonction des totaux marginaux qu'on a fixés. Cette seconde méthode est décrite à la partie 6.3.3 et dépend du degré de détail du tableau.

Le même facteur de bruit de *base* s'appliquerait à un établissement dans tous les tableaux. Toutefois, puisque chaque tableau ferait l'objet d'une itération, la somme de bruit *nette* (après itération) applicable à la contribution d'un établissement à différentes cellules pourrait varier d'un tableau à l'autre, voire d'une colonne (ou rangée) à l'autre dans le même tableau. Il pourrait s'ensuivre des problèmes d'incohérence entre tableaux. Pour assurer une certaine uniformité, il est essentiel qu'un groupe de cellules sans bruit dans un tableau le soient également dans tous les autres tableaux où elles apparaissent. Sans cela, la même cellule pourrait présenter des valeurs différentes dans des tableaux distincts. Par conséquent, on serait contraint de garder la trace des cellules dont la valeur a déjà été fixée au fil des tableaux. Ces cellules comprendraient les cellules non dénaturées lors de la production des tableaux spéciaux. Pareille méthode aurait une complexité analogue à celle qui consiste à prendre en note les cellules supprimées. Si jamais on y recourt, il faudra sans doute la réserver aux enquêtes à petite échelle dont les tableaux présentent relativement peu de liens entre eux.

6.3.3 Itération unique avant la mise en tableaux

Pour éviter que la même cellule présente des valeurs différentes dans des tableaux distincts, on pourrait procéder à une seule itération de toutes les estimations. Avant de produire les tableaux, l'analyste déterminerait la série de cellules dont la valeur devrait correspondre à la valeur véritable. Toutes les autres cellules des tableaux publiés feraient ensuite simultanément l'objet d'une itération par rapport aux cellules dont la valeur a été fixée.

On procéderait à cette itération unique en construisant d'abord une «supermatrice» $n \times n$ où n correspondrait au nombre total de variables catégoriques apparaissant dans un tableau quelconque. Après addition du bruit, on affecterait un établissement à une cellule *précise* de la supermatrice, selon la valeur des n variables catégoriques de l'établissement. Ensuite on réaliserait une itération pour n_0 dimensions, n_0 représentant le nombre de variables catégoriques pour lesquelles une partie ou l'ensemble des totaux marginaux correspond à la valeur véritable ($n_0 \leq n$). L'itération établirait le facteur d'ajustement de chaque cellule, indiquant par quel pourcentage la valeur de la cellule a changé consécutivement à l'itération. Ce facteur serait appliqué à chaque établissement se retrouvant dans la cellule. Le facteur de bruit net de l'établissement sera

égal au produit du facteur de bruit original et du facteur de correction résultant de l'itération.

Le facteur de bruit net serait lié à l'établissement pour la production des tableaux ordinaires et spéciaux. De cette façon, on garantirait la cohérence des estimations d'un tableau à l'autre, car les mêmes facteurs de bruit s'appliqueraient à la même série d'établissements dans une cellule, ce qui donnerait constamment la même estimation. La production de tableaux spéciaux ne réclamerait aucune méthode particulière; il suffirait simplement de mettre en tableaux les valeurs de chaque établissement, de les multiplier par le facteur combinant bruit et itération, puis de marquer les cellules pertinentes. On n'aurait pas à se préoccuper des estimations qui apparaissent dans les tableaux antérieurs, ni du bruit qui dénature la valeur.

Le principal inconvénient de cette approche est qu'elle restreint le degré de détail à partir duquel les estimations correspondront à leur valeur véritable. Sans un grand nombre *n* de variables catégoriques, les cellules intérieures de la supermatrice seront généralement peu peuplées. Si on tente d'appliquer l'itération aux cellules intérieures en fonction d'un trop grand nombre de totaux marginaux fixés, on pourrait contraindre certaines cellules à reprendre leur valeur véritable, à seule fin de garantir l'additivité (parfois banale) dans toutes les dimensions de la supermatrice, ce qui annulerait l'effet du bruit. Les analystes devraient aussi limiter le nombre de totaux marginaux fixes; sans quoi l'itération éliminerait la protection assurée par l'addition de bruit aux cellules intérieures, ce qui rendrait bon nombre de cellules sensibles vulnérables.

CONCLUSION

L'addition de bruit aux microdonnées sur les établissements présente clairement des avantages sur la suppression des cellules pour ce qui est d'assurer la protection voulue aux répondants. À mesure que se multiplient les produits de données spéciaux et les tableaux définis par l'utilisateur, cette technique nous donnerait la souplesse voulue pour répondre à des demandes fort variées, sans qu'on ait à s'inquiéter de la divulgation accidentelle des valeurs d'un répondant. Contrairement à la méthode de suppression des cellules, il ne faudrait pas garder en note les demandes antérieures pour éviter qu'un nouveau produit dévoile des renseignements délicats.

Le principal argument contre la nouvelle approche est que toutes les estimations en seraient affectées, et pas seulement celles qu'il faut protéger. Plusieurs modifications ont été proposées afin de résoudre ce problème. Parmi elles, l'idée d'ajouter du bruit, puis de procéder à une itération en fonction des valeurs véritables, aux degrés d'agrégation supérieurs, paraît la plus prometteuse; c'est cette option que nous avons l'intention d'approfondir. Si l'itération apporte une solution satisfaisante au problème de la qualité des données dénaturées, l'addition de bruit aux microdonnées pourrait fort bien devenir la méthode privilégiée au *Census Bureau* pour ce qui est d'éviter la divulgation des données sur les établissements dans les tableaux.

Une version plus détaillée de cette présentation paraîtra dans la série de rapports de la Division de la recherche statistique du *Census Bureau* (Evans et Zayatz).

BIBLIOGRAPHIE

Cox, L.H., et Zayatz, L. (1993). Setting an agenda for research in the Federal Statistical System: Needs for statistical disclosure limitation procedures, *Proceedings of the Section on Government Statistics, American Statistical Association*, 121-126.

Evans, B.T., et Zayatz, L. (1996). Using noise for Disclosure Limitation of Establishment Tabular Data, *Statistical Research Division Report Series*, Bureau of the Census, à paraître.

Federal Committee on Statistical Methodology (1994). *Statistical Policy Working Paper 22: Report on Statistical Disclosure Limitation Methodology*, Washington, DC: U.S. Office of Management and Budget.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

ÉVALUATION ET RÉDUCTION DU RISQUE DE DIVULGATION DANS DES FICHIERS DE MICRODONNÉES À VARIABLES DISCRÈTES

J.-R. Boudreau[1]

RÉSUMÉ

L'estimation du nombre d'éléments uniques dans une population à partir d'un échantillon, et en particulier de la probabilité conditionnelle d'être un élément unique dans la population étant donné d'être un élément unique dans l'échantillon, est essentielle à l'élaboration de règles de confidentialité efficientes pour tout genre de produits de diffusion. Pour le cas des microdonnées à variables discrètes, nous trouvons la relation exacte entre les éléments uniques dans la population et ceux dans l'échantillon. De plus, nous donnons un estimateur sans biais du nombre d'éléments uniques dans la population. Sa grande variabilité échantillonnale pour de petites fractions de sondage nous force à envisager de modéliser cette relation. Après avoir observé cette probabilité conditionnelle pour quantité de populations réelles, nous en donnons une formulation paramétrique. Cette formulation n'est qu'empirique, elle n'a pas de justification théorique. Ce modèle permet cependant d'élaborer des règles de confidentialité. Nous terminons cet article par la description d'une méthode d'identification d'enregistrements possiblement problématiques. Cette technique permet aux concepteurs de microdonnées d'introduire du bruit dans les données qu'aux enregistrements qui en ont réellement besoin.

MOTS CLÉS: Risque de divulgation; confidentialité; microdonnées; unicité; identification.

1. INTRODUCTION

Considérons le problème d'appariement suivant. Un échantillon aléatoire simple (fichier A) est tiré d'une population. Nous voulons apparier ce fichier avec un autre fichier (fichier B) provenant de cette même population en utilisant toutes les variables discrètes communes aux deux fichiers. Nous supposons que les erreurs de saisies et de réponses sont négligeables. Si un appariement biunivoque est obtenu entre deux enregistrements, quel est le «niveau de confiance» que nous pouvons accorder à l'énoncé : «ces deux enregistrements proviennent de la même unité de la population»?

L'auteur y voit une application immédiate. Ce niveau de confiance peut servir à évaluer le risque de divulgation du fichier A. En effet, une agence statistique diffuse un fichier de microdonnées A. Un individu ou une entreprise, ayant en possession un fichier de microdonnées B ayant d'une part une clé unique (par exemple: noms et adresses) et d'autre part quelques variables communes avec A, peut opérer un appariement avec ce dernier dans le but d'identifier la provenance de certains enregistrements. Sous cet angle, il faut que l'agence statistique s'assure avant la diffusion de son fichier que le niveau de confiance soit le plus bas possible; ceci afin d'enlever toute incitation à apparier le fichier diffusé avec d'autres fichiers.

Une condition nécessaire pour avoir un haut niveau de confiance est d'imposer que le fichier B couvre bien la population. Sous cette hypothèse, le niveau de confiance est directement relié à la probabilité conditionnelle d'être un élément unique dans la population (par rapport aux variables d'appariement) étant donné d'être un élément unique dans l'échantillon. Comme nous allons le voir par la suite, la détermination de cette probabilité est liée en partie à l'estimation du nombre d'éléments uniques dans la population à partir de l'échantillon.

L'estimation du nombre d'éléments uniques dans la population a fait l'objet de beaucoup de recherches ces dernières années. Greenberg and Zayatz donnent deux façons d'estimer le nombre d'éléments uniques. La première consiste à ré-échantillonner l'échantillon selon le même plan de sondage. L'estimateur est construit en supposant que les relations entre les éléments uniques de

[1] Jean-René Boudreau, Statistique Canada, Division des méthodes d'enquêtes sociales, Édifice R.H. Coats, 15ième étage, Section P, Parc Tunney, Ottawa (Ontario), K1A 0T6.

la population et du premier échantillon sont les mêmes entre ceux du premier et du deuxième échantillon. La deuxième façon proposée par ces auteurs utilise la structure de la population, c'est-à-dire la description de la population en termes du nombre de cellules définies par les variables d'appariement ayant exactement une unité, deux unités, etc... Ce qu'ils appellent «classes d'équivalences». Nous utiliserons cette technique comme point de départ. Ces deux techniques donnent de bons résultats si la fraction de sondage est supérieure à 10%. Une autre façon de procéder est d'essayer de modéliser la structure de la population à partir d'un échantillon. Bethleem, et coll. ont tenté de modéliser la proportion du nombre d'éléments uniques dans la population à l'aide d'un modèle dérivé de la loi Poisson-Gamma. Ce modèle souffre d'un manque d'ajustement important. Skinner et coll. a modélisé la proportion d'uniques dans la population en utilisant la loi Poisson-Log-normal. Il obtient des résultats qui collent beaucoup plus à la réalité.

L'auteur propose d'utiliser la théorie de l'échantillonnage pour déterminer exactement la forme de la relation entre les éléments uniques dans la population et ceux dans l'échantillon. Nous démontrerons qu'il est pratiquement futile de vouloir solutionner le problème pour les petites fractions de sondage en n'utilisant que la théorie de l'échantillonnage. Par conséquent, nous essaierons de modéliser cette relation pour de petites fractions de sondage. Nous donnerons aussi une façon de procéder, quoiqu'elle soit encore au niveau des conjectures, pour borner supérieurement la probabilité conditionnelle : permettant ainsi de contrôler le risque de divulgation d'un fichier de microdonnées.

2. DÉTERMINATION DE LA PROBABILITÉ CONDITIONNELLE

Nous avons une population de N éléments ou unités. Le contenu, c'est-à-dire les variables d'appariement, partitionne cette population en m sous-populations de taille $N_1, ..., N_m$. La structure de la population est donnée par le vecteur $(U_1, ..., U_N)$ où U_j = card $\{k: N_k=j\}$. Nous prenons un échantillon de taille n tiré d'une manière aléatoire simple de cette population. Nous observons le vecteur aléatoire $(n_1, ..., n_m)$ dont les composantes sont respectivement le nombre d'unités échantillonnées de la sous-population k $(k = 1, ... ,m)$. La structure de l'échantillon est le vecteur aléatoire $(u_1 ,...,u_n)$ où u_j = card $\{k: n_k=j\}$.

Un élément sera dit unique dans la population s'il appartient à une sous-population de taille unité. Une unité échantillonnée sera dite unique dans l'échantillon si elle est la seule unité échantillonnée à appartenir à sa sous-population. Puisqu'un élément unique dans la population qui est échantillonné est nécessairement unique dans l'échantillon, nous obtenons que la probabilité conditionnelle d'être unique dans la population étant donné d'être unique dans l'échantillon est le rapport entre les proportions des éléments uniques dans la population et dans l'échantillon. Donc, nous voulons avoir une estimation de

$$P = \frac{\dfrac{U_1}{N}}{\dfrac{E\{u_1\}}{n}} = f\,\frac{U_1}{E\{u_1\}}$$

où f est la fraction de sondage et l'espérance mathématique est celle établie par le plan de sondage. L'espérance est nécessaire pour obtenir un paramètre au niveau de la population. Ce paramètre, qui par abus de langage sera tout de même considéré comme une probabilité conditionnelle, n'est pas loin de l'idée du risque de divulgation ou du niveau de confiance expliqué à la section précédente. Nous avons un premier résultat.

Théorème A. *Si un échantillon aléatoire simple de taille n est tiré d'une population de taille N possédant la structure $(U_1, ... , U_N)$, alors*

$$E\{u_j\} = \frac{\binom{N-j}{n-j}}{\binom{N}{n}}\,U_j + \sum_{i=1}^{\infty} \frac{\binom{j+i}{j}\binom{N-j-i}{n-j}}{\binom{N}{n}}\,U_{j+i}\,.$$

Démonstration. La somme est en réalité finie. Puisque u_j est à valeurs entières, nous pouvons utiliser l'identité

$$E\{u_j\} = \sum_{i=1}^{\infty} P\{u_j \geq i\}\,.$$

Posons $A_k = \{(n_1 ,...,n_n): n_k=j\}$. Nous avons l'identité suivante

$$P\{u_j \geq i\} = P\left\{ \bigcup_{k_1<\cdots<k_i} A_{k_1} \cdots A_{k_i} \right\}.$$

Nous pouvons montrer facilement que

$$\sum_{i=1}^{\infty} P\{u_j \geq i\} = \sum_{k=1}^{m} P\{A_k\}\,.$$

En effet, il suffit de déterminer la probabilité de chaque union et de réaliser que tous les termes s'annulent sauf la somme des probabilités des événements A_k. Maintenant, $P\{A_k\}$ vaut

$$P\{A_k\} = \frac{\binom{N_k}{j}\binom{N-N_k}{n-j}}{\binom{N}{n}}.$$

Donc l'espérance de u_j vaut

$$E\{u_j\} = \sum_{\substack{k=1 \\ j \le N_k \le N-n+j}}^{m} \frac{\binom{N_k}{j}\binom{N-N_k}{n-j}}{\binom{N}{n}}$$

$$= \sum_{i=j}^{\infty} \frac{\binom{i}{j}\binom{N-i}{n-j}}{\binom{N}{n}} U_i .$$

Ce qu'il fallait démontrer.

En particulier pour $j = 1$, nous avons

$$E\{u_1\} = f\,U_1 + \sum_{i=1}^{\infty} (i+1)\frac{\binom{N-1-i}{n-1}}{\binom{N}{n}} U_{1+i},$$

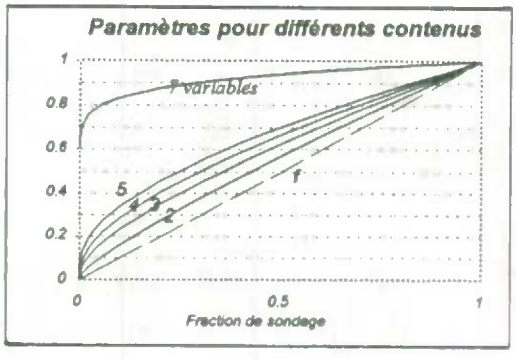
qui peut s'écrire comme

$$E\{u_1\} = f\,U_1 + n\sum_{i=1}^{N-n} \frac{(i+1)}{N-i}\left(1-\frac{n}{N}\right)\cdots\left(1-\frac{n}{N-i+1}\right)U_{1+i}$$

$$\approx f\left(U_1 + \sum_{i=1}^{N(1-f)} (i+1)\,(1-f)^i\,U_{1+i}\right)$$

si N est suffisamment grand. Ainsi la probabilité conditionnelle devient

$$P = \frac{1}{1 + \sum_{i=1}^{N(1-f)} (i+1)\,(1-f)^i \frac{U_{1+i}}{U_1}}.$$



Ce graphique donne la relation entre P et la fraction de sondage pour différents contenus. La taille de la population est près de 800 000. Nous avons extrait sept variables de cette population. Le premier contenu est défini par les deux premières variables extraites ; le deuxième est défini par les trois premières variables et ainsi de suite. Plus le nombre de variables est important, plus la probabilité conditionnelle est élevée. La région intéressante est sans contredit l'intervalle [0, 0.1]. La valeur de P à l'origine est donnée en posant $n = 1$ dans la formule exacte ou en posant $f = 0$ dans la formule approximative de P. La valeur de P à ce point donne $P=U_1/N$, qui est exactement la proportion d'éléments uniques dans la population. Le comportement de la courbe à l'origine est analysée en développant P autour de 0. Pour ce faire, introduisons la loi de probabilité $p(\cdot) = (p_1, p_2, \ldots)$ définie par

$$p_i = \frac{i\,U_i}{N}$$

pour $i \ge 1$. Si nous notons l'espérance de cette loi par $E_p\{\cdot\}$, la probabilité conditionnelle peut être écrite sous la forme

$$P = p_1 \frac{1-f}{E_p\{(1-f)^X\}}$$

pour tout $f \le \min_{1 \le k \le m} (1 - N_k/N)$ et $X(i) = i$.

Sous cette forme, nous voyons aisément que les dérivées d'ordre quelconque de P évaluées à l'origine sont des sommes de moments centrés à l'origine de la loi p (le dénominateur étant la fonction génératrice des moments). Les premiers termes du développement sont

$$P = p_1 + p_1(\mu_p - 1)\,f + \frac{p_1}{2}\left(\mu_p^2 - \mu_p - \sigma_p^2\right)f^2 + o(f^3).$$

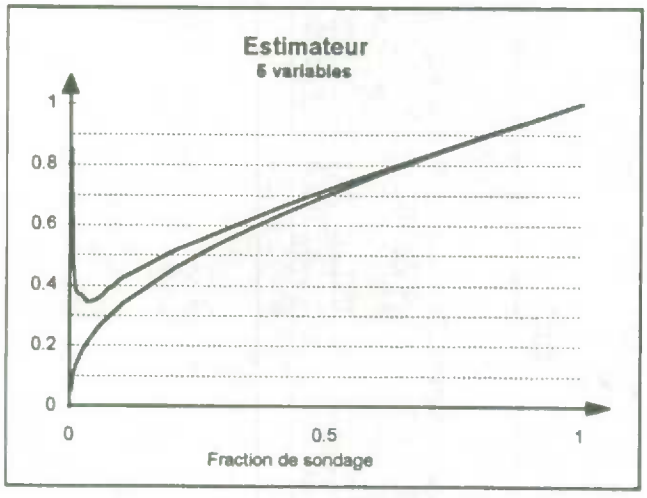
Cette expression nous indique que P est croissante dans le voisinage de 0 et est convexe ou concave dépendant si le coefficient de variation de $p(\cdot)$ est plus petit ou plus grand que $1 - 1/\mu_p$. Il semble qu'une courbure concave soit un fait général pour les populations réelles. Cet énoncé est très important. La relation entre P et f peut donner un peu n'importe quoi si seulement la condition $N = U_1 + 2U_2 + \ldots + NU_N$ est vérifiée. En fait, cette dernière condition n'est pas suffisante pour que la population ayant cette structure puisse avoir le qualificatif de «réelle» ou «observée». Nous ne savons pas comment les populations réelles sont simulées mais il appert que la distribution de $p(\cdot)$ entraîne la concavité de la relation.

3. ESTIMATEUR NATUREL DE P

La forme de l'expression pour P suggère, comme estimateur de cette quantité, de remplacer la structure de la population par la structure de l'échantillon, c'est-à-dire l'estimateur

$$\hat{P} = \frac{1}{1 + \sum_{i=1}^{N(1-f)} (i+1)\,(1-f)^{i}\,\frac{u_{1+i}}{u_1}},$$

estimateur qui se trouve dans la ligne de pensée de Greenberg et Zayatz.



L'estimateur converge vers 1 lorsque la fraction de sondage tend vers 0 au lieu de converger vers la proportion d'éléments uniques de la population. Ce comportement désastreux est dû au fait que la proportion d'éléments uniques dans l'échantillon devient disproportionnée, causé par la petitesse de la taille d'échantillon. Nous allons donner à la section suivante une mesure de la divergence entre les structures de la population et de l'échantillon pour de petites fractions de sondage. Ces résultats nous forceront à modéliser P sur ce domaine.

4. ESTIMATEUR SANS BIAIS DE U_j

Pour voir comment la structure de l'échantillon s'éloigne de celle de la population lorsque la fraction de sondage tend vers 0, nous proposons de construire un estimateur sans biais pour les composantes de la structure de la population. Ce résultat se trouve dans Goodman. La divergence des structures sera transformée en variation échantillonnale de l'estimateur.

Pour construire cet estimateur, nous avons besoin du lemme suivant.

Lemme B. *Pour $1 \le k \le r$, nous avons l'identité suivante*

$$\sum_{i=1}^{k} (-1)^{i+1} \binom{k}{i} \binom{r+i-1}{k-1} = \binom{r-1}{k-1}. \quad (1)$$

Démonstration. Nous savons que le terme de gauche de (1) vaut

$$\frac{1}{(k-1)!} \sum_{i=1}^{k} (-1)^{i+1} \binom{k}{i} \frac{i(i+r-1)\cdots(i+r-k+1)}{i} \quad (2)$$

Posons $g(x) = x(x + r - 1) \ldots (x + r - k + 1)$. $g(\cdot)$ est un polynôme de degré k. Nous pouvons donc appliquer l'identité (Mercier, Corollaire 1, p. 141)

$$\sum_{i=1}^{k} (-1)^{i+1} \binom{k}{i} \frac{g(i)}{i} = g'(0) + g(0) \sum_{i=1}^{k} \frac{1}{i}.$$

Puisque $g(0) = 0$ et $g'(0) = (r - 1)! / (r - k)!$, nous obtenons que (2) vaut le coefficient binomial du terme de droite de (1). Ce qui complète la démonstration.

Nous avons le théorème suivant.

Théorème C. *Un échantillon aléatoire simple de taille n est tiré d'une population ayant la structure $(U_1, \ldots, U_N)$. Si $n \ge N_k$ $(k = 1, \ldots, m)$, alors un estimateur sans biais de U_j est donné par*

$$\hat{U}_j = \frac{\binom{N}{n}}{\binom{N-j}{n-j}} u_j - \frac{\binom{N}{n}}{\binom{N-j}{n-j}} \sum_{i=1}^{\infty} (-1)^{i+1} \frac{\binom{j+i}{i}\binom{N-n+i-1}{i}}{\binom{n-j}{i}} u_{j+i}$$

Démonstration. D'après le théorème A, nous avons que

$$E\{u_j\} = \sum_{i=j}^{\infty} \frac{\binom{i}{j}\binom{N-i}{n-j}}{\binom{N}{n}} U_i.$$

Si nous remplaçons les u_j par leur espérance, nous obtenons

$$E\{\hat{U}_j\} = U_j + \sum_{k=1}^{\infty} \frac{\binom{k+j}{j}\binom{N-k-j}{n-j}}{\binom{N-j}{n-j}} U_{j+k}$$

$$-\sum_{i=1}^{\infty}\sum_{r=i+j}^{\infty} (-1)^{i+1} \frac{\binom{j+i}{i}\binom{N-n+i-1}{i}}{\binom{n-j}{i}} \frac{\binom{r}{i+j}\binom{N-r}{n-i-j}}{\binom{N-j}{n-j}} U_r$$

$$= U_j + \sum_{k=1}^{\infty} \frac{\binom{k+j}{j}\binom{N-k-j}{n-j}}{\binom{N-j}{n-j}} U_{j+k}$$

$$-\sum_{k=1}^{\infty}\sum_{i=1}^{k} (-1)^{i+1} \frac{\binom{j+i}{i}\binom{N-n+i-1}{i}}{\binom{n-j}{i}} \frac{\binom{k+j}{i+j}\binom{N-k-j}{n-i-j}}{\binom{N-j}{n-j}} U_{j+k}.$$

Les termes binomiaux de la deuxième somme se simplifient pour donner

$$\sum_{k=1}^{\infty} \frac{\binom{k+k}{j}}{\binom{N-j}{n-j}} \frac{(N-j-k)!\,(k-1)!}{(N-n-1)!\,(n-j)!}$$

$$\left\{\sum_{i=1}^{k} (-1)^{i+1} \binom{k}{i} \binom{N-n+i-1}{k-1}\right\} U_{j+k}.$$

Ce qui donne, d'après le lemme B,

$$\sum_{k=1}^{\infty} \binom{j+k}{j} \binom{N-n-1}{k-1} \frac{(N-j-k)!\,(k-1)!}{(N-n-1)!(n-j)!} U_{j+k}$$

$$= \sum_{k=1}^{\infty} \frac{\binom{j+k}{j}\binom{N-j-k}{n-j}}{\binom{N-j}{n-j}} U_{j+k}.$$

Les deux sommes s'annulent et nous obtenons le résultat recherché.

Si N est suffisamment grand, nous pouvons utiliser l'approximation suivante

$$\hat{U}_j \approx f^{-j}\left\{u_j - \sum_{i=1}^{\infty} (-1)^{i+1} \binom{j+i}{i} (f^{-1}-1)^i u_{j+i}\right\}$$

à la place de l'estimateur. En particulier, pour $j = 1$, nous obtenons

$$\hat{U}_1 \approx f^{-1}\left\{u_1 - \sum_{i=1}^{\infty} (-1)^{i+1} (i+1) (f^{-1}-1)^i u_{1+i}\right\}.$$

Pour $f < 0{,}5$, nous voyons clairement que cet estimateur est inutilisable car sa variance échantillonnale grandit exponentiellement lorsque N grandit. Notons que la condition $n \geq N_k$ pour tout k est nécessaire pour avoir un estimateur sans biais. Si nous gardons cet estimateur même si la condition n'est pas vérifiée, un biais sera introduit qui ne réduira pas l'erreur quadratique moyenne de façon substantielle car le problème se situe à la grandeur de la taille d'échantillon. Par exemple, pour une population de 1 000 000 et une fraction de sondage de 0,001, la taille de l'échantillon est de 1 000. Nous avons ainsi de très fortes chances que le plus grand des indices i pour lequel $u_{1+i} > 0$ fera en sorte que l'estimation ne voudra plus rien dire (si $u_5 = 1$, un des termes de l'estimateur sera de l'ordre 10^{15}).

Ceci montre bien que pour estimer la probabilité conditionnelle ou une composante de la structure de la population pour de petites fractions de sondage, la structure de l'échantillon n'a que peu de valeur. Cela jette un froid à ceux ou celles voulant une solution non paramétrique. Tournons-nous maintenant vers les possibilités de paramétrisation de la relation entre P et f.

5. PARAMÉTRISATION DE LA PROBABILITÉ CONDITIONNELLE

Au cours des dernières années, plusieurs personnes ont tenté avec plus ou moins de succès de modéliser la structure de la population (en particulier le nombre d'éléments uniques dans la population). La première tentative connue de l'auteur est celle de Bethleem et coll. Ils ont supposé, sans justification outre celle de la simplicité des techniques, que la structure d'une population pourrait être simulée à partir d'un modèle Poisson-Gamma. Sous cette hypothèse, on en vient facilement à trouver une expression paramétrique pour la proportion d'éléments uniques dans la population. En fait, l'expression est donnée par

$$E_m\{U_1/N\} = \left(\frac{1}{1+\beta N}\right)^{1+\alpha},$$

où α et β sont les paramètres de la loi Gamma du modèle ($\alpha, \beta > 0$). Dès que l'on essaie, à partir d'un échantillon, d'estimer ces paramètres, on s'aperçoit vite que le modèle souffre d'un manque d'ajustement. Le paramètre α est invariablement estimé à une valeur non significativement différente de zéro. Même les techniques classiques de compensation ne permettent pas de stabiliser le modèle. Nous verrons plus loin que le problème se situe au niveau de l'étendue de définition

de α. Skinner et coll., propose une approche basée sur la théorie de la classification. Selon cette théorie, la structure de la population serait simulée par un modèle Poisson-Lognormal. Ce modèle est beaucoup plus difficile à maîtriser que celui énoncé précédemment ; en particulier, l'estimateur de la proportion d'éléments uniques dans la population est la solution implicite d'une équation intégrale n'ayant pas de primitive. D'après les résultats obtenus par Skinner sur des populations italiennes, ce modèle semble très bien coller à la réalité. L'approche développée dans cet article tente de modéliser directement la relation entre P et f au lieu de modéliser la structure sous-jacente. Cette approche, de nature purement empirique, a l'avantage de coller à la réalité si on est en mesure de pouvoir observer un grand nombre de populations réelles. Un désavantage toutefois de cette méthode est qu'elle ne donne aucun renseignement sur comment ces populations sont générées. Autrement dit, elle ne donne aucune justification théorique ni n'en suggère.

Cette approche empirique ne suppose aucune hypothèse probabiliste sauf celle de la sélection d'un échantillon aléatoire simple. La technique consiste à étudier la relation entre P et f pour plusieurs populations obtenues par voie de recensements, de tenter de décanter les ressemblances, de proposer une formulation paramétrique de P en fonction de f, et de proposer une méthode d'estimation des paramètres en utilisant un échantillon. Le but ultime de cette démarche est de proposer une série de règles de confidentialité pour assurer, si le modèle tient bon et si la méthode d'estimation est satisfaisante, une probabilité conditionnelle la plus basse possible.

5.1 Formulation de la relation entre P et f

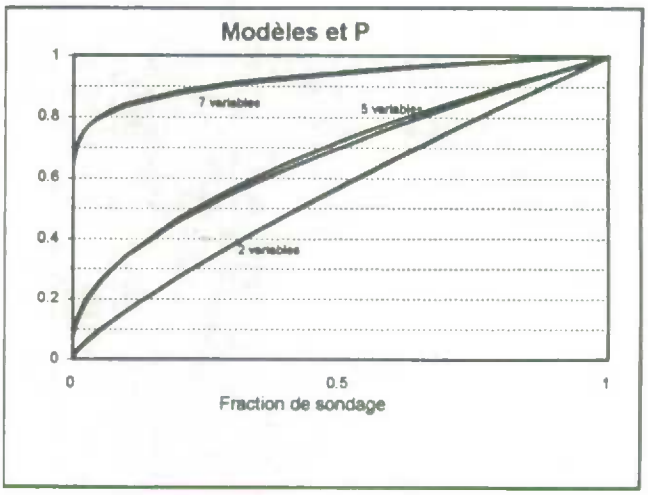
La formulation de cette relation qui colle très bien avec l'observation est donnée par l'expression suivante

$$P_M = \left| \frac{f + \gamma}{1 + \gamma} \right|^{\alpha},$$

où $0 < \alpha < 1$ et $\gamma > 0$. Le paramètre α influe directement sur la concavité observée de la relation ; γ est lié directement au taux d'unicité dans la population. Pour la population et les contenus donnés en exemples à la section précédente, nous obtenons les valeurs des paramètres α et γ. Ces valeurs ont été obtenues par la méthode des moindres carrés.

Contenu	Alpha	Gamma
2 variables	0,805167	0,002085
3 variables	0,643916	0,002484
4 variables	0,548488	0,001771
5 variables	0,471880	0,003014
7 variables	0,075045	0,001915

Les graphiques suivants montrent que le modèle cité peut s'ajuster assez bien pour différents contenus.



Nous ne pouvons pas utiliser directement la relation entre P et f pour estimer les paramètres α et γ puisque la probabilité conditionnelle n'est pas observable. Les seules quantités d'intérêt observables sont les composantes de la structure de l'échantillon. Essayons de dériver l'espérance du nombre d'éléments uniques dans un échantillon à partir de P et f.

Théorème D : *Si la formulation paramétrique entre P et f est correcte avec paramètres α et γ, alors l'espérance, au sens du modèle, de la proportion du nombre d'éléments uniques dans l'échantillon est donnée par*

$$Q_n = E_m \{ u_1/n \} = \left(\frac{1 + \beta}{1 + \beta n} \right)^{\alpha},$$

où β est le réciproque de la multiplication de γ par la taille de la population.

Démonstration : Par définition, la probabilité conditionnelle recherchée est le quotient des proportions d'éléments uniques dans la population et dans l'échantillon respectivement. Puisque la formulation entre P et f est correcte, nous avons

$$P = \left(\frac{\frac{n}{N} + \gamma}{1 + \gamma} \right)^{\alpha} = \left(\frac{1 + \frac{n}{\gamma N}}{1 + \frac{N}{\gamma N}} \right)^{\alpha}$$

$$= \left(\frac{1 + \beta n}{1 + \beta N} \right)^{\alpha} = \frac{Q_N}{Q_n} .$$

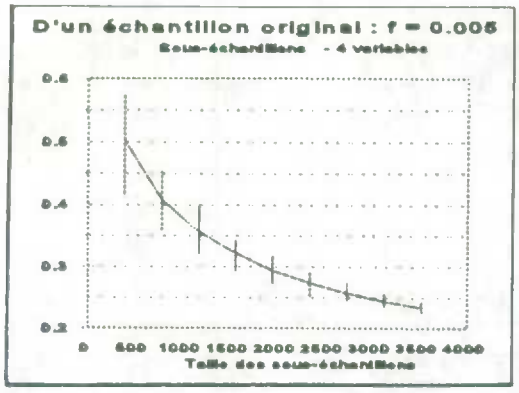
Ce qui donne que

$$Q_n = K \left(\frac{1}{1 + \beta n} \right)^{\alpha} .$$

Puisque $Q_1 = 1$, nous obtenons le résultat recherché.

Ce théorème nous dit pourquoi le modèle Poisson-Gamma a un si grand manque d'ajustement. Le modèle Poisson-Gamma implique une relation convexe entre P et f. Mais nous savons par observation que la relation est concave. Ainsi, vouloir ajuster le modèle Poisson-Gamma a la réalité ne peut que donner quelque chose proche de la linéarité ($\alpha = 0$). C'est exactement ce qu'on peut lire dans la littérature.

On peut se demander si la relation entre Q_n et n colle à la réalité. Nous avons pris comme exemple la population de taille 800 000 avec un contenu de quatre variables duquel nous avons sélectionné un premier échantillon avec une fraction de sondage de 0,005. À partir de cet échantillon, nous avons sélectionné d'une manière complètement indépendante 900 échantillons : 100 échantillons avec une fraction de sondage de 0,1 ; 100 avec une fraction de 0,2 ; ... ; 100 échantillons avec une fraction de 0,9. Puisque l'échantillon premier et les autres sont tirés selon le plan aléatoire simple, tous ces échantillons sont tirés par un plan aléatoire simple (seule la fraction de sondage change). Le graphique suivant nous donne la courbe empirique de la relation entre Q_n et n:



Les droites verticales donnent une indication de la variabilité de la proportion d'éléments uniques dans les échantillons. La loi régissant les écarts est extrêmement difficile à modéliser. Par contre, la courbe des moyennes de ces proportions reflète bien le modèle.

5.2 Estimation des paramètres α et γ

Le modèle présenté à la sous-section précédente semble bien coller à la réalité. Il ne reste qu'à trouver une méthode d'estimation des paramètres. Le modèle s'écrit de la manière suivante :

$$u_1/n = \left(\frac{1 + \beta}{1 + \beta n} \right)^{\alpha} + \epsilon ,$$

où ϵ est régi par la loi des écarts des u_1. Les méthodes standards d'estimation des paramètres de la sorte dépendent lourdement de cette loi. Puisque nous ne la connaissons pas et nous ne sommes pas en mesure d'émettre des hypothèses, nous allons plutôt utiliser les réalisations des moyennes Q_n et de supposer que ces points seront près de la courbe si les moyennes sont basées sur plusieurs expériences (par ex : 100 échantillons). La méthode est la suivante :

I. Sélectionner des échantillons aléatoires simples répétés de l'échantillon original selon plusieurs fractions de sondage (ex : 0,1, 0,2, ... , 0,9). Le nombre de répétitions pour chacune de ces fractions de sondage doit être élevé.
II. Pour chacune des fractions de sondage, calculer les moyennes du nombre d'éléments uniques dans l'échantillon (Q_n).
III. Utiliser une méthode numérique[2] pour déterminer les paramètres α et β qui collent le plus à l'observation.
IV. Déterminer γ à partir de β.
V. Calculer les probabilités conditionnelles à partir du modèle.

---

[2] Nous avons utilisé l'algorithme NEWTON programmé dans la procédure NLIN du progiciel SAS (version 6.10).

Avec les données des exemples antérieurs, si l'échantillon original a une fraction de sondage de 0,005, nous trouvons les estimations suivantes pour la probabilité conditionnelle :

	Alpha		Gamma		Probabilité conditionnelle	
Contenu	Vraie valeur	Estimation	Vraie valeur	Estimation*	Vraie valeur	Estimation
2 variables	0,805167	0,538648	0,002085	0,0000222	0,0219	0,0577
3 variables	0,643916	0,422989	0,002484	0,0000981	0,0501	0,1072
4 variables	0,548488	0,374097	0,001771	0,0000961	0,0744	0,1387
5 variables	0,471880	0,274312	0,003014	0,0001914	0,1146	0,2361
7 variables	0,075045	0,068793	0,001915	0,0011052	0,6949	0,7040

* Même si les estimations sont proches de zéro, il sont tout de même significativement différents de zéro si l'on se fie aux diagnostiques de l'algorithme NEWTON.

Deux critiques doivent être immédiatement être mentionnées sur cette façon d'estimer. Premièrement, il n'y a rien dans la méthode qui puisse contrecarrer les variations des proportions du nombre d'éléments uniques dans l'échantillon. Peut-être serait-il sage d'utiliser une variante de la méthode des moindres carrés en pondérant la distance entre le modèle et les points obtenus par observation. Deuxièmement, ce qui est encore plus sérieux, est l'extrapolation nécessaire pour estimer la probabilité conditionnelle. En effet, nous devons faire le rapport entre Q_N et Q_n. Estimer Q_N revient à extrapoler d'une manière indue. Par contre la méthode sous-estime systématiquement les paramètres et surestime la probabilité conditionnelle. Il est clair que cette méthode a des lacunes importantes à plusieurs points de vue, mais avant de la jeter aux oubliettes, il serait bon d'examiner si la surestimation des probabilités est fortuite ou systématique pour les populations réelles. Car si cela n'est pas fortuit alors la méthode d'estimation est alors permissible puisqu'elle donnerait des évaluations conservatrices du risque de divulgation. Faisons la conjecture suivante :

Conjecture E : *La méthode d'estimation proposée donne toujours une surestimation de la probabilité conditionnelle pour les population réelles.*

La suite de cette section suppose que cette conjecture est vraie.

5.3 Règles de confidentialité

Une règle de confidentialité est définie comme toute mesure appliquée sur les données afin d'en réduire le risque de divulgation. Dans le cas des microdonnées discrètes, ces mesures sont divisées en deux groupes : la suppression et l'introduction de bruit dans les données. Une règle est dite globale si elle s'applique à tous les enregistrements ; autrement, elle est dite locale. Un exemple d'introduction de bruit globale est tout regroupement de valeurs pour une ou plusieurs variables. L'imputation de nouvelles valeurs pour certains enregistrements est une méthode d'introduction de bruit locale. La suppression d'une donnée est le recodage de cette donnée par une nouvelle catégorie signifiant «suppression pour cause de confidentialité». L'élimination d'un enregistrement ou la suppression d'une donnée ou plusieurs données pour cet enregistrement est une suppression locale. Une suppression globale est l'élimination d'une variable du fichier. Puisque le risque de divulgation et la probabilité conditionnelle sont intimement liés, les règles de confidentialité doivent réduire le plus possible cette probabilité.

En regardant la relation entre P et f, nous pouvons faire les observations suivantes :

I. Pour un même contenu, le fichier le plus sécuritaire, c'est-à-dire celui qui donne la plus petite probabilité conditionnelle, est celui ou la taille d'échantillon est l'unité.

II. Même pour de très petites fractions de sondage, les fichiers ayant des contenus «explosifs» peuvent avoir des risques de divulgation très grands.

Un exemple de contenu explosif est celui donné par sept variables. Les deux dernières variables, sixième et septième, sont respectivement les codes d'industrie et d'occupation à quatre chiffres pour les particuliers dans la population active. Bien entendu qu'il n'est pas possible d'avoir ce niveau de détail dans un fichier sécuritaire. Par contre, il se peut très bien que pour des fichiers longitudinaux, par exemple, le risque de divulgation de chaque partie transversale soit acceptable mais que le risque soit inacceptable pour le fichier en entier. Prenons comme exemple l'état matrimonial au Canada d'un particulier. Cette variable a cinq valeurs : célibataire, marié, séparé, divorcé et veuf. Seuls, ces cinq valeurs ne peuvent pas rendre cette variable très «dangereuse», mais si cette variable est demandée lors d'une enquête longitudinale sur dix ans, l'information apportée par cette variable pour évaluer le risque doit contenir également toutes les transitions des valeurs de cette variable dans le temps. Ainsi, au lieu d'avoir une variable à cinq valeurs, nous devons considérer plutôt une variable représentant les transitions de l'état matrimonial. Cette nouvelle variable peut contenir plusieurs centaines de valeurs : faisant exploser ainsi le contenu du fichier.

Donc des petites fractions de sondage et des contenus non explosifs garantissent de faibles probabilités conditionnelles. Si le modèle et la conjecture tiennent bon, une proportion d'éléments uniques dans l'échantillon pas trop élevée et une estimation de α pas trop basse sont suffisantes pour garantir un contenu non explosif. Le théorème suivant nous permet de déterminer une borne supérieure pour la proportion du nombre d'éléments uniques dans l'échantillon.

Théorème F : *D'après le modèle, nous avons*

$$P = \left(\frac{f}{1 - Q_n^{1/\alpha}\,(1-f)} \right)^{\alpha},$$

Ainsi, si la conjecture est vraie, en spécifiant une probabilité conditionnelle P^* à ne pas dépasser, l'expression suivante

$$\left(\frac{1 - \frac{f}{(P^*)^{1/\hat{\alpha}}}}{1-f} \right)^{\hat{\alpha}},$$

$\hat{\alpha}$ étant l'estimation de α, donne une limite supérieure pour l'espérance de la proportion d'éléments uniques dans l'échantillon.

Démonstration. La première identité se démontre ainsi.

$$Q_n = \left(\frac{1+\beta}{1+\beta n} \right)^{\alpha} = \left(\frac{1+\frac{1}{\gamma N}}{1+\frac{n}{\gamma N}} \right)^{\alpha}$$

$$= \left(\frac{\gamma + \frac{1}{N}}{\gamma + f} \right)^{\alpha} \approx \left(\frac{\gamma}{\gamma + f} \right)^{\alpha}.$$

Nous obtenons ainsi l'expression suivante pour γ,

$$\gamma = \frac{f Q_n^{1/\alpha}}{1 - Q_n^{1/\alpha}}.$$

Si nous remplaçons γ par cette expression dans la définition du modèle, nous obtenons

$$P = \left(\frac{f + \frac{f Q_n^{1/\alpha}}{1 - Q_n^{1/\alpha}}}{1 + \frac{f Q_n^{1/\alpha}}{1 - Q_n^{1/\alpha}}} \right)^{\alpha} = \left(\frac{f}{1 - Q_n^{1/\alpha}\,(1-f)} \right)^{\alpha}.$$

Pour ce qui est de la borne supérieure, il faut noter premièrement que la fonction, pour x et y entre 0 et 1,

$$F(x,y) = \left(\frac{f}{1 - x^{1/y}\,(1-f)} \right)^{y}$$

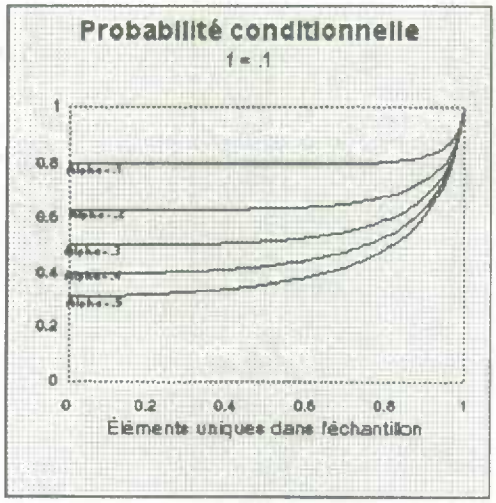
est croissante pour x (y fixé) et décroissante pour y (x fixé). Ainsi nous avons

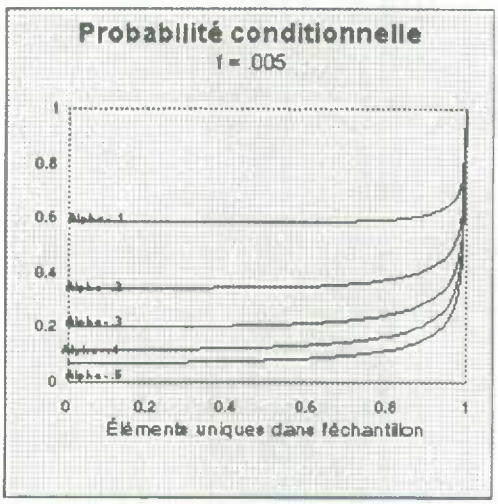
$$P = \left(\frac{f}{1 - Q_n^{1/\alpha}\,(1-f)} \right)^{\alpha} \le \left(\frac{f}{1 - Q_n^{1/\hat{\alpha}}\,(1-f)} \right)^{\hat{\alpha}}$$

$$\le \left(\frac{f}{1 - \left(\frac{1 - \frac{f}{(P^*)^{1/\hat{\alpha}}}}{1-f} \right)(1-f)} \right)^{\hat{\alpha}} = P^*.$$

La première inégalité vient de la sous-estimation supposée de α, la deuxième vient de l'application de la borne pour Q_n.

Les graphiques suivants donnent la relation entre P et Q_n pour différents contenue et fractions de sondage.





Ces graphiques nous montrent clairement que nous pouvons tolérer un grand nombre d'éléments uniques dans l'échantillon pour de petites fractions de sondage. Que faut-il faire, cependant, si la proportion d'éléments uniques dans l'échantillon est trop élevée ? La section suivante tentera de répondre à cette question.

6. TRAITEMENTS DANS LES DONNÉES

La section précédente donnait des moyens pour évaluer le risque de divulgation. Encore une fois, les conditions suffisantes d'avoir un risque peu élevé sont :

I. Une petite fraction de sondage :
II. Une proportion d'éléments uniques dans l'échantillon pas trop élevé :
III. Une estimation de α pas trop petite.

Dès qu'une de ces conditions n'est pas respectée, ou bien il faut échantillonner de nouveau pour réduire la fraction de sondage ou il faut modifier le contenu. Toute modification de contenu sera appelée un traitement dans les données. On peut effectuer soit un traitement global ou un traitement locale. Un traitement global est appliqué à tous les enregistrements, comme, par exemple, un regroupement de valeurs d'une des variables d'appariement. Un traitement local, par opposition, n'est appliqué qu'à une partie des enregistrements. Il y a plusieurs méthodes de traitement globales ou locales. Elles ont toutes leurs points forts ou faibles. L'objectif de cette section n'est pas d'expliciter la liste des méthodes avec leur performance. Nous aimerions plutôt répondre à la question suivante : lorsqu'on privilégie un traitement local, quels sont les enregistrements qui devraient être traités pour optimiser le traitement ?

En théorie, le but du traitement est de réduire le nombre d'éléments uniques dans la population qui se trouvent dans l'échantillon. Donc si nous voulons optimiser, nous devons trouver un moyen d'identifier ces enregistrements et deuxièmement d'appliquer un traitement qui ne les rend plus uniques dans la population. Pour ce qui est du traitement, nous pouvons supposer que les changements nécessaires de certaines valeurs des variables d'appariement sont minimes. Nous supposerons donc que dès qu'un traitement est appliqué à un enregistrement, ce dernier sera considéré comme sécuritaire. Reste donc la question du choix des enregistrements à traiter. Puisque les éléments uniques dans la population sont nécessairement uniques dans l'échantillon, nous devons nous concentrer premièrement seulement sur les uniques dans l'échantillon. Mais cela ne suffit pas. Il faut être en mesure de pouvoir filtrer les éléments uniques dans la population de ceux qui ne sont uniques que dans l'échantillon. C'est ici que le concept de la multiplicité d'un enregistrement s'insère dans la pratique.

Comment peut-on faire pour filtrer les uniques dans la population des autres ? Il faut arriver à pouvoir évaluer le «degré d'unicité» des enregistrements. À pouvoir dire qu'un enregistrement est plus unique qu'un autre. Comment faire ? Nous allons énoncer un postulat et nous verrons vers où il nous amène.

Postulat G : *La plupart des éléments uniques dans la population sont également uniques dans la population pour un sous-ensemble restreint de variables d'appariement.*

Ce postulat dit que l'attribut d'unicité dans la population dépend surtout d'une combinaison très rare de valeurs d'un petit nombre de variables d'appariement. Cela dit, si nous recherchons les uniques dans la population avec, par exemple, seulement trois variables

d'appariement, peut-être certains éléments seront déjà classés comme uniques. En cherchant les uniques pour toutes les combinaisons de trois variables parmi le nombre de variables d'appariement et en additionnant, pour chaque élément, le nombre de fois que ce dernier est unique, on en vient à une notion quantitative d'unicité. Le nombre de fois qu'un élément est unique dans un tableau à trois dimensions est appelé la multiplicité de cet élément. Nous pouvons dire que plus un élément a une multiplicité élevé, plus cet élément a un risque d'identification élevé. Que se passe-t-il lorsque nous n'avons qu'un échantillon ? Nous avons constaté que si nous calculons la multiplicité seulement avec l'échantillon, elle définit une partition de l'échantillon dont les différentes parties ont des proportions d'éléments uniques dans la population très différentes. Nous avons simulé un petit exemple pour montrer l'utilité de la multiplicité.

Nous avons pris un échantillon aléatoire simple avec une fraction de sondage de 0,009 d'une population de taille 781 825 éléments. Ce qui donne une taille échantillonnale de 7 037. Le fichier contient cinq variables d'appariement. Le nombre d'éléments uniques dans la population est 35 718 (4,5%). Le nombre d'éléments uniques dans l'échantillon s'élève à 2 301 (32,7%). Le nombre d'éléments dangereux (uniques dans la population qui se trouvent dans l'échantillon) s'élève à 321 (4,5%). La probabilité conditionnelle s'établit à 14%. Si nous choisissons au hasard parmi les éléments uniques dans l'échantillon, seulement 14% de ces enregistrements (en moyenne) sont dangereux. Beaucoup de traitement est par conséquent appliqué à des enregistrements qui n'en ont pas besoin. Si nous calculons la multiplicité des enregistrements, nous obtenons le tableau suivant.

Nous pouvons voir aisément que la partition générée par la multiplicité nous aide grandement à choisir les enregistrements à traiter. Par exemple, si nous décidons de traiter tous les enregistrements ayant une multiplicité supérieure à trois, nous éliminons 83,4% (268 éléments) des enregistrements dangereux en ne traitant que 10,3% des enregistrements, ce qui est plus performant que d'y aller au hasard. Nous avons essayé cette technique avec des fichiers de dix ou quinze variables d'appariement et, bien que le filtre n'était pas aussi performant que celui présenté ci-dessus, les résultats sont qu'en même surprenants. La recherche se poursuit maintenant vers une détermination de la multiplicité minimale où un traitement serait nécessaire. Cette multiplicité, appelée «le seuil de singularité» indiquerait, si le pourcentage de traitement est trop élevé, qu'il faut envisager plus de mesures globales.

Résultats de la simulation

Multiplicité	# éléments	# uniques	%
10	18	15	83,3
9	41	23	56,1
8	64	33	51,6
7	45	26	57,8
6	191	61	31,9
5	220	77	35,0
4	140	33	23,5
3	388	32	8,2
2	294	17	5,8
1	472	3	0,6
0	5 164	1	0,0
Total	7 037	321	4,5

7. CONCLUSION

Nous vous avons donné dans cet article l'avancement de la recherche à Statistique Canada sur l'évaluation du risque de divulgation des fichiers de microdonnées à variables discrètes. Pour que le modèle énoncé dans cet exposé puisse ressortir, il faut qu'il puisse être justifié théoriquement. Les recherches amorcées par Skinner sur le modèle Poisson-Lognormal sont encourageantes. Si ces modèles collent à la réalité, ils doivent converger quelque part. Aussi, la multiplicité des enregistrements est un concept qui selon nous est essentiel pour un traitement efficace du fichier de microdonnées.

BIBLIOGRAPHIE

Greenberg, B. V., et Zayatz, L. (1992). Strategies for measuring risk in public use microdata files, Statistica Neerlandica, 1992.

Bethleem, J. G., Keller, W. J., et Pannekoek, J. (1990). Disclosure control of microdata, *JASA*, 85, 38-45.

Skinner, C. J., et Holmes, D. J. (1992). Modelling population uniqueness, Paper presented at International Seminar on Statistical Confidentiality, Dublin.

Goodman, L. A. (1949). On the estimation of the number of classes in a population, *AMS*, 20, 572-579.

Mercier, A. (1984). Quelques identités de l'analyse combinatoire, *Discrete Mathematics*, North Holland, 49, 139- 149.

SESSION 6

Rendre les données accessibles au grand public

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

DIFFUSION DE L'INFORMATION EN VUE DES ÉTUDES DE MARCHÉ ET DE L'ANALYSE SPATIALE

C. Sewards et L. Li[1]

RÉSUMÉ

Le prototype de la version électronique du Recueil statistique des études de marché assure un accès convivial à la vaste gamme de données intégrées qu'on retrouve dans la publication du même nom de Statistique Canada, ce qui facilite l'analyse des produits et des marchés assistée par ordinateur. Le prototype consiste en un système intégré exploitant les avantages de trois sortes de logiciel (hypertexte, chiffrier et cartographie). Il est entièrement compatible avec les logiciels de traitement de texte et les chiffriers les plus populaires. Le produit consiste en un manuel électronique pourvu d'une table des matières consultable, associée aux tableaux secondaires du chiffrier et aux cartes de la visionneuse géographique. La série de logiciels permet à l'utilisateur d'effectuer des analyses statistiques et d'afficher les données sous forme de graphiques ou de cartes. Une présentation intégrée de ce genre pourrait devenir un moyen de diffusion peu coûteux pour un très large éventail de données.

MOTS CLÉS: Études de marché; système d'information géographique; analyse spatiale; intégration des données.

1. INTRODUCTION

Le Recueil statistique des études de marché (RSEM) est l'une des publications vedette de Statistique Canada. On y trouve des données très agrégées sur maints sujets, des habitudes d'achat des consommateurs au commerce international, en passant par le portrait socioéconomique de diverses villes canadiennes, et ainsi de suite. L'utilisateur peut ainsi se faire une idée de la situation commerciale générale sur le marché canadien. Par ailleurs, le Recueil vient en aide aux analystes en leur permettant d'obtenir les données qui appuieront leurs études de marché puisqu'on y retrouve des renseignements sur presque tous les domaines auxquels s'intéresse Statistique Canada.

On a récemment entrepris une étude-pilote pour examiner la conversion éventuelle du RSEM au format électronique. Les résultats ont fait ressortir les possibilités d'un moyen de diffusion électronique qui tirerait parti des points forts d'un logiciel hypertexte, d'un chiffrier électronique et d'un système d'information géographique (SIG), et qui proposerait à l'utilisateur un instrument souple mais puissant, en mesure de répondre à bon nombre de ses exigences commerciales. Parallèlement, cette technologie habilitante ramène la question de l'harmonisation des données au premier plan.

Le présent article donne les résultats de l'étude-pilote. On débute par un examen des besoins de l'utilisateur, identifiés par la consultation de la clientèle et les commentaires des groupes de réflexion. On passe ensuite aux questions techniques et aux problèmes d'intégration des données qu'a posés le développement du prototype. Suit une description détaillée de la solution retenue, qui combine logiciel hypertexte, chiffrier et système d'information géographique (SIG) pour fournir un jeu d'ensembles de données intégrées touchant de nombreux domaines examinés à Statistique Canada. Les auteurs terminent par quelques réflexions sur les possibilités que réserve une telle approche pour l'avenir.

2. BESOINS DE L'UTILISATEUR

Le RSEM est destiné à une clientèle très diversifiée. Après avoir pris connaissance des commentaires des clients et avoir établi l'orientation qu'on souhaite donner au produit, aux fins de l'étude, on a circonscrit les principales applications de ce dernier, qui aux fins de l'étude, correspondent aux suivantes:

[1] Crystal Sewards et Larry Li, Division de la géographie, Statistique Canada, Ottawa (Ontario), K1A 0T6.

- examiner la situation sur le marché;
- évaluer le potentiel du marché;
- chercher des clients potentiels;
- analyser les tendances générales du marché.

Les commentaires des groupes de réflexion et des clients ont démontré l'utilité du RSEM. Toutefois, avec plus de 600 pages sous sa forme actuelle, le format imprimé nuit à l'utilisation du produit en raison de son manque de souplesse. Les clients ont souligné qu'il était difficile de repérer les données voulues, et que les informations relatives à ces données n'étaient pas suffisamment détaillées. Les clients préféreraient une version électronique intégrant certaines fonctions pour la manipulation et l'analyse des données. Ils ont entre autres demandé des données plus nombreuses (plus de données longitudinales, ventilation géographique plus poussée et catégorisation accrue des produits), des données plus souples et plus détaillées, des notes plus abondantes sur les sources, de meilleures références aux sources complémentaires de données, la possibilité de recouper les totalisations et de produire des résultats analytiques, des graphiques, des tableaux et des cartes, et la capacité d'exporter les données vers un chiffrier ou un autre logiciel en vue de leur inclusion dans un rapport. Les clients souhaitaient un produit de première qualité en mesure de satisfaire leurs besoins et étaient prêts à payer plus cher pour l'obtenir.

Les souhaits précités ont déterminé les spécifications fondamentales du prototype. L'équipe chargée de la gestion du projet voulait aussi préserver des liens étroits entre la publication existante et sa version électronique. Les membres de l'équipe estimaient qu'ils pourraient ainsi adopter une approche autorisant le perfectionnement graduel du produit électronique, sachant qu'il est improbable que la version initiale inclue toutes les données réclamées par les utilisateurs. En outre, pour mieux aiguiller le développement du produit, on s'est fixé pour cible un prix de 500$. Pareil prix situerait le produit dans la gamme des produits analogues présentement offerts sur le marché. Sur cette somme, on a réservé 100$ pour le logiciel.

3. À LA RECHERCHE D'UNE SOLUTION

Les travaux ont débuté par l'examen des logiciels susceptibles de conférer au produit les qualités désirées. Quatre types de logiciels ont été envisagés: hypertexte, chiffrier, base de données et SIG de bas niveau.

Le logiciel hypertexte, comme Folio Views, présente d'excellentes possibilités pour ce qui est de structurer l'information sous forme de livre électronique. Les logiciels de ce genre acceptent les entrées des logiciels de traitement de texte courants, pour créer une base de données narratives à partir de laquelle on peut entreprendre des recherches au moyen de mots clés. Bon nombre de ces logiciels acceptent l'intégration de liens qui permettent à l'utilisateur de sauter rapidement d'une section à l'autre. On procède de la même façon pour associer des renvois et des notes explicatives à certains mots ou syntagmes. Malheureusement, les logiciels hypertexte s'avèrent moins efficaces avec les tableaux de données. En règle générale, ils sont aussi dépourvus des fonctions servant à la production de cartes et de graphiques ou à la manipulation statistique des données.

Les tableurs sont extrêmement bien adaptés au traitement des tableaux numériques. Bon nombre d'entre eux incluent de puissantes fonctions mathématiques et statistiques. On peut créer des graphiques et des diagrammes au simple déclic d'un bouton. Certains logiciels populaires autorisent aussi désormais la production de cartes rudimentaires. Quelques logiciels spécialisés, notamment Ivision, peuvent traiter des tableaux très volumineux. Ils réorganisent la position des variables dans un tableau; ils déplacent l'axe des abscisses et celui des ordonnées d'un tableau et modifient très efficacement la «dimension» illustrée à l'écran. Par contre, ils tolèrent moins les données en format «texte» que les logiciels hypertexte, sont moins efficaces dans les recherches de type narratif et associent plus difficilement les explications à des mots clés ou à des syntagmes.

Les logiciels de base de données comme dBASE ou FoxPro gèrent des données numériques et des textes courts avec aisance. Des opérateurs puissants leur permettent d'indexer et de chercher l'information, surtout à l'intérieur de champs donnés. Ces logiciels peuvent exécuter des calculs pour combiner les données de champs différents et forger des liens entre les tableaux. Toutefois, ils ne sont pas expressément conçus pour le traitement de texte ou la cartographie, en général.

Les systèmes d'information géographique (SIG) regroupent habituellement une base de données et un logiciel de cartographie. Le premier permet le traitement des attributs, comme le font les logiciels de base de données. À cela s'ajoutent cependant d'autres fonctions, par exemple la production de cartes, des recherches spatiales et l'intégration de données spatiales par superposition de différentes cartes ou couches de données. À l'instar des logiciels de base de données, les SIG sont plus gauches avec les descriptions narratives.

En outre, ils manquent fréquemment d'opérateurs statistiques.

Après ce tour d'horizon, on s'est rendu compte qu'aucun logiciel actuellement offert sur le marché ne répondrait à toutes les exigences de l'utilisateur sans une programmation importante. Par bonheur, l'arrivée récente d'ARCView-1 dans le monde des «gratuiciels» (quoique toujours sous licence) et les conditions intéressantes associées à la distribution de masse, sous licence, des logiciels Folio Views et Ivision nous ont permis d'envisager la combinaison de ces trois logiciels dans la version électronique du RSEM, dans les limites du budget établi. On a donc pris la décision d'articuler le prototype sur les trois logiciels intégrés. De cette façon, il sera possible d'exploiter les points forts de chacun pour satisfaire aux diverses exigences de la liste de spécifications.

4. LE FORMAT DE PUBLICATION INTÉGRÉ

Le format de publication intégré ressemble beaucoup à un livre électronique. Le logiciel hypertexte a donné au prototype la structure intuitive d'un livre, mais ses instruments de navigation perfectionnés facilitent la recherche de l'information et font surgir des notes qui éclairent le contexte. Pour permettre le traitement des tableaux et la manipulation des données, on a exporté les tableaux dans le tableur retenu, soit Ivision. Les utilisateurs voulaient aussi être en mesure de produire des cartes et d'effectuer une analyse spatiale de l'information sur les marchés. Une présentation spatiale des données figurait au premier rang des caractéristiques les plus utiles pour le traitement de l'information au niveau infra-provincial, où interviendront bon nombre de cellules de données et où on s'intéressera à l'emplacement des clients ou des concurrents. Cette exigence a été satisfaite grâce au système d'information géographique (SIG) ARCView-1, qui a été intégré à l'environnement de la publication.

Bien que les utilisateurs s'intéressent au coût et aux particularités techniques de la structure de la publication électronique, la convivialité et l'efficacité de cette dernière revêtent encore plus d'importance pour eux. Dans le prototype, l'utilisateur accède au RSEM par Folio Views. Le premier écran qu'il aperçoit est la table des matières, qui dresse la liste des sujets de tous les tableaux faisant partie de la base de données et de la base d'information du logiciel Folio.

Folio VIEWS - Table des matières - RSEM

Faillites d'entreprises, par province, 1991 et 1992
Revenu total, secteur des services aux entreprises, par province, 1988 et 1989
\+ Structure des dépenses, familles et personnes seules, 1986 - 1991
\+ Nombre de magasins de détail appartenant à une chaîne, et ventes, par branche d'activité, 1989 et 1990
Ventes au détail estimatives, pour certaines catégories, par province, 1989

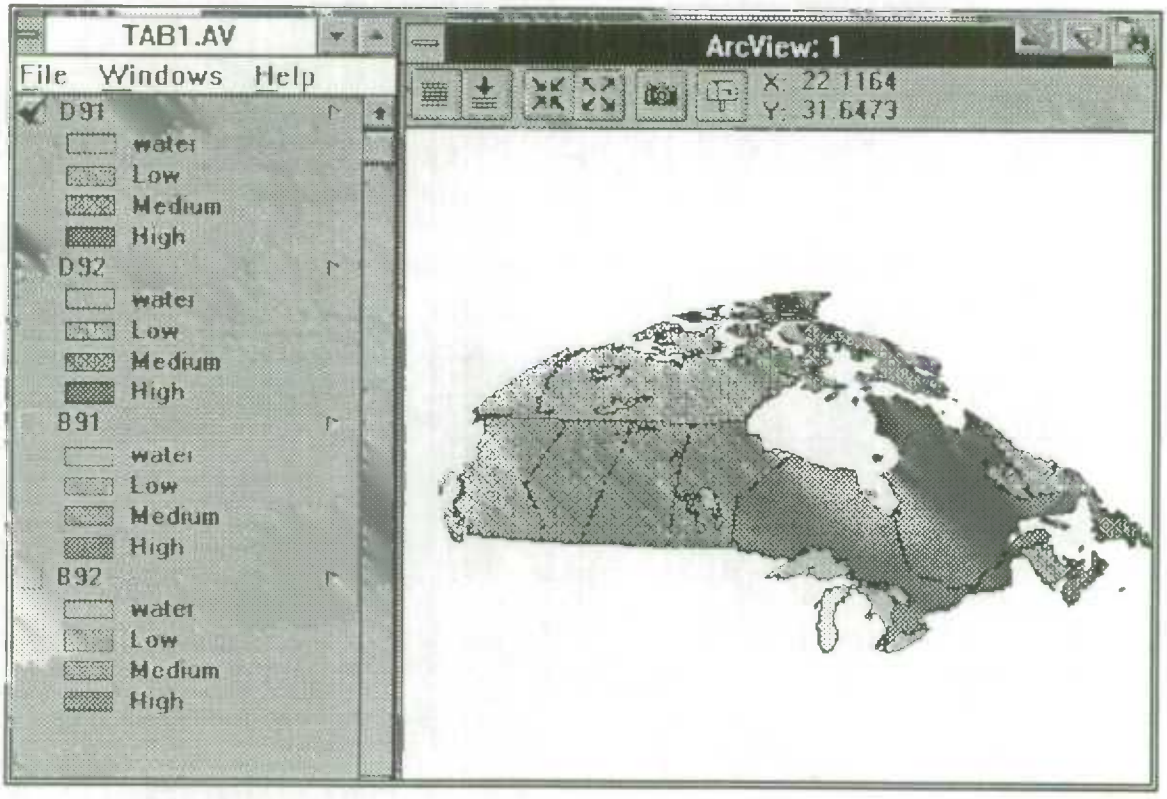
Folio VIEWS - Table des matières - RSEM

Faillites d'entreprises, par province, 1991 et 1992
Revenu total, secteur des services aux entreprises, par province, 1988 et 1989
\- Structure des dépenses, familles et personnes seules, 1986 - 1991
Calgary (Alberta)
St. John's (Terre-Neuve)
\- Nombre de magasins de détail appartenant à une chaîne, et ventes, par branche d'activité, 1989 et 1990
Calgary (Alberta)
St. John's (Terre-Neuve)
Ventes au détail estimatives, pour certaines catégories, par province, 1989

C:\IBUILDER\DATA\TABLE1.IVT

Year: 1991

Location	Units	Cases	Deficiencies
Canada		13496	3856859
Newfoundland		149	36232
Pr. Ed. Is.		23	17202
Nova Scotia		662	59703
New Brunswic		227	39371
Quebec		5217	1708718
Ontario		3629	1367263
Manitoba		410	83571
Saskatchewar		583	66734
Alberta		1304	231352
Br. Colum.		1288	246377
Yk./NWT		4	334



La table des matières peut être déployée ou comprimée, si bien que l'utilisateur peut voir autant ou aussi peu d'information qu'il le désire. Le signe plus (+) à côté d'une entrée indique l'existence de niveaux secondaires. Le signe moins (-) signale qu'il est impossible de développer l'entrée davantage, mais qu'on peut la réduire. Pour accéder à un tableau de la liste, il suffit de cliquer deux fois sur le titre. Chaque entrée de la table des matières est automatiquement liée à la partie correspondante du document. Sous le titre de chaque tableau apparaissent deux icônes: celle d'Ivision et celle d'ArcView-1. En cliquant deux fois sur l'une ou sur l'autre, on lance le logiciel et obtient un tableau de données, avec Ivison, ou une carte, avec ArcView-1. Lorsqu'il ferme Ivision ou ArcView-1 après avoir examiné le tableau ou la carte, l'utilisateur retourne automatiquement au titre du tableau dans Folio Views. Pour examiner les sources et les renvois d'un tableau dans Folio Views, il suffit de cliquer deux fois sur les notes qui se déroulent sous le titre du tableau.

Après avoir abandonné le titre du tableau dans Folio Views pour le tableau proprement dit dans Ivision, l'utilisateur peut trier les données relatives à une variable déterminée par ordre croissant ou décroissant, modifier la position d'une dimension quelconque, emboîter les dimensions, voire calculer la variation d'une année à l'autre en pour cent, en quelques déclics du bouton de la souris. Les données peuvent ensuite être transformées rapidement en graphique qui fera ressortir les changements et facilitera une analyse comparative. L'utilisateur a aussi la possibilité de convertir les données en cartes, en cliquant sur l'icône ARCView-1 s'il souhaite dégager les tendances spatiales.

La fonction recherche de Folio Views permet à l'utilisateur de repérer un mot clé précis, «faillites» ou «Calgary» par exemple, dans la base de données du logiciel. Le nombre de fois où ce mot apparaît dans la base de données s'inscrit dans la fenêtre et il suffit de cliquer sur le bouton «*apply to all*» pour se porter instantanément au premier endroit où se présente le mot dans la base de données. Le bouton «*next*» amène l'utilisateur à l'inscription suivante.

ARCView-1 gère les fonctions de cartographie et d'analyse spatiale. Le programme présente les données d'un tableau quelconque sous forme de carte. L'utilisateur peut effectuer un panoramique, un gros plan ou un plan éloigné et examiner le tableau de données correspondant à la carte. S'il le désire, l'utilisateur chevronné pourra aussi sauter directement dans ARCView-1 et recourir à des fonctions plus avancées puisque la version intégrale du logiciel fait partie du produit.

5. ORGANISATION DES DONNÉES

Pour préserver un lien solide entre la publication ordinaire et sa version électronique, on a reproduit la structure et l'organisation fondamentale des tableaux du RSEM dans le prototype. Sept tableaux ont été extraits de la publication existante et intégrés au produit électronique afin d'en illustrer la fonctionnalité. L'utilisateur a donc eu l'occasion de vérifier la valeur du concept par des recherches réalistes portant sur les diverses questions fondamentales qu'on se pose lors d'une étude de marché: «Qui sont mes clients?, Où sont mes clients?, Comment puis-je en trouver d'autres?».

Le prototype a aussi mis en évidence qu'on peut intégrer des données plus nombreuses et plus détaillées au produit. On l'a démontré en incluant les données chronologiques d'un plus grand nombre d'années au chiffrier Ivision. Pour accroître la précision géographique, on a annexé des cartes aux variables des données provinciales, ce qui offre un aperçu infra-provincial du sujet. L'utilisateur pouvait accéder aux informations correspondantes de la base de données à partir de la carte infra-provinciale.

Au fil des perfectionnements apportés au prototype, il est devenu manifeste que la structure existante empêcherait l'utilisateur d'exploiter pleinement certaines capacités offertes par les logiciels quant à la recherche, à l'intégration et au classement des données. Le stockage des données dans un si grand nombre de petits tableaux, comme c'est le cas dans la publication actuelle, constitue l'un des principaux obstacles. Une structure aussi compartimentée freine la recherche puisque cette dernière ne s'effectue qu'à l'intérieur d'une base de données déterminée. La manipulation des données d'un tableau à l'autre soulève aussi des difficultés, l'utilisateur devant ouvrir différents tableaux et en découper des fragments pour procéder à l'analyse. Peut-être devra-t-on donc combiner les tableaux afin d'élargir la base de données.

Bien que la portée de l'étude nous ait empêché d'approfondir ces questions, les résultats préliminaires révèlent que l'intégration de données beaucoup plus nombreuses en une seule base pourrait aider l'utilisateur à effectuer rapidement un nouveau tri et à regrouper toutes les cellules de données pour un lieu ou une période précis dans le même tableau, ce qui faciliterait d'autant l'analyse des tendances et des conditions du marché. Il semble donc qu'on pourrait mieux répondre à diverses exigences de l'utilisateur; néanmoins, on n'y arrivera pas sans difficulté. Parmi les défis et les problèmes qui se sont manifestés lorsqu'on a tenté de restructurer les données, il convient de mentionner la

disponibilité et la compatibilité de ces dernières. Par exemple, pourrait-on intégrer un tableau sur la population par province en 1986 et 1991 à un tableau sur les dépenses des ménages par province en 1990? Pourrait-on combiner un tableau des mises en chantier de constructions résidentielles par province en 1992 à un tableau du revenu des ménages par région métropolitaine de recensement en 1992? Devrait-on grouper les données selon des paramètres géographiques, par domaine, selon l'année ou au moyen d'une autre variable? Certaines données relatives à une région particulière ont été recueillies à des années différentes; comment pourrait-on intégrer les tableaux correspondans ou recouper les totalisations? Les utilisateurs avaient aussi proposé que certaines données soient offertes à un niveau géographique plus détaillé. Certaines sont disponibles pour des régions plus vastes comme la province ou la région métropolitaine de recensement (RMR), mais pas au niveau de la subdivision de recensement (SDR). Serait-il acceptable de fournir des données différentes ou moins exhaustives aux niveaux inférieurs de groupement géographique. Quelles autres couches de données trouve-t-on déjà dans le RSEM pour les mêmes régions? La présentation des données soulève aussi des difficultés; une version électronique du RSEM permettrait peut-être de ne diffuser que les données d'une province ou d'une RMR. Ces questions et beaucoup d'autres devront être soigneusement examinées dans l'avenir.

6. CONCLUSIONS

Les résultats ont dévoilé le potentiel d'un instrument de diffusion électronique qui exploiterait les points forts d'un logiciel hypertexte, d'un tableur et d'un système d'information géographique pour proposer à l'utilisateur un produit efficace et adaptable qui répondra à bon nombre de ses exigences professionnelles. Outre sa convivialité, pareil produit fournirait beaucoup plus de données et donnerait à l'utilisateur la possibilité de les transformer en fonction de ses besoins d'analyse, tout cela à un prix attrayant.

Parallèlement, la technologie habilitante a fait ressortir les différences conceptuelles qui existent dans la définition des variables de divers ensembles de données employés à Statistique Canada. Ainsi, on ne définit pas «famille» de la même façon d'un tableau à l'autre. Dans l'Enquête sur les dépenses des familles, ce terme désigne un groupe de personnes occupant le même logement et effectuant ses principaux achats à partir d'un revenu commun ou combiné. Les tableaux reprenant les données du Recensement, par contre, ont pour définition de «famille» un couple marié ou en union libre avec ou sans enfant jamais marié ou un parent seul avec au moins un enfant jamais marié occupant le même logement.

Les efforts déployés pour créer une version électronique intégrée du RSEM ont aussi fait surgir les problèmes d'harmonisation des données, ainsi que de rationalisation des prix et des produits. On tente déjà d'approfondir différents aspects de ces difficultés, ce qui devrait en fin de compte aider Statistique Canada à mieux répondre aux besoins des utilisateurs.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

GAGNER LA CONFIANCE DES JOURNALISTES : *LE QUOTIDIEN* DE STATISTIQUE CANADA ET LES MÉDIAS

W.R. Smith[1]

RÉSUMÉ

Statistique Canada a réalisé un projet de deux ans visant à améliorer ses communications avec le grand public par l'intermédiaire des médias. La réalisation du principal objectif de ce projet — utiliser les médias pour diffuser des informations statistiques — exigeait un changement des mentalités au sein du Bureau et l'appui des responsables au plus haut niveau. Grâce à des programmes officiels ou non et en procédant parfois également par tâtonnements, on a modifié de fond en comble les méthodes de rédaction des communiqués en s'attachant surtout à décrire les tendances et à présenter des analyses dans un langage accessible, avec des tableaux et des graphiques clairs et faciles à comprendre. Les démarches utilisées incluent les délibérations d'un comité supérieur de rédaction, la publication de lignes directrices, l'appel à des services de consultation et l'organisation de cours de rédaction axés sur les médias. Les succès remportés sont difficiles à mesurer, mais il semble qu'on ait réussi jusqu'à maintenant à accroître la portée et la qualité de la couverture médiatique tout en bénéficiant de retombées telles qu'une attitude plus positive et plus enthousiaste dans les rapports avec les médias.

MOTS CLÉS : Médias; consultations; analyse; comité de rédaction; communications; lignes directrices.

1. INTRODUCTION

1.1Servir le public

Au cours des deux dernières années, Statistique Canada a consacré une grande part de ses énergies à l'amélioration des communications avec le grand public par l'intermédiaire des médias. On est en effet convaincu que le grand public constitue une clientèle cible importante pour un bureau national de statistique.

Toutefois, pour communiquer efficacement avec le grand public, il est essentiel de faire appel aux médias. Les journalistes deviennent donc ainsi un chaînon intermédiaire incontournable entre le bureau de statistique et le public. Dans une société démocratique, les journalistes sont férocement indépendants: on ne peut ni les acheter, ni leur donner des ordres, ni les coopter. Il faut plutôt s'employer à gagner leur soutien et leur confiance pour en faire des agents de diffusion des informations du bureau de statistique.

L'organe principal de communication de Statistique Canada avec les médias est son bulletin officiel: *Le Quotidien*. Lorsque de nouvelles données sont sur le point d'être diffusées, l'organisme a pour politique d'en faire l'annonce au préalable dans *Le Quotidien*. Publié chaque jour ouvrable, *Le Quotidien* constitue donc un outil idéal pour améliorer nos communications avec le grand public par l'intermédiaire des médias.

Dans le présent article, nous décrivons le contexte et les mentalités qui caractérisaient Statistique Canada lorsque notre projet a été élaboré. Nous décrivons les objectifs fixés et les méthodes utilisées pour réaliser ce projet. Finalement, nous exposons brièvement les résultats obtenus jusqu'à maintenant.

1.2 Les bureaux de statistique et le grand public

Si Statistique Canada souhaite communiquer plus efficacement par l'intermédiaire des médias, c'est qu'il est convaincu que le grand public est une des clientèles importantes d'un bureau de statistique. Dans une société à démocratie libérale, les membres du public, c'est-à-dire les citoyens ordinaires et les acteurs du secteur économique, ont besoin d'informations sur la population, la société, l'économie et la culture de leur pays. Ces informations les aident à faire leur travail, à élever leur famille, à faire des achats, à juger leurs gouvernants, à voter et à faire chaque jour une foule de choix à incidences économiques. Une population bien informée améliore à la fois l'efficacité politique et économique

[1] Wayne R. Smith, Division des communications, Statistique Canada, Ottawa (Ontario) Canada K1A 0T6.

d'un pays.

Cependant, les conséquences d'un vote individuel ou d'une décision économique personnelle ne sauraient justifier que l'on procède chaque fois à une vaste recherche d'informations. Rares sont ceux qui visiteront le bureau de statistique pour dresser un bilan économique national avant d'aller voter. De toute manière, les bureaux de statistique pourraient à peine répondre à la demande. Par ailleurs, ces derniers n'ont pas les moyens de communiquer directement chaque jour avec tous les citoyens.

Les médias offrent donc aux bureaux de statistique un moyen unique de s'acquitter d'un élément critique de leur mandat. Les citoyens s'informent en lisant des articles de journaux et en parcourant les manchettes; ils sont nombreux à écouter chaque jour une ou plusieurs émissions de télévision ou de radio qui influent aussi sur leurs opinions. L'utilisation que les bureaux de statistique pourront faire des médias pour communiquer efficacement aura une incidence énorme sur leur aptitude à informer le grand public.

Outre l'avantage direct que présentait à ses yeux une population mieux informée, Statistique Canada entrevoyait un certain nombre d'avantages connexes. Grâce à un programme de communications publiques mieux ciblé, la population deviendrait plus consciente de l'existence du bureau de statistique ainsi que de l'importance et de la pertinence de ses programmes. Les répondants verraient plus clairement les avantages de collaborer aux sondages effectués par l'organisme, et les clients principaux seraient mieux renseignés sur les informations disponibles. Bref, Statistique Canada anticipait une augmentation de l'appui du public, une amélioration de ses rapports avec les répondants et certains avantages commerciaux.

1.3 Le chaînon essentiel

Au Canada, comme dans la plupart des pays à démocratie libérale, les médias défendent férocement leur autonomie de rédaction. Pour être publiées, des informations doivent d'abord mériter d'être signalées, ou mériter plus que d'autres informations concurrentes d'être signalées. Les journalistes transmettent les informations dans leurs propres mots et non dans ceux de la source, et la qualité de la couverture est un reflet de leur aptitude à comprendre l'information. Les journalistes constituent un chaînon incontournable dans la transmission des informations entre les bureaux de statistique et le grand public.

Le mérite que présente une information d'être signalée est relatif. L'espace disponible — celui qui n'est pas déjà occupé par la publicité — est limité et a eu tendance à diminuer au cours des récentes années. Par ailleurs, le nombre de groupes d'intérêts et d'organisations qui cherchent à attirer l'attention du public sur leurs préoccupations et sur leurs actions par l'intermédiaire des médias ne cesse d'augmenter. Même si les journalistes, au Canada à tout le moins, reconnaissent que Statistique Canada constitue une source importante d'informations, la couverture qu'ils lui accorderont dépendra largement de notre aptitude à les convaincre de la valeur de nos informations, ainsi que du volume et de la nature des informations concurrentes et de l'espace disponible. Bref, pour bénéficier d'une bonne couverture, il faut pouvoir démontrer la pertinence de l'information que l'on souhaite diffuser. En outre, le degré de couverture pourra varier en fonction du mérite attribué à chacun des éléments d'information.

Les journalistes savent que *Le Quotidien* offre un aperçu unique et exhaustif de nouvelles informations en provenance de Statistique Canada. Les médias surveillent donc de près cette publication. Les diverses agences de presse se livrent une concurrence féroce lorsque *Le Quotidien* est livré par messager à la Tribune de la presse, au centre-ville d'Ottawa, à l'heure de diffusion officielle de 8 h 30.

L'obtention d'une bonne couverture n'est que le premier des obstacles à franchir pour obtenir une communication efficace par l'intermédiaire des médias. Le miroir que constitue les médias peut parfois être déformant. La plupart des journalistes doivent respecter des délais extrêmement serrés, et produire trois à cinq articles par jour. Même si l'on admet qu'ils soient capables de faire une analyse indépendante des données brutes, le temps leur manquera. À l'exception des rédacteurs d'articles spécialisés, les journalistes n'auront donc pas tendance à faire une analyse poussée des informations fournies; ils sont en fait de moins en moins capables, aujourd'hui, d'effectuer une telle analyse. Par ailleurs, les budgets de rédaction réduits vont entraîner une réduction du nombre de journalistes assignés à un secteur particulier et qui sont donc en mesure d'acquérir une connaissance spécialisée. Aujourd'hui, la couverture est un reflet de la compréhension que peut avoir un journaliste de l'information fournie et si cette dernière n'est pas claire, l'article qui en découlera risquera fort de contenir des erreurs dont tous finiront par souffrir.

Ainsi, pour obtenir une couverture médiatique positive, exacte et informative dans la plupart des sociétés à démocratie libérale, il est essentiel d'obtenir la collaboration des journalistes. Nous devons démontrer *pourquoi* notre information est suffisamment importante pour justifier l'attention. Si nous voulons qu'elle soit présentée comme autre chose que des faits

divers, nous devons inciter les journalistes à lui porter l'attention voulue en fournissant des analyses en contexte et en montrant les tendances qui en dénotent la véritable importance. Nous devons également communiquer dans un langage simple si nous souhaitons que notre information soit correctement transmise au grand public.

Ces activités répondent à nos propres besoins ainsi qu'à ceux du grand public, mais non aux besoins des médias. Ces derniers bénéficient certainement de cette symbiose, mais les bureaux de statistique et le public en profitent encore plus.

1.4 Objectifs et contexte interne

La haute direction de Statistique Canada considérait depuis longtemps que l'organisme pouvait et devait mieux gérer la transmission de ses informations par l'intermédiaire des médias. Elle jugeait important d'améliorer la quantité et la qualité de la couverture médiatique de l'information diffusée par le Bureau. C'est ce qui l'a conduite à porter une attention toute spéciale à l'amélioration du bulletin *Le Quotidien* et, en particulier, à la section «Principaux communiqués».

Pour réaliser un tel objectif, on a jugé essentiel de modifier de fond en comble les méthodes de rédaction des communiqués publiés dans *Le Quotidien*. On a déterminé clairement que les médias devaient pour cela devenir la clientèle cible, et que les autres utilisateurs du *Quotidien*, y compris la clientèle générale, deviendraient secondaires.

Avec le temps, Statistique Canada a mis au point un mécanisme extrêmement ordonné de diffusion des informations aux médias. En vertu de la politique en vigueur, tous les nouveaux communiqués doivent passer par l'organe officiel de Statistique Canada: *Le Quotidien*. Les communiqués les plus importants sont largement étoffés et paraissent sous la rubrique «Principaux communiqués». Ceux qui sont moins importants font l'objet d'articles plus sommaires.

Presque 30 divisions participent à la rédaction des communiqués destinés au *Quotidien*, en plus d'une douzaine d'analystes et de gestionnaires. Les textes sont préparés par les diverses divisions spécialisées, le personnel des Communications étant chargé des révisions mineures, de la présentation, de la normalisation et du contrôle de la qualité. La mise en oeuvre de la nouvelle vision nécessitait un changement d'attitude de tous ces intervenants. Savoir ce qui devait être fait n'était pas suffisant en soi pour exécuter le travail.

1.5 Un nouveau modèle

Statistique Canada décidé d'améliorer ses communiqués en élargissant l'analyse pour montrer l'importance et la pertinence des informations et en décrire le contexte, en utilisant un langage moins technique ainsi que des graphiques et des tableaux. Il a également entrepris de montrer plus clairement les rapports existant entre les données d'un communiqué à l'autre. Si les journalistes pouvaient constater la valeur des informations et en comprendre la teneur, ils seraient alors en mesure d'en faire un écho exact au grand public.

Le personnel des Communications a consulté des journalistes, en portant une attention particulière aux grandes agences de presse et aux grandes chaînes de journaux, pour déterminer comment on pourrait rendre les communiqués de l'organisme plus utiles pour eux. Les journalistes nous ont assuré que Statistique Canada constitue pour eux une source majeure de nouvelles nationales. Ils ont cependant insisté sur la nécessité d'identifier plus clairement les informations les plus importantes dans les communiqués, et d'étayer ces informations avec des analyses et une description suffisante du contexte pour leur permettre de rédiger leurs propres articles. Ils ont réclamé des communiqués plus brefs, rédigés dans un langage plus clair, avec des textes et des tableaux simplifiés. En tenant compte de ces réactions, le personnel des Communications a établi et a diffusé l'ébauche d'un guide sur la rédaction efficace des communiqués.

Les rédacteurs des divisions spécialisées ont été invités à utiliser une trame plus explicite pour leurs articles et à mettre l'accent sur l'analyse et l'explication du contexte. Ils ont été priés d'adopter un style et une structure plus journalistiques, et à utiliser un langage, des graphiques et des tableaux plus simples. Il s'agissait là d'un changement révolutionnaire par rapport aux analyses «baromètre» qu'on affectionnait jusque là et aux énumérations de faits saillants disparates, présentées de la même façon d'un numéro à l'autre avec des tableaux et des graphiques standards.

Les analystes ont généralement accueilli avec enthousiasme les nouvelles règles, même si certains de ceux employés dans les divisions spécialisées et certains cadres intermédiaires se montraient peu enclins à adopter le nouveau modèle. La résistance la plus forte a été observée chez les cadres intermédiaires et les gestionnaires. Ce conservatisme prenait ses racines dans la crainte et le ressentiment parfois justifiés qu'on éprouvait à l'endroit des médias. On s'inquiétait en outre des réactions que pourraient susciter des analyses plus poussées au sein des ministères chargés de l'établissement de politiques et on se préoccupait de l'utilité des articles pour la clientèle générale de Statistique Canada par rapport à celle des médias. Du

point de vue pratique, les divisions s'inquiétaient équalment des besoins en ressources et en temps qu'entraînerait le nouveau modèle, et se montraient préoccupées par l'ampleur des révisions qui seraient confiées au personnel des Communications.

Les pratiques de diffusion historiques, figées dans des énoncés de politique passablement stricts, confortaient les éléments conservateurs. La réaction traditionnelle aux préoccupations profondément enracinées concernant la limite à ne pas transgresser entre les commentaires analytiques légitimes et les commentaires politiques inappropriés consistait à prendre carrément le parti de la prudence.

Il est très vite devenu clair qu'il ne suffirait pas de simplement annoncer la nouvelle approche pour obtenir un changement. On a donc élaboré un plan en plusieurs volets abordant les questions de la culture de l'organisation, des mesures d'encouragement et des techniques nécessaires pour faire évoluer les mentalités et mettre en oeuvre la nouvelle stratégie de communication.

2. DIFFUSION DU MESSAGE

2.1 Changement à plusieurs niveaux

Pour réaliser les changements radicaux nécessaires, il fallait intervenir à un certain nombre de niveaux. Une première tentative des directeurs des communications fondée uniquement sur les mérites de la nouvelle méthode et portant une attention particulière aux analystes responsables de la rédaction des communiqués n'a pas réussi à vaincre la résistance des gestionnaires responsables de la supervision du procédé. Il a donc fallu mettre en place un programme plus ambitieux.

Le premier défi a consisté à modifier la façon même d'envisager les communiqués au sein de l'organisme et à gagner l'appui et la collaboration des cadres. Pendant deux ans, on a profité de toutes les occasions qui se présentaient pour proposer le nouveau modèle et en expliquer les principes sous-jacents. Pour lancer le processus, le comité des politiques comprenant le statisticien en chef et les statisticiens en chef adjoints a discuté du plan. Les statisticiens en chef adjoints se sont assurés chacun de leur côté de la collaboration de leurs gestionnaires. Le statisticien en chef a lui-même fait la promotion du plan dans une note de service, lors de la présentation de son «bilan» annuel aux cadres de l'organisation et lors de son entrevue annuelle publiée dans le bulletin des employés, *SCAN*. Les cours de formation sur l'analyse incorporaient le message dans leur plan; en fait, le principal cours de Statistique Canada sur l'analyse et la présentation des données comprenait deux jours d'exposés sur les médias et sur la rédaction de communiqués. Des journalistes éminents ont été invités à venir décrire leurs conditions de travail et leurs exigences. Le message était omniprésent et montrait clairement que les changements envisagés pour *Le Quotidien* étaient devenus une priorité qui bénéficiait de l'appui de la haute direction.

2.2 Le comité supérieur de rédaction

Dans une grande organisation, de nombreuses priorités se disputent l'attention des cadres. On a donc mis sur pied pour *Le Quotidien* un comité supérieur de rédaction présidé par le statisticien en chef et composé des statisticiens en chef adjoints et d'analystes principaux pour faire en sorte que cette priorité devienne et demeure la première dans l'esprit des gestionnaires. Le comité se réunissait toutes les semaines pour étudier les communiqués du *Quotidien* et pour rédiger des lignes directrices sur la rédaction efficace de communiqués à l'intention des gestionnaires et des analystes. Tous les communiqués principaux étaient soumis à un membre du comité de rédaction pour subir un examen critique. Les gestionnaires et les analystes des divisions spécialisées responsables de la rédaction de communiqués étaient invités à participer aux discussions du comité de rédaction portant sur leurs communiqués respectifs. Lorsque les améliorations apportées étaient jugées suffisantes, les communiqués étaient exemptés d'examen. L'exemption du processus d'examen est ainsi devenue un objectif à atteindre. Il a fallu 12 mois avant que l'ensemble des communiqués principaux finissent par être exemptés d'examens.

La création du comité supérieur de rédaction a été la clé du succès. La grande influence et la participation du comité au niveau de l'ensemble de Statistique Canada ont contribué à donner au processus la crédibilité dont il avait besoin pour entraîner l'évolution voulue des mentalités. Les lignes directrices sur la rédaction efficace des communiqués, un sous-produit des travaux du comité, ont également été ainsi légitimées. Elles ont fourni et maintenu un haut niveau d'attention et de concentration jusqu'à l'obtention d'un résultat probant. Pour être exemptés du cycle d'examen, les gestionnaires n'avaient d'autre choix que de se plier au processus d'examen : une mesure d'encouragement extrêmement efficace. Les cadres des Communications évoluant aux niveaux inférieurs de l'organisation n'avaient pas la légitimité ni l'influence nécessaires pour réaliser un changement d'une telle ampleur.

2.3 Aide directe

Pendant que la campagne d'information et les activités du comité supérieur de rédaction créaient le climat, la légitimité et les conditions propices à l'amélioration des communiqués du *Quotidien*, les divisions spécialisées avaient besoin d'une aide directe pour faire la transition.

C'est le comité de rédaction lui-même qui, dans une large mesure, a fourni cette aide grâce à ses conseils détaillés et à ses suggestions explicites pour la refonte des communiqués. En outre, en élaborant et en publiant des lignes directrices générales sur la rédaction efficace des communiqués, le comité a aidé les divisions à faire la transition. Toutefois, on avait malgré tout besoin d'une aide supplémentaire à l'extérieur de l'atmosphère chargée des réunions du comité de rédaction. Cette aide pratique a été fournie de trois façons.

2.3.1 Groupes de travail

Lors des premières étapes du travail du comité, on a mis sur pied pour plusieurs publications des groupes de travail relativement officiels constitués d'analystes principaux, de membres du personnel des Communications et de membres des divisions spécialisées intéressées. Ces groupes de travail ont transmis leurs conclusions accompagnées de modèles révisés de communiqués au comité supérieur de rédaction. Cette méthode a aidé le comité à déterminer rapidement les principes généraux de la rédaction efficace des communiqués et à en entreprendre l'application.

2.3.2 Services de consultation

Le deuxième moyen de venir en aide aux divisions a été de créer un service de consultation au sein de la Division des communications, en prise directe avec le personnel de la rédaction du *Quotidien* et accessible à toutes les divisions spécialisées. La Division des communications a recruté un journaliste qui avait récemment été directeur du bureau d'Ottawa d'un important quotidien canadien, et lui a adjoint les services d'un analyste subalterne. On a ainsi pu établir un certain équilibre entre l'orientation du journaliste, fortement axée sur les médias, et la sensibilité de l'analyste face aux problèmes analytiques, aux limites des données et aux limites qu'un bureau de statistique se doit de ne pas transgresser lorsqu'il émet ses commentaires. Ces deux personnes ont participé à toutes les réunions du comité de rédaction et élaboré les lignes directrices à partir des discussions du comité. De nombreuses divisions ont tiré profit de ce service, qui est maintenant devenu un élément permanent du programme de communications. Des gestionnaires ont vu dans ce service l'outil le plus utile aux fins du travail de refonte de leurs communiqués.

2.3.3 Cours de formation

Finalement, les divisions spécialisées ont également pu compter sur un cours intitulé « Rédiger un communiqué efficace pour *le Quotidien*.» Ce cours, élaboré à partir des lignes directrices du comité supérieur de rédaction, comprend des informations sur l'importance du projet, sur les conditions de travail et les contraintes des médias, et propose des exercices de rédaction dans un style journalistique. Il est organisé division par division et s'adresse à tous les analystes appelés à écrire pour *le Quotidien*. Les gestionnaires chargés de l'approbation des articles destinés à la publication sont invités à participer au cours avec leur personnel. Les analystes ont en effet fréquemment déploré les réticences des gestionnaires chargés d'approuver les communiqués rédigés selon la nouvelle formule. Les membres du comité supérieur de rédaction sont invités à présenter un exposé d'introduction et les instructeurs sont choisis, autant que possible, parmi les membres du personnel des Communications et les analystes avec lesquels les membres des divisions spécialisées seront en fait appelés à travailler pour la préparation de leurs communiqués. En plus d'inculquer aux stagiaires de nouvelles techniques, le cours donne la possibilité de bâtir des réseaux et d'établir un climat de confiance parmi les participants et les instructeurs. Lorsque toutes les divisions intéressées y auront participé, ce cours deviendra partie intégrale du programme de formation régulier.

3. SITUATION ACTUELLE

3.1 Suspension des travaux du comité

En septembre 1995, la plupart des communiqués étant dorénavant exemptés de l'examen critique du comité supérieur de rédaction, ce dernier a décidé de suspendre ses travaux. Même si on admet avoir observé certains reculs, on tient pour acquis que le maintien du processus d'examen des communiqués ne donnerait lieu à aucune amélioration sensible à court terme. Toutefois, le choix de simplement «suspendre» les travaux a été délibéré, et la remise en marche du processus d'examen n'est pas exclue. Le statisticien en chef continue à faire part de ses commentaires aux divisions spécialisées concernant les communiqués qu'elles produisent, lorsqu'il le juge opportun.

Le service de consultation en matière de communiqués de la Division des communications et les cours de formation qui continuent à se donner dans les divisions spécialisées contribuent toujours au processus de refonte du contenu du *Quotidien*.

4. RÉSULTATS

4.1 Des objectifs largement atteints

En suspendant son travail d'examen, le Comité de rédaction a clairement laissé entendre que ses objectifs avaient été largement atteints. Une comparaison des communiqués du *Quotidien* publiés avant la mise en marche du projet à d'autres communiqués publiés aujourd'hui laisse constater une évolution radicale.

La réaction des médias a été extrêmement positive. Fait intéressant, les journalistes, qu'il s'agisse de chroniqueurs spécialisés en matière d'économie ou d'affaires sociales, de reporters spécialisés ou de reporters généraux, s'entendent tous pour affirmer que la qualité générale et la cohérence des articles, l'analyse sous-jacente et la présentation plus claire facilitent leur travail. Nombre de lecteurs généraux du *Quotidien* ont également laissé savoir qu'ils trouvaient les nouveaux communiqués plus accessibles, plus intéressants et plus enrichissants. Ces commentaires sont particulièrement encourageants compte tenu du fait que la diffusion électronique du *Quotidien* nous permet aujourd'hui de rejoindre un public beaucoup plus vaste. Néanmoins, Statistique Canada a conclu que *Le Quotidien* ne pourrait pas servir efficacement deux maîtres. Les divisions spécialisées s'étaient inquiétées de ce que le nouveau style des communiqués ne satisferait pas leurs clientèle générales et leurs craintes ont été partiellement confirmées. Même si les médias nous permettent de rejoindre des millions de Canadiens, la liste des abonnés directs du *Quotidien* se limite à quelques centaines de personnes. *Le Quotidien* continuera à accorder la priorité aux médias, et les divisions spécialisées ont été invitées à explorer de nouveaux mécanismes, par exemple des services de diffusion ponctuelle par télécopie, qui leur permettraient de répondre aux besoins des autres clients.

4.2 Évaluation des résultats

Les efforts consentis pour améliorer l'efficacité du *Quotidien* ont connu un franc succès. Les preuves quantitatives de l'élargissement de la couverture médiatique — par exemple, l'augmentation du nombre de lignes ou du nombre d'articles parus et portant sur les communiqués l'organisme — ne sont pas toutefois évidentes. Les facteurs qui influent sur la qualité de la couverture sont trop nombreux pour autoriser une analyse globale. Par exemple, un communiqué, même mal rédigé et diffusé il y a deux ans au sujet d'une découverte extrêmement importante bénéficierait quand même d'une couverture beaucoup plus large qu'un communiqué bien rédigé, publié aujourd'hui, mais n'annonçant rien de particulièrement important. Un communiqué mal rédigé peut fort bien bénéficier d'une meilleure couverture, mais cette couverture risque d'être nuisible. Par ailleurs, la couverture accordée à deux communiqués identiques publiés à des moments différents subira l'influence d'événements concurrentiels ou complémentaires. Finalement, la couverture accordée à deux communiqués identiques dépendra évidemment des efforts déployés par le personnel des Communications pour maintenir leurs rapports avec les médias. Une analyse globale quantitative des incidences du projet sur la couverture médiatique ne saurait être possible à défaut de moyens de contrôler ces variables externes.

4.3 Études de cas

Les études de cas portant sur des communiqués particuliers présentent les mêmes problèmes de comparaison, mais nous offrent tout de même une certaine mesure de contrôle sur les facteurs externes. La Division des communications a comparé des communiqués publiés avant et après la réforme et provenant de trois séries économiques régulières : les permis de bâtir, le taux d'utilisation de la capacité industrielle et les comptes économiques provinciaux.

L'article sur les permis de bâtir publié avant la réforme n'a donné lieu qu'à un seul article d'importance mineure, publié dans un seul des journaux recensés par Statistique Canada. Après avoir subi une refonte majeure, ce communiqué a bénéficié d'une large couverture dans deux journaux d'affaires importants et il a également été cité dans un article de fond rédigé par le chroniqueur économique d'une grande chaîne de journaux. En outre, le nombre de citations reprises textuellement était plus grand.

Le communiqué sur les taux d'utilisation de la capacité industrielle publié avant la réforme n'avait lui aussi donné lieu à la publication que d'un seul article dans un journal financier d'importance. Après la refonte, il a bénéficié d'une large couverture dans huit quotidiens, y compris un article en première page de la section économique d'un quotidien qui se targue d'être le «journal national des Canadiens».

Le cas du communiqué sur les comptes économiques provinciaux est spécial. Avant la refonte,

il avait fait l'objet de deux articles dans des quotidiens importants par suite, notamment, de la confusion créée par un communiqué connexe. Après la refonte, les deux communiqués ont été combinés. En outre, à cause de l'intérêt régional qu'ils présentaient, on a fait un effort particulier pour les soumettre à l'attention des médias régionaux. L'effort a porté fruits puisque ce communiqué a donné lieu à la publication de 22 articles dans des quotidiens importants et a fait l'objet de 15 journaux diffusés dans les médias recensés par Statistique Canada.

Sans être concluantes en soi, ces trois études de cas semblent indiquer que, toutes choses étant égales par ailleurs, le nouveau style de rédaction a en fait contribué à accroître le nombre et la longueur des articles publiés à partir de nos communiqués. Pour ce qui est de la qualité de la couverture, les informations analytiques diffusées dans *Le Quotidien* sont clairement — quoique sélectivement — reproduites dans les articles des médias, souvent sous forme de citations intégrales.

4.4 Avantages supplémentaires

La réforme a également eu pour avantage connexe de promouvoir un accueil plus positif et même plus enthousiaste des journalistes appelés à travailler avec nos analystes. La multiplication des contacts avec des journalistes dans un contexte plus convivial a incité les analystes à se montrer plus disposés à fournir des éclaircissements ou des commentaires supplémentaires lors de la diffusion de leurs communiqués. Cette collaboration plus étroite entre les analystes et les journalistes a également contribué à accroître et à améliorer la couverture des communiqués ainsi qu'à augmenter l'utilisation des informations de Statistique Canada dans la préparation d'autres articles.

5. CONCLUSION

Les changements apportés dans les communiqués du *Quotidien* semblent avoir encouragé les médias à faire un meilleur usage des informations préparées à l'intention du public canadien par Statistique Canada sous une forme à la fois pertinente et utile. Statistique Canada a donc réussi implicitement à utiliser les médias pour mieux informer le public.

Nos contacts avec d'autres bureaux nationaux de statistique semblent indiquer qu'un certain nombre d'entre eux cherchent également à apporter des améliorations semblables à leurs communiqués. L'exemple de Statistique Canada pourrait fournir ou, à tout le moins, suggérer les moyens d'apporter des changements radicaux en cette matière. Les changements réalisés n'ont pas découlé d'un plan élaboré a priori. Ils ont plutôt été l'aboutissement d'une période d'expérimentation par tâtonnements. Les expériences concluantes ont été retenues alors que celles qui ne fonctionnaient pas étaient abandonnées. Finalement, il est devenu évident que les résultats désirés ne pourraient être atteints ni par une démarche descendante, ni par une démarche ascendante.

Dans un autre ordre d'idée, l'expérience de Statistique Canada nous a fourni une étude de cas fascinante sur les démarches nécessaires à l'instauration de changements radicaux et rapides dans un contexte décentralisé. Elle a illustré les multiples degrés de soutien et l'énergie soutenue qu'il est nécessaire de consacrer à la modification de la culture d'une organisation. Beaucoup des leçons apprises sont applicables à d'autres programmes assez différents. Pour réaliser des changements, il ne suffit pas toujours de savoir ce qu'il y a à faire.

SESSION 7

Stockage informatique des données

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

NOUVELLES TECHNIQUES DE COLLECTE ET DE DIFFUSION DES DONNÉES

W.J. Keller et W.F.H. Ypma[1]

RÉSUMÉ

Les auteurs décrivent brièvement certains progrès réalisés par Statistics Netherlands dans le domaine de la technologie de l'information. Ils donnent d'abord une vue d'ensemble des effets de ces progrès sur le processus de production, puis s'intéressent à deux volets particuliers, à savoir l'échange électronique de données (EED), pour la collecte des données, et l'usage d'Internet, pour leur diffusion. Parmi les nombreux projets en cours à Statistics Netherlands, ils décrivent le «Projet-pilote EED 2» qui vise à appliquer l'EED aux comptes financiers des entreprises. Ils discutent aussi du système de pages d'accueil dynamiques, baptisé WITCH, mis au point par l'organisme. En ce qui a trait tant à la collecte qu'à la diffusion des données, ils soulignent le rôle de la méta-information comme instrument de contrôle du processus. Ils montrent comment le progrès technologique modifie ce rôle et permet de produire des méta-données plus efficaces.

MOTS CLÉS : Statistiques officielles; collecte des données; diffusion des données; EED; Internet; méta-information.

1. INTRODUCTION

Plusieurs changements sont en cours à Statistics Netherlands. À l'instar de ce qui se passe ailleurs, l'Institut ne fonctionne plus comme un organisme public intouchable. L'efficacité et l'offre de services axés sur le marché sont les nouveaux mots d'ordre. Nous devons non seulement réduire nos coûts de production mais aussi ceux que nous faisons encourir à nos fournisseurs de données. Ces efforts devraient se concrétiser par un produit qui, même s'il n'est pas encore commercialisé, répond aux exigences des clients.

Nous devons également tenir compte des progrès réalisés dans le domaine de la technologie de l'information. Ces progrès nous permettront, en effet, de créer les instruments nécessaires pour répondre aux nouvelles demandes. Dans une telle conjoncture, il est essentiel que tout institut national de la statistique prenne des décisions stratégiques judicieuses.

2. INFLUENCE DE LA DEMANDE

D'une part, le processus de production est influencé par les exigences de plus en plus nombreuses des clients et des répondants. Les politiciens demandent avec insistance que nous diminuions le fardeau de réponse afin d'alléger la charge administrative des entreprises. Statistics Netherlands envoie 1,25 million de questionnaires par an aux entreprises et à d'autres établissements. Les grandes et les moyennes entreprises reçoivent jusqu'à 50 questionnaires par an, y compris les questionnaires répétitifs des sondages mensuels et trimestriels. En particulier, les grandes entreprises du secteur manufacturier font l'objet d'une grande variété d'enquêtes. La conclusion est claire : Statistics Netherlands doit s'efforcer d'alléger le fardeau de reponse qu'entrainent les nombreux formulaires à remplir. En outre, les compressions budgétaires nous obligent à devenir plus efficaces et à augmenter la productivité.

Les clients, quant à eux, demandent que nous leur offrions des produits plus faciles à utiliser. Ils souhaitent notamment que nous produisions des données, dans l'ensemble, plus cohérentes. Qui plus est, ils veulent pouvoir utiliser les nouveaux supports de données qu'offre la technologie de l'information.

---

[1] W.J. Keller, Statistics Netherlands, Director of Division Research and Development, P.O. Box 4000, 2270 JM Voorburg, Pays-Bas; W.F.H. Ypma, Department Statistical Methods, P.O. Box 4000, 2270 JM Voorburg, Pays-Bas.

3. POUSSÉE TECHNOLOGIQUE

D'autre part, nous bénéficions des progrès accomplis dans le domaine de la technologie de l'information, c'est-à-dire de la poussée technologique. En premier lieu, ces progrès élargissent l'éventail des possibilités techniques, autrement dit nous donnent les moyens de créer de nouveaux instruments de production. Aussi entrevoyons-nous d'importantes améliorations en ce qui a trait au traitement, au stockage et à la transmission des données. Ce dernier élément, qui concerne la communication de données entre les répondants et l'Institut national de la statistique (INS), d'une part, et entre l'Institut et ses clients, d'autre part, est sans doute celui qui aura l'incidence la plus marquée sur nos travaux.

En second lieu, le progrès technologique crée, en soi, une demande. La nouvelle technologie sera bientôt en application partout. Nos fournisseurs de données l'utiliseront, au même titre que nos clients. Dorénavant, ces intervenants ne seront plus satisfaits de communiquer avec nous selon les anciennes méthodes, c'est-à-dire par écrit. Nos fournisseurs produisent leurs données selon des méthodes électroniques et, afin de réduire leurs dépenses au minimum, voudront se servir de ces méthodes pour nous les transmettre. De même, nos clients traitent électroniquement les données que nous leur fournissons. Par conséquent, ils exigeront de pouvoir sélectionner et recevoir ces données à l'aide des instruments que la technologie de l'information met à leur disposition.

Ces deux facteurs nous obligent à conclure que l'Institut national de la statistique devra s'efforcer de prendre, au sujet du processus de production, les décisions stratégiques qui lui permettront de profiter au mieux des possibilités offertes par la technologie de l'information.

4. DÉCISIONS STRATÉGIQUES

Les nouvelles demandes et les nouveaux instruments auront des répercussions sur tous les aspects de notre processus de production. Avant de décrire ces répercussions, il convient de distinguer trois étapes du processus de production. La phase d'entrée est celle durant laquelle les données sont collectées auprès des répondants. Durant la phase de traitement, ces données sont manipulées afin de produire des renseignements ayant les caractéristiques souhaitées. Durant la phase de sortie, les renseignements sont offerts aux clients et diffusés.

Commençons par la phase d'entrée, c'est-à-dire la collecte des données. Examinons d'abord la collecte des données auprès des particuliers et des ménages. Nous pouvons déclarer sans exagération que, dans ce domaine, Statistics Netherlands a déjà réalisé des progrès considérables. Nous avons introduit toute une gamme de méthodes d'interview assistée par ordinateur (IAO) et, à cette fin, mis au point le système BLAISE. (Il va sans dire que les capacités de BLAISE ne se limitent pas à l'élaboration et à la présentation de questionnaires électroniques.) Ces innovations ont permis, avant tout, d'augmenter la productivité ou l'efficacité. Le nombre d'employés nécessaires pour coder, saisir et vérifier les données a diminué de façon spectaculaire. La production beaucoup plus rapide des résultats des enquêtes témoigne également de ce gain d'efficacité. Toutefois, il serait possible d'augmenter davantage les gains, non seulement en ce qui concerne le processus de production proprement dit, mais aussi sur le plan statistique, grâce à l'introduction de nouvelles méthodes d'interviews (AIAO, auto-interview assistée par ordinateur) et, bien que la question ne soit pas directement liée à la technologie de l'information, grâce à de meilleurs plans d'échantillonnage.

En revanche, beaucoup de progrès restent à accomplir en ce qui concerne la collecte de données auprès des entreprises. Dans ce domaine, les demandes sont plus insistantes. Le fardeau de réponse, qui est devenu problématique, est l'élément autour duquel s'articulent nos décisions stratégiques. Constatant parallèlement que, presque partout, l'automatisation et la technologie de l'information font désormais partie intégrante des systèmes de comptabilité utilisés par les répondants, il nous parait évident que, dans un avenir proche, nous serons obligés d'intégrer l'échange électronique de données (EED) à nos opérations de collectes de données auprès des entreprises. L'EED sera à la collecte de données auprès des entreprises ce qu'est l'IAO à l'interview des membres des ménages. Nous approfondirons la question de l'EED avec les entreprises plus loin.

En ce qui concerne la phase de traitement, nous recherchons des moyens plus efficaces de traiter nos données. L'IAO et l'EED rendent, certes, une grande partie des opérations de vérification superflues, car les erreurs sont moins nombreuses. Néanmoins, nous espérons apporter d'importantes améliorations grâce à une approche plus efficace et plus rationnelle du processus de vérification. Ici, l'élément essentiel est le traitement des données. Nous avons décidé de ne plus vérifier chaque enregistrement à l'avenir. Il devrait être possible de s'appuyer sur l'ordinateur pour déceler les

erreurs les plus graves et essayer de les corriger. Simultanément, la vérification par ordinateur nous évitera de consacrer du temps et de l'argent à la correction d'erreurs non importantes. Les gains relèvent ici, avant tout, de la productivité.

Venons-en enfin à la phase de sortie des données. C'est à ce stade du processus que les innovations attirent sans doute le plus l'attention du public. De nouveaux supports de données au moyen desquels les renseignements peuvent être présentés aux utilisateurs font leur apparition. Les publications imprimées conserveront vraisemblablement leur place, mais, les utilisateurs, particulièrement les plus professionnels, voudront avoir la capacité de sélectionner et de recevoir par voie électronique les données qui les intéressent. Par conséquent, Statistics Netherlands est en train d'élaborer et de mettre en oeuvre des solutions telles que la production de données sur CD-ROM ou la diffusion de données dans Internet pour satisfaire ce type de demande. Une question plus importante, et peut-être plus difficile à résoudre, est celle de savoir comment il convient de présenter les données sur ces nouveaux supports, car la quantité d'information sera beaucoup plus importante que celle qui figure dans nos publications imprimées. À ce stade, la gestion de la méta-information devient un facteur essentiel.

C'est dans cet esprit que Statistics Netherlands a décidé de mettre au point la base de données STATLINE. Cet effort devrait aboutir à la création d'une base de données conçue en fonction de l'utilisateur final et permettant à ce dernier d'avoir accès à «toutes» nos données. Fait peu surprenant, la principale difficulté consiste à structurer les données. Parallèlement, nous devons résoudre la question de leur manque d'uniformité dûe au défaut de coordination statistique. STATLINE est destinée à jouer un rôle essentiel dans le processus de diffusion des données. Aussi avons-nous pris la décision stratégique de définir une structure de données qui permettra de produire toutes les publications et d'effectuer toutes les autres diffusions de données par le biais de cet instrument.

Nous discuterons de la base de données STATLINE plus en détails, en particulier dans le contexte d'Internet, aux sections 11 et suivantes.

5. RESTRUCTURATION DU PROCESSUS DE PRODUCTION

À la section précédente, nous avons décrit les décisions stratégiques prises quant aux diverses phases du processus de production. Ces décisions, qui dépassent le cadre de la simple mise au point d'un nouvel instrument, auront une incidence sur la structure même du processus de production. Aussi devons-nous nous préparer à faire face aux conséquences inévitables de ces changements. À l'heure actuelle, selon la façon de faire «démodée», le processus de production est structuré en fonction des diverses catégories de statistiques. Pour chacune de celles-ci, c'est-à-dire pour chaque produit final, on conçoit un nouveau questionnaire, on sélectionne des répondants, on traite des données et on publie les résultats, procédé inefficace s'il en est, surtout en ce qui concerne la sortie des données.

Quant la nouvelle technologie sera mise en place, autrement dit dans plus de dix ans, le processus de collecte de données sera tout spécialement visé par la restructuration. Ce ne sera plus la demande, mais l'approvisionnement de données, c'est-à-dire les ensembles de données réels disponibles, qui dictera l'organisation des sources. Chacune de celles-ci sera exploitée une seule fois, de façon exhaustive, afin d'en soutirer toutes les données d'une utilité éventuelle au sein de l'Institut national de la statistique. La collecte des données sera adaptée à chaque source, tant sur le plan technique que conceptuel. (Nous donnerons certains renseignements sur la nature de ces sources dans les sections qui suivent.)

Après avoir collecté les données, nous devrons éventuellement les traduire en concepts appropriés du point de vue de la statistique, les intégrer et, enfin, les transmettre aux divers utilisateurs. Ces derniers sont soit membres de l'INS, comme les systèmes d'intégration dont celui des comptes nationaux est un exemple, soit des clients externes. Autrement dit, nous devons rassembler les données à un point donné aux fins de les diffuser. L'étape de sortie des données peut être illustrée de la façon suivante:

Processus : ancien c. nouveau (2000+)

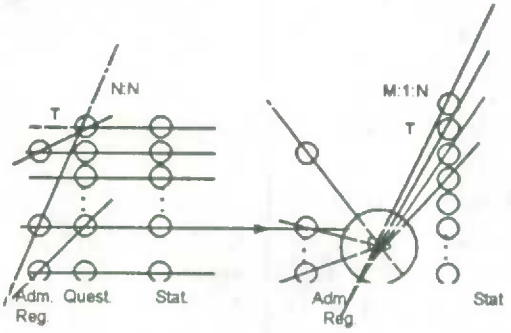


Figure 1

La portion gauche de l'illustration représente l'ancienne situation, avec une ligne de production distincte pour chaque statistique. La portion droite représente la future situation, où toutes les sources possibles contribueront à une base de données centrale contenant des renseignements pertinents. Les statistiques réelles seront produites à partir de la base de données, après combinaison des renseignements pertinents. Il est évident qu'afin de pouvoir combiner les données, on doit veiller à ce que les caractéristiques de ces dernières soient telles que les combinaisons aient un sens. La méta-information précise ces caractéristiques.

6. ÉCHANGE ÉLECTRONIQUE DE DONNÉES (EED)

Concentrons-nous maintenant sur l'EED avec les entreprises et les établissements. Les INS collectent des données afin de produire des renseignements statistiques. Donc, les données collectées auprès des répondants doivent être converties en données de sortie. Cette traduction comporte plusieurs étapes. La première est parfois laissée aux soins des répondants, ce qui, le cas échéant, se traduit par un alourdissement du fardeau de réponse.

La première étape de la traduction comprend deux volets. En premier lieu, il convient d'effectuer une traduction d'ordre conceptuel, c'est-à-dire établir la correspondance entre les concepts de la source, autrement dit les concepts administratifs, et ceux utilisables par l'INS. C'est là l'étape la plus difficile. Les données qui figurent dans les dossiers des entreprises diffèrent non seulement des données statistiques, mais aussi les unes par rapport aux autres. La deuxième étape de la traduction est d'ordre technique. En effet, nous aimerions recevoir des données dont la présentation est appropriée d'un point de vue technique. Plus précisément, nous, et nos répondants, souhaiterions ne pas devoir saisir des données.

7. MODES DE L'EED

L'échange électronique de données sera l'un des instruments stratégiques qui nous permettront de relever le défi que pose l'allégement du fardeau de réponse et l'amélioration de la productivité. Nous devrons examiner séparément chaque situation afin de décider s'il y a lieu d'appliquer l'EED et, le cas échéant, selon quel mode. Nous allons donc décrire plusieurs modes d'EED et les évaluer d'après leur effet sur le fardeau de réponse. Pour chacun, nous préciserons la nature de la traduction à effectuer et à qui en incombe la responsabilité. Nous nous concentrons ici sur la traduction conceptuelle.

7.1 EED stockées dans des régistres centralisés

Ce mode de collecte de données n'exige aucun contact avec les répondants. Il s'agit de renseignements sur des unités distinctes, stockés dans des banques de données centralisées, qui sont collectés à d'autres fins que la production de statistiques, mais qui présentent néanmoins un intérêt pour le statisticien. En soi, ce mode de collecte de données ne crée aucun fardeau de réponse.

Il présente néanmoins certains inconvénients, le plus important étant le choix très limité quant au contenu conceptuel des données reçues par l'INS. Autrement dit, ce dernier ne peut exiger qu'une traduction limitée à des concepts statistiques et doit effectuer lui-même le gros du travail.

Le deuxième problème, étroitement lié au premier, concerne les unités et les populations. Ici encore, nous n'avons pas d'autre choix que d'accepter les données que le registre est capable de nous fournir. Or, il se peut que les unités utilisées ne correspondent pas aux unités statistiques. Il en est de même de la classification de ces unités. La question qui se pose est celle de savoir comment nous pouvons relier la population du registre à notre population statistique globale.

Un troisième problème a trait à la stratégie d'échantillonnage. Si le registre nous fournit annuellement des données sur, disons, 70% d'une population que nous avions l'habitude de décrire au moyen d'une enquête par panel, avec groupes de renouvellement de 1 sur 5, quelle stratégie devons-nous adopter en ce qui concerne les 30% restants?

Il existe aux Pays-Bas plusieurs registres dont les données sont utilisables. Certains sont des banques centralisées de données sur les entreprises affiliées aux chambres de commerce. La bande magnétique de ces registres alimente directement notre registre d'unités statistiques. Des données statistiques peuvent aussi être tirées de dossiers fiscaux (impôt sur le revenu des entreprises, TVA) ou sur la sécurité sociale. Dans plusieurs cas (chambres de commerce, impôt sur le revenu des entreprises et TVA), nous exploitons effectivement les sources disponibles ou nous en étudions la possibilité.

7.2 Cabinets d'experts-comptables

Une autre méthode consiste à utiliser les renseignements recueillis par les cabinets d'experts-comptables. Ces derniers gardent dans leurs archives les registres financiers ou les registres de paie d'un nombre

parfois élevé d'entreprises. Ce mode de collecte de données présente, en outre, l'avantage d'avoir accès à un grand nombre de répondants en n'établissant qu'une seule liaison. Qui plus est, le cabinet d'experts-comptables peut fournir d'autres renseignements que, par exemple, ceux contenus dans les dossiers fiscaux. Un des inconvénients de la méthode tient au fait que, pour répondre aux questions de l'INS, le cabinet facturera vraisemblablement des frais que tous ses clients ne seront pas disposés à payer.

Ayant reconnu que les cabinets d'experts-comptables détiennent souvent une grande partie des renseignements dont a besoin l'INS, il existe deux options quant à l'exécution de la traduction des données. Le choix dépend des résultats de l'analyse de rendement. Par exemple, à Statistics Netherlands, les livres de 40% des entreprises étudiées par une des divisions sont tenus par un même cabinet. Dans ce cas, il est profitable que l'INS se charge d'effectuer les traductions requises. Par contre, dans d'autres cas, nous proposons au cabinet de lui fournir le logiciel de traduction.

7.3 EED avec des répondants individuels

Quand on ne peut avoir accès aux types de sources susmentionnés, il est nécessaire d'entrer directement en contact avec les répondants. Le cas échéant, il ne faut pas perdre de vue qu'il convient parfois de distinguer, au sein d'une unité statistique, souvent une entreprise, plusieurs ensembles de dossiers administratifs. Comme nous le verrons plus loin, ces sous-ensembles doivent être abordés séparément et différemment. Les entreprises commerciales tiennent des registres financiers, des registres logistiques (commerce avec l'étranger, état des stocks) et des registres des salaires et de l'emploi. Aux Pays-Bas, la tenue des registres financiers et celle des registres de paie sont strictement séparées.

Examinons maintenant la situation en fonction de l'entité qui effectue la traduction des données.

7.3.1 Traduction des données par l'INS

Un de nos projets d'EED - EFLO - s'inscrit dans ce cadre. Il concerne les données fournies par les municipalités hollandaises. Ces dernières transmettent un ensemble d'enregistrements tirés directement de leur ensemble complet d'enregistrements. La traduction est effectuée à Statistics Netherlands. En ce qui concerne le fardeau de réponse, les avantages sont évidents. Par ailleurs, bien que l'INS fournisse un travail supplémentaire, ce dernier peut être perçu comme un investissement qui dépend de la stabilité des schémas de traduction. Selon nous, une fois ces derniers établis, l'adoption de cette forme d'EED permettra d'améliorer la productivité. Ici, le point important est que nous avons affaire à un nombre limité de répondants (600).

7.3.2 Tranduction des données par le répondant, conformément à un cliché d'enrégistrement standardisé

Dans ce cas-ci, on commence par établir un cliché standardisé d'enregistrement des renseignements. La normalisation touche les aspects tant conceptuel que technique. La conception du logiciel nécessaire à la production des enregistrements est confiée au répondant. Toutefois, il n'est pas toujours possible de travailler avec un cliché d'enregistrement standardisé. Ce dernier ne peut être utilisé que si les données sont déjà normalisées dans une certaine mesure d'un répondant à l'autre. En outre, pour pouvoir utiliser un cliché d'enregistrement standardisé, l'INS doit parfois se rapprocher des concepts appliqués par le répondant. Le cas échéant, une grande partie de la traduction aux statistiques de sortie doit être exécutée par l'INS.

Ce mode d'EED a une incidence manifestement favorable sur le fardeau de réponse, particulièrement quand le cliché d'enregistrement standardisé peut être intégré au logiciel de comptabilité utilisé et mis à jour régulièrement par le répondant.

On trouve deux exemples. Le premier est le système IRIS, c'est-à-dire le système d'EED sur les échanges commerciaux au sein de la Communauté européenne. Le cliché d'enregistrement standardisé mis au point ici a été intégré à plus de 40 progiciels vendus sur le marché hollandais, après certification par Statistics Netherlands. Le second, le projet EGUSES, a trait aux renseignements sur les salaires. Ce sous-ensemble de registres d'entreprise étant hautement réglementé aux Pays-Bas, il a été possible de définir un cliché d'enregistrement standardisé.

7.3.3 Traduction des données par le répondant. Sans cliché d'enregistrement standardisé

Malgré l'application des méthodes de collecte de données susmentionnées, une grande partie des renseignements que nous recherchons ne sont pas saisis, parce que le répondant les possède sous une forme qui s'écarte conceptuellement et techniquement des données que recherche l'INS et de celles que détiennent les autres répondants. Pour obtenir ces données, on dispose des moyens suivants:

- Questionnaires imprimés

De toute évidence, cette solution n'entre pas dans le cadre de l'EED. Nous la mentionnons uniquement par souci de rigueur et pour souligner le fait que le

répondant effectue lui-même toute la traduction des données et qu'il doit répéter l'opération lors de chaque enquête.

- Questionnaires électroniques

Même si, strictement parlant, il ne s'agit, au mieux, que d'un EED partiel, la méthode donne d'excellents résultats quand on l'applique avec le logiciel IRIS sur les échanges commerciaux au sein de la Communauté européenne. (IRIS offre l'option du cliché d'enregistrement standardisé ou de la saisie de données.) Grâce à des fonctions d'aide supplémentaires et à la possibilité d'adapter le questionnaire à un répondant particulier, elle permet aussi d'amoindrir le fardeau de réponse.

- EED «complet»

La dernière option est celle selon laquelle l'INS fournit au répondant le logiciel qui lui permet d'élaborer le schéma de traduction tant technique que conceptuelle des données. Une fois établi, et à condition qu'aucun changement ne survienne, le schéma de traduction peut être utilisé pour produire des données qui seront fournies à l'INS. Le Projet-pilote EED2, qui vise les registres financiers des entreprises, illustre cette situation et est décrit à la section suivante.

Avant d'entamer cette description, résumons les caractéristiques de plusieurs options d'EED avec des entreprises individuelles.

(Sous-)ensembles de registres	Financiers Salariaux Logistiques Tous les registres
Responsable de la traduction	INS Répondant
Données de sortie du répondant	Données non converties Enregistrement standardisé Enregistrement non standardisé Saisie de données: questionnaire électronique questionnaire imprimé

8. PROJET-PILOTE EED 2

À titre d'exemple, nous allons maintenant décrire le projet-pilote EED 2 visant les registres financiers d'entreprises individuelles. Ce projet donne une idée des difficultés à surmonter. Durant sa description, nous pouvons nous reporter au schéma de la section précédente.

Le projet-pilote EED 2 vise les données de comptes financiers individuels. Aux Pays-Bas, ces derniers ne sont qu'un des éléments des comptes d'une entreprise. En particulier, les comptes relatifs aux salaires et à l'emploi en sont exclus. Statistics Netherlands n'a d'autre choix que de se plier à cette division des comptes, qui résulte de l'organisation des systèmes de tenue de livres dans notre pays. Donc, en omettant les questions détaillées sur les salaires, nous avons réuni dans le programme-pilote toutes les questions visant les comptes financiers pour aboutir à un questionnaire combiné.

Le contenu du questionnaire combiné est dicté par les renseignements qui figurent dans les comptes financiers. Aussi réglementée que puisse être notre société, les comptes financiers des entreprises varient fortement en ce qui a trait à l'organisation interne et aux concepts appliqués. En premier lieu, cette situation signifie que nous devons adapter nos questions aux capacités des systèmes automatisés des entreprises. Dans certains cas, cette situation obligera l'INS à effectuer plus d'opérations statistiques pour atteindre la même production. Si nous voulons plus de données, il sera probablement nécessaire de demander aux répondants qu'ils fournissent explicitement des renseignements supplémentaires, par saisie de données. En second lieu, la diversité des répondants signifie qu'il faudra établir et tenir à jour un programme de traduction unique pour chacun d'eux.

Les comptes financiers se distinguent aussi par leur présentation technique. Les entreprises utilisent toute une gamme de progiciels de tenue de livres. Or, il n'existe à l'heure actuelle aucun cliché d'enregistrement standardisé qui permette de sélectionner électroniquement des données à partir de ces logiciels et il est probable qu'on ne puisse en définir un dans un avenir proche. Comme le projet-pilote 2 vise avant tout à diminuer le fardeau de réponse, nous avons décidé qu'il était nécessaire de réduire au minimum la saisie de données.

Nous avons donc dû faire preuve d'ingéniosité pour créer la liaison automatisée souhaitée. La solution consiste à se servir des rapports, ou sorties imprimées, produits par le logiciel. Toutefois, au lieu d'être imprimés, ces rapports sont envoyés dans un fichier d'impression afin d'être lus par le logiciel de traduction (traducteur) qui constituera l'élément principal du module exécuté par l'ordinateur du répondant et qui est mis au point à l'heure actuelle en tant qu'élément du projet-pilote 2. La présentation des rapports, donc des fichiers d'impression, est relativement constante. Le répondant communique cette présentation au traducteur en définissant les rangées et les colonnes du rapport.

Puis, il lui indique comment manipuler ces rangées et ces colonnes afin de convertir les données du rapport aux données statistiques nécessaires pour répondre au questionnaire combiné. Les enregistrements résultants sont envoyés à Statistics Netherlands.

Le traducteur

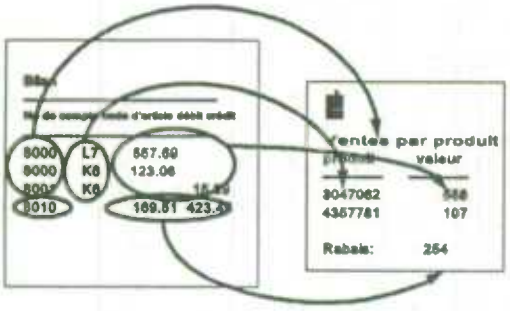


Figure 2

La figure 2 illustre les deux éléments du programme de traduction. Le premier correspond à la présentation des fichiers d'impression qui servent à la traduction technique. Le second définit la traduction conceptuelle des données qui figurent dans le fichier d'impression aux données statistiques nécessaires pour répondre au questionnaire combiné.

La question finale est celle de savoir qui exécutera le programme de traduction. Un des principes du projet-pilote 2 veut que «le répondant exécute la traduction». Autrement dit, il incombe au répondant d'établir le programme de traduction. Cette situation rend le système moins convivial pour le répondant, mais il semble que Statistics Netherlands ne puisse établir ces programmes. Il est évident que la mise au point de tels programmes ne sera pas une tâche facile pour le répondant. Cela signifie, d'une part, que nous devons former un groupe d'assistance compétent et offrir des services sur le terrain relativement importants et, d'autre part, que, même dans le cadre du projet-pilote 2, nous n'atteindrons pas l'ultime convivialité en matière d'EED.

Selon nous, le programme de traduction sera relativement stable, autrement dit, les modifications techniques ou conceptuelles seront peu fréquentes. Une fois établi, le traducteur pourra être utilisé lors de sondages subséquents pour produire les données statistiques demandées. Répondre au questionnaire combiné sera alors une question de minutes plutôt que d'heures, et la tâche pourra être exécutée par du personnel moins qualifié. Ce sont ces éléments qui rendent le concept intéressant et justifie l'investissement initial aux yeux du répondant.

9. PORTÉE DU PROJET-PILOTE EED 2

Comme nous l'avons mentionné, le projet-pilote 2 vise les comptes financiers. En principe, tout produit de Statistics Netherlands qui utilise des données tirées de ces comptes les obtiendra grâce au projet-pilote 2 à condition que l'extraction automatique soit possible. En pratique, cela signifie que plusieurs grands produits passeront complètement au mode de collecte de données par EED. En ce qui concerne le secteur de l'industrie, notre cible principale, le projet-pilote 2 produira les données suivantes:

- Statistiques mensuelles sur le chiffre d'affaires total
- Statistiques mensuelles sur le commerce extérieur par produit
- Statistiques trimestrielles sur le chiffre d'affaires par produit
- Statistiques annuelles sur l'investissement brut
- Statistiques annuelles sur le processus de production
- Statistiques annuelles sur le processus financier, y compris les bilans

La participation du commerce extérieur constitue un projet-pilote au sein du projet-pilote 2. En effet, Statistics Netherlands a non seulement déjà mis en place un programme d'EED qui donne de bons résultats dans ce domaine, grâce à IRIS, mais doit aussi déterminer s'il est possible d'obtenir suffisamment de données sur le commerce extérieur en visant en premier lieu les comptes financiers.

En ce qui concerne les statistiques susmentionnées, certaines questions, comme celles concernant la quantité d'énergie utilisée, posées pour obtenir les statistiques sur la production, ont été éliminées, parce que la forme d'EED en question ne permet pas de les aborder. Il faudra vraisemblablement envoyer un questionnaire imprimé distinct sur ce sujet.

Par ailleurs, certaines questions en provenance d'autres produits et visant principalement d'autres sujets et comptes (par exemple les comptes relatifs à l'emploi et aux salaires) ont été incluses, parce que les réponses sont habituellement obtenues à partir des comptes financiers de l'entreprise.

Le domaine de l'EED regroupe les entreprises commerciales qui ont établi des registres financiers au moyen d'un logiciel satisfaisant certaines spécifications techniques. En pratique, cela signifie que nous nous adressons au secteur de profits au sein du secteur de l'industrie, du commerce et des services. Nous visons l'industrie en premier lieu, car c'est dans ce domaine que

les progrès quant à l'allégement du fardeau de réponse seront les plus importants. Les plus petites entreprises individuelles ne sont pas incluses, car leur capacité est vraisemblablement faible en ce qui a trait à la comptabilité et à l'automatisation. Étant donné que la quantité d'information demandée est relativement faible, nous prévoyons tirer plus de données sur ces entreprises des banques de données centralisées (TVA, impôt sur le revenu des entreprises) et auprès des cabinets d'experts-comptables qui tiennent souvent les livres de centaines de petites entreprises. Les très grandes entreprises sont exclues également. En raison de leur complexité, elles doivent faire l'objet d'un examen séparé, qui, en dernière analyse, sera effectué par EED, mais «personnalisé» dans ce cas.

En ce qui concerne le nombre d'entreprises visées par le projet, il convient de mentionner que 12 répondants ont participé au projet-pilote1 et continuent d'y participer. Le projet-pilote 2 débutera en mars, par un essai sur le terrain visant 20 répondants. Puis, à partir de septembre 1996, nous nous adresserons à un plus grand nombre d'entreprises. À la fin de 1996, le projet-pilote 2 devrait traiter les données de plusieurs centaines de répondants. Nous nous servirons également de ce programme pour obtenir les données des cabinets d'experts-comptables, de sorte que nous augmenterons le nombre d'unités statistiques décrites grâce à une seule liaison EED. Si le projet-pilote donne de bons résultats, en 1997, nous nous efforcerons d'obtenir des données auprès de 25 000 unités au moyen de cet instrument, en partie, par l'intermédiaire des cabinets d'experts-comptables.

S'il est couronné de succès, le projet-pilote EED 2 aura avant tout pour résultat d'alléger le fardeau de réponse. Les gains de productivité ne seront pas très importants. En premier lieu, toute une gamme d'activités devront encore être accomplies. Tous les répondants n'accepteront pas de participer, les données devront être vérifiées, etc. En second lieu, le programme suscitera de nouvelles activités puisqu'il faudra élargir le groupe d'assistance et le groupe de services sur le terrain, et que leurs membres devront résoudre non seulement des problèmes de comptabilité, mais aussi d'automatisation.

10. CONTRÔLE DU PROJET-PILOTE EED 2: LE MÉTA-SYSTÈME

Éventuellement, Statistics Netherlands vise à atteindre plusieurs milliers de répondants. Cette entreprise exige évidemment la mise en place d'un système de contrôle destiné à s'assurer que le questionnaire approprié est produit et envoyé aux répondants, à vérifier la rapidité de production des réponses, à vérifier et à stocker les données d'entrée, à contrôler les rétroactions éventuelles, etc. Autrement dit, il est nécessaire de tenir à jour une grande quantité de données (des méta-données) sur les répondants.

Un autre aspect de la méta-information est lié au contenu du questionnaire combiné. À titre d'exemple, nous nous concentrerons sur cet aspect.

Pour concevoir le questionnaire combiné, il est nécessaire de coordonner non seulement les différents produits qui tirent des données des comptes financiers, mais aussi les pratiques comptables des répondants. Ce dernier élément devait déjà être pris en compte auparavant, mais il devient plus explicite dans le cadre de l'EED. Il a donc fallu entreprendre certaines négociations. Il est évident qu'une fois l'EED bien établi, les divers produits perdront une grande partie de l'autonomie dont elles jouissaient antérieurement, particulièrement en ce qui concerne le questionnaire.

Le module contenant le traducteur nous permet de fournir plus facilement qu'avant la méta-information au répondant. Il comprend les fonctions habituelles d'aide en direct. Les explications sont reliées grâce à l'hypertexte. En outre, nous établirons probablement un système plus détaillé de fonctions d'aide et d'explications à l'intention du groupe d'assistance et du groupe de services sur le terrain. Le système contiendra non seulement des renvois, mais aussi des règles de calcul simples qui permettront, par exemple, d'effectuer des totalisations.

À cette fin, nous avons implanté un ensemble de variables dans une base de données, avec les noms, le texte des questions, les explications et, au besoin, les relations de calcul avec d'autres variables. Les variables, le texte des questions, les explications, etc. sont tirés de cette base de données et combinés au questionnaire. Les répondants sont répartis en groupes selon la taille de l'entreprise, le secteur d'activités et le type de registres financiers tenus. Parfois, les registres des ventes sont tenus par les entreprises, tandis que les bilans annuels sont préparés par un cabinet d'experts-comptables. Le cas échéant, l'ensemble des renseignements nécessaires pour l'unité statistique en question devra être obtenu au moyen de deux questionnaires différents, s'adressant à deux unités de déclaration distinctes. Un questionnaire combiné est établi pour chaque groupe.

11. EED : CONCLUSIONS

En procédant ainsi, nous voyons se dégager un vaste ensemble de méta-données sur les concepts. Cette méta-information contrôle le processus de collecte des

données. Une question visant les comptes financiers ne peut arriver à destination qu'en passant par la base de données centralisée. Au moment d'entrer la variable dans la base de données, il est nécessaire de préciser sa relation avec le reste du document. Autrement dit, elle doit s'intégrer à l'ensemble.

En premier lieu, nous notons que le caractère de la méta-information évolue. La plupart des auteurs décrivent la méta-information comme un élément purement descriptif, disponible uniquement si le statisticien a eu le temps de le produire, généralement après la production des données statistiques demandées par l'utilisateur. Plus tard, rien n'empêche le statisticien de s'écarter de la méta-information établie et rien ne garantit que celle-ci sera mise à jour.

En revanche, ici, les méta-données doivent être établies avant que le processus de production ne démarre. Le statisticien n'a pas d'autre choix que d'utiliser le système de méta-information. Ce dernier est devenu partie intégrante du processus de production. Descriptive auparavant, la méta-information devient désormais normative. Précédemment, nous avons observé le même phénomène en ce qui concerne la collecte des données auprès des ménages grâce au système BLAISE.

Néanmoins, ici, le retentissement est plus profond. Remontons aux premières sections de l'article. Nous y avons mentionné les nouvelles exigences imposées à Statistics Netherlands, notamment celle visant l'allégement du fardeau de réponse. Y donner suite représente l'objectif principal du projet-pilote EED 2. Toutefois, nous découvrons aussi comment la poussée technologique nous donne l'occasion de répondre à une autre exigence, à savoir la production de données plus cohérentes. Le mode de mise en oeuvre de l'EED que nous avons choisi mènera indubitablement à une plus grande coordination statistique (sur le plan conceptuel). Nous avons souligné les possibilités du méta-système et nous observons également, dans le cadre de l'EED, le regroupement de plusieurs statistiques produites antérieurement selon des procédés distincts et indépendants. Fait remarquable, au lieu d'être due à des directives de centralisation, cette amélioration de la coordination statistique est le sous-produit des instruments utilisés au cours du processus de production. Nous ne prétendons pas pouvoir résoudre tous les problèmes de cohésion que présentent nos produits finis, autrement dit tous les problèmes de coordination statistique, en mettant simplement au point l'instrument approprié. Cependant, nous pensons que nous réaliserons d'autres améliorations en la matière si nous tirons judicieusement parti des progrès technologiques.

12. DIFFUSION

Tournons-nous maintenant vers la phase de sortie du processus statistique. À l'heure actuelle, la plupart des bureaux de la statistique produisent des données statistiques agrégées sous diverses formes, mais principalement sur support papier. L'impression étant un procédé de diffusion relativement lent et coûteux, de plus en plus de personnes voient dans l'autoroute électronique (c.-à-d. Internet) un moyen aisé et bon marché de diffuser l'information statistique. Dans le présent article, nous nous concentrons sur les répercussions de cette tendance sur les statistiques officielles. Nous argumenterons qu'outre les considérations technologiques liées à la publication sur le réseau Internet, les principaux problèmes, d'ordre conceptuel, ont trait à la coordination et à l'intégration statistiques.

Nous examinerons certains projets de diffusion électronique en cours à Statistics Netherlands (SN). Nous parlerons de Statline (notre base de données statistiques équipée d'un instrument traditionnel d'interrogation en direct installé dans le système DOS du client) et de sa nouvelle version expérimentale, Statline-Witch, avec ses pages d'accueil dites «dynamiques» sur Internet. Nous montrerons qu'en combinant la facilité d'utilisation et d'accès d'Internet aux bases de données multidimensionnelles utilisées en statistique, il est possible d'élargir considérablement le champ des possibilités en matière de diffusion des données.

13. PUBLICATION ÉLECTRONIQUE DANS INTERNET

À l'heure actuelle, nos publications prennent diverses formes: imprimés, disquettes, télécopies et disques compacts, réponse vocale automatique ou humaine, communiqués de presse, vidéotexte, etc. Ces diverses formes de publication s'appuient sur des données statistiques (agrégées), souvent exploitables par machine, comme les données de sortie des systèmes de traitement des données d'enquête. Par conséquent, nous devrions créer, entre les systèmes de traitement interne et le monde extérieur, une banque «centralisée» de données à diffuser, capable de produire des données sur des supports très variables, régulièrement, dans les délais prévus et efficacement. En plus de l'accès en direct à la base de données, un tel système fournirait automatiquement les données pour d'autres supports, tels que les disquettes, les abonnements au courrier

électronique, les télécopies et les publications sur disque compact. Toutefois, l'un des objectifs les plus importants d'un tel système consisterait à faciliter pour nos clients l'accès en direct à la mine de données que possèdent les bureaux de la statistique. Selon nous, à cet égard, Internet jouera un rôle très important dans un avenir proche.

En raison de l'apparition des navigateurs graphiques (Mosaic, Nestscape) sur le réseau Internet, ce dernier s'est agrandi de façon spectaculaire l'année dernière. Il n'a fallu que quelques mois pour que toute entreprise respectable installe son propre «serveur Web» dans le World Wide Web (WWW). Le réseau Internet, dont la popularité monte en flèche et dont l'infrastructure devient par conséquent immense, relie déjà des dizaines de millions de personnes partout dans le monde, son accès devenant plus aisé et les communications, pratiquement gratuites (en Hollande, à la fin de 1995, pratiquement n'importe qui pourra se raccorder à Internet d'un simple appel téléphonique local, à la vitesse de 28 kilobits/seconde). Le réseau Internet permet aux statisticiens d'effectuer plus efficacement non seulement la collecte des données (voir notre article sur l'EED), mais aussi la diffusion des données statistiques agrégées, à un coût marginal de reproduction et de distribution pratiquement nul. (On compte déjà 27 000 internautes abonnés gratuitement au serveur Top Ten List de David Letterman. Imaginez la situation si nos communiqués de presse bénéficiaient d'une telle diffusion!)

À l'heure actuelle, plusieurs bureaux de la statistique publient des données sur Internet par la voie du World Wide Web. Les serveurs du US Census Bureau, de Statistique Canada, d'Eurostat et de Statistique Netherlands, pour n'en nommer que quelques-uns, sont au nombre des serveurs Web bien connus. Quiconque est raccordé au réseau Internet et équipé d'un navigateur tel que Netscape peut avoir accès à ces serveurs n'importe où dans le monde. Toutefois, la plupart du matériel publié grâce à ces serveurs Web ne contient pas réellement des données statistiques, mais plutôt des listes de publications, des communiqués de presse et des renseignements généraux à l'intention du public. Le nombre limité de données statistiques diffusées de cette manière est souvent présenté de façon documentaire, c'est-à-dire sous forme de copie électronique des pages imprimées des publications traditionnelles.

Cette approche, typique des *pages d'accueil statiques*, rend difficile la manipulation des chiffres statistiques en tant qu'information structurée, puisque l'utilisateur n'a accès qu'à des documents, c.-à-d. du texte (formaté). Il serait donc vraiment nécessaire d'avoir accès (grâce à Internet) à une vraie *base de données* regroupant diverses sources statistiques sous forme de système intégré. Une fois que nos données statistiques seront disponibles sous une forme structurée, exploitable par machine, nous pourrons les manipuler et les présenter sous n'importe quelle forme, non structurée (comme un texte) ou structurée (par exemple, comme un tableur). La création d'une base de données structurées est également nécessaire si l'on veut offrir des renseignements statistiques mieux coordonnés et intégrés.

Statline, créée par Statistics Netherlands, est un exemple de base de données statistiques. Elle est fondée sur le modèle client-serveur, selon lequel le frontal (un ordinateur personnel, qui peut se situer en-dehors de Statistics Netherlands) est séparé du processeur dorsal (le serveur base de données, situé à Statistics Netherlands). À l'heure actuelle, le frontal et le dorsal sont raccordés au moyen d'installations traditionnelles de communication de données comme des réseaux locaux, à l'interne, ou de simples lignes asynchrones (formées de lignes téléphoniques et de modems), à l'externe. Afin d'optimiser la performance de la base de données multidimensionnelles (voir la section 4), STATLINE comprend une base de données non relationnelle, de conception «maison», basée sur des fichiers non indexés. Le système frontal articulé sur DOS est rendu convivial grâce à un système comparable au bureau de Windows avec souris où les résultats des recherches sont affichés dans une sorte de tableur multidimensionnel, avec des vues graphiques supplémentaires, y compris des cartes thématiques. Le logiciel frontal de STATLINE est identique au logiciel d'interface que nous utilisons pour nos publications sur disquettes (ou sur disques compacts). Pour le moment, STATLINE ne s'articule pas sur Internet, mais la situation changera quand nous introduirons le concept des *pages d'accueil dynamiques*.

14. PAGES D'ACCUEIL DYNAMIQUES : LA COMBINAISON D'INTERNET À DES BASES DE DONNÉES

Comme nous l'avons mentionné plus haut, les données statistiques qui apparaissent dans Internet grâce aux pages d'accueil ordinaires sont difficiles à manipuler de façon structurée, en raison du caractère documentaire (non numérique) des pages d'accueil. En outre, de nature statique, ces pages doivent être préparées d'avance en mémorisant leur image (documentaire) dans le serveur Web. Ne serait-il pas merveilleux de pouvoir allier la puissance des bases de données en direct, telles que

Statline, à la facilité d'utilisation et d'accès du World Wide Web? C'est ici qu'entre en jeu la *page d'accueil dynamique*. L'idée consiste à utiliser une navigateur, comme Netscape, en tant que frontal de systèmes tels que Statline. Chaque fois qu'un utilisateur demande des données, une interface spéciale, appelée WITCH, convertit la demande à la structure de Statline et produit une page d'accueil sur le vif pour présenter le résultat fourni par Statline.

Un exemple de page d'accueil produite par WITCH, au moyen de Netscape 1.1 avec le support - tableau HTML3, figure ci-dessous.

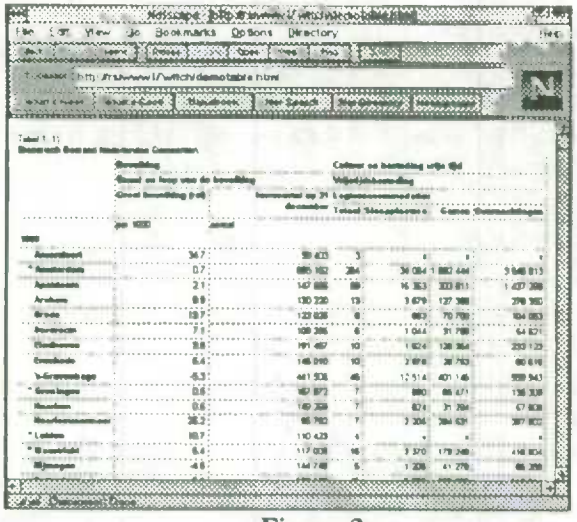


Figure 3

L'utilisation d'un navigateur Web en tant que frontal d'une base de données contenant des données structurées permet de se servir d'autres outils Web. Par exemple, en plus de fournir l'information dans une présentation Web, on peut la télécharger ou utiliser d'autres «visionneurs» dans le navigateur, p. ex., pour visionner des feuilles de calcul, des graphiques ou des cartes à partir d'Internet. En plus des pages d'accueil dynamiques, WITCH produit aussi des présentations telles que des tableurs. De cette façon, l'utilisateur peut mémoriser l'information sous une forme structurée afin de la manipuler plus tard.

Les avantages de cette approche sont multiples. Premièrement, nous ne sommes pas obligés de construire notre propre outil frontal, comme cela a été le cas quand nous avons articulé STATLINE sur DOS. Avec un navigateur Web convenable, n'importe qui peut avoir accès à STATLINE, n'importe où dans le monde. Deuxièmement, en utilisant le navigateur Web, qui est disponible couramment, STATLINE devient accessible immédiatement sur diverses plate-formes (Windows, Mac, Unix). Troisièmement, s'il sait se servir du navigateur Web, l'utilisateur ne doit pas se familiariser avec une nouvelle interface. Enfin, nous pouvons nous servir d'Internet comme support de communication, avec tous les avantages qu'il présente: grande largeur de bande (28,8 kilobits/seconde par modem ou même mieux, dans le cas de la liaison RNIS ou T1) et accessibilité excellente (comme nous l'avons mentionné plus haut, aux Pays-Bas, à la fin de 1995, pratiquement n'importe qui pourra se raccorder au réseau Internet grâce à un simple appel téléphonique local).

Le fait que des millions d'utilisateurs potentiels puissent avoir accès à nos bases de données statistiques en direct grâce à Internet pose de nouveaux défis. Selon nous, la coordination statistique de l'information fournie sera l'élément le plus problématique.

15. COORDINATION STATISTIQUE

La plupart des bureaux de la statistique produisent des centaines de publications statistiques en se basant sur les données de centaines d'enquêtes. Cela représente, globalement, des millions de chiffres, des milliers de totalisations et un très, très grand nombre de sources d'information. Or, mises à part quelques publications spécialisées (comme celles sur les comptes nationaux), chaque publication ne traite que d'un sujet très précis et les utilisateurs se trouvent devant une «mine de renseignements» inaccessible, présentant de multiples facettes. Par exemple, une personne qui s'intéresse à l'automobile doit consulter plus d'une douzaine de publications pour obtenir un tableau complet, englobant la production des automobiles, les exportations et les importations, leur utilisation (en heure et en milles), la consommation d'énergie, les accidents de la circulation, les effets environnementaux, etc. Rechercher toutes ces données peut devenir une tâche pénible, en particulier, parce que les services de statistique ne se concentrent que sur les sujets qui relèvent de leur domaine et sur leurs publications. Parallèlement, l'institut national de la statistique ne vend qu'un nombre très limité d'exemplaires de chaque publication, souvent en ne récupérant pas complètement le coût de la diffusion, sans parler du coût de la collecte. Enfin, si les utilisateurs apprécient notre impartialité et notre précision, ils se plaignent du manque d'actualité de nos données statistiques.

Si toute l'information statistique disponible était offerte gratuitement sur Internet, des millions d'utilisateurs pourraient et voudraient y avoir accès. Comparativement, quelque centaines de personnes seulement lisent nos publications imprimées. Cependant, en plus d'avoir des conséquences

étourdissantes en ce qui concerne la diffusion, une telle situation aurait des conséquences d'ordre conceptuel considérables et vraisemblablement très problématiques. Pour quelle raison? Parce que, si on leur offre l'accès illimité à toutes les données statistiques, les utilisateurs commenceront par demander de meilleures voies d'accès (avec recherche par mot-clé et interrogation multidimensionnelle, en temps réel, par secteur d'activité, par région, etc.). Puis, quand nous leur aurons fournis ces outils, ils découvriront que nos données ne sont pas toujours coordonnées, encore moins intégrées. Des discordances, enfouies dans des centaines de publications imprimées, deviendront visibles dans Internet et les utilisateurs commenceront à réclamer des données non seulement plus nombreuses, mais aussi mieux coordonnées et mieux structurées.

L'adoption de la formule des systèmes, comme dans le cas des comptes nationaux, représente un moyen de répondre à la demande de données mieux coordonnées. Un autre, moins ambitieux, consiste à coordonner les classifications, les domaines et les définitions utilisés dans diverses publications statistiques. C'est sur cette dernière notion que repose le projet de créer une nouvelle base de données, articulée sur des tableaux multidimensionnels ou *compartiments*.

Comme dans le cas des systèmes d'interview assistée par ordinateur (IAO), nous pouvons faire la distinction entre les données proprement dites et leur description, c'est-à-dire les méta-données. Tandis que les systèmes d'IAO mettent l'accent sur les données individuelles et sur les méta-données traitées au stade de la collecte et de la vérification des données, notre base de données à diffuser sera axée sur les données agrégées et sur leurs méta-données. Ces dernières comprennent la syntaxe (structure) et la sémantique (définition des variables publiées, comme celle du «nombre d'employés») des données, ainsi que la description de l'enquête proprement dite, des sources et des méthodes de calcul de divers éléments. La première étape en vue de coordonner les publications statistiques consiste à normaliser les définitions des variables.

Chaque élément (p. ex., le nombre d'employés) est souvent disponible pour divers domaines, définis par le recoupement de variables nominales discrètes, comme le secteur, la région ou le temps. La normalisation de ces variables nominales, qui mènent à des classifications, par exemple, par secteur d'activités, par produit ou par région, représente un autre mécanisme important de coordination de la diffusion des données statistiques. La présentation fondamentale de l'information utilisée dans une telle base de données est donc la matrice multidimensionnelle (parfois appelée «compartiment») où une dimension représente les diverses variables (p. ex., nombre d'employés, bénéfices, prix), une deuxième, l'axe du temps (discret) (p. ex., années et mois), tandis que d'autres correspondent à diverses classifications (industrie, produit, région, etc.). Les éléments qui figurent dans la matrice reflètent les mesures («nombre de») de certaines variables («employés») dans le domaine défini par le recoupement des catégories représentées sur l'autre axe («dans l'industrie x, dans la région y, au temps t»). Fréquemment, les catégories sont classées selon divers systèmes de détails (p. ex., une classification des industries à n chiffres, avec n=1...9), souvent hiérarchisés, pour aboutir à divers niveaux de classification.

Les méta-données (descriptions) qui figurent dans cette base de données à compartiments peuvent se rapporter à la matrice dans son ensemble, aux axes et à leurs variables et catégories, ou aux éléments individuels au sein de la matrice. Des problèmes particuliers aux méta-données surviennent quand la définition de certaines catégories (comme les régions, les industries, les produits) varie selon le domaine, en particulier au fil du temps. À titre d'exemple, examinons le cas d'une municipalité. Non seulement le nombre d'habitants que comptait Amsterdam en 1980 se distingue de celui enregistré en 1991, mais la définition de la municipalité proprement dite diffère également pour ces deux années (p. ex., à cause d'une modification des limites). Des problèmes similaires surviennent quand des éléments ne sont disponibles que pour certaines catégories ou certains niveaux de classification, rendant parfois les comparaisons entre domaines impossibles.

Comme nous l'avons expliqué plus haut, dans les bases de données statistiques, l'objet le plus important est celui de type multidimensionnel ou *compartiment*. Une base de données statistiques contient un très grand nombre de compartiments, qui ont parfois des classifications en commun le long de certains axes. En plus de ces objets multidimensionnels, on doit mémoriser et présenter dans la base de données de simples données croisées bidimensionnelles («plates»), comme celles qui apparaissent dans la plupart des publications statistiques classiques, ainsi que des textes (unidimensionnels) comme des communiqués de presse. Toute cette information est décrite (méta-données) dans la base de données, à divers niveaux (allant de l'objet dans son ensemble aux éléments distincts ou cellules). La classification des objets dans des domaines statistiques bien connus (comme les statistiques économiques ou socio-démographiques) et leur sous-classification (p. ex., selon la production, l'environnement, le marché du travail, le bien-être etc.)

facilite la navigation dans cette immense base de données. À cet égard, l'installation d'un système de recherche par mots clés (dictionnaire analogique) à l'échelle de la base de données et permettant à l'utilisateur d'attribuer rapidement le bon objet constitue un instrument de recherche d'information très puissant.

Dans plusieurs pays, on est en train de créer, ou on utilise déjà, des bases de données statistiques articulées autour du modèle des compartiments. Les systèmes les mieux connus incluent PC-Axis, de Statistics Sweden, la base de données ABS, de l'Australian Bureau of Statistics, et la base de données Statline susmentionnée, de Statistics Netherlands. Statline est non seulement un système interne, mais aussi un système ouvert avec accès en direct à la base de données, mis à la disposition des clients en dehors de Statistics Netherlands. À l'interne, elle sera la «force motrice» de la publication des données sur divers supports (papier, disquette, vidéotexte, etc.), dans la mesure du possible, sans intervention humaine (p. ex., avec composition complètement automatique des imprimés). Statline est aussi utilisée pour répondre à toutes les demandes internes (p. ex., pour le soutien aux clients) et pour permettre aux clients de jeter, occasionnellement ou régulièrement, un coup d'oeil sur son contenu, (p. ex., grâce à l'abonnement automatique, par télécopieur ou par courrier électronique, aux publications de données «prises»). Nous l'utiliserons aussi, fait peut-être encore plus important, comme véhicule de normalisation (donc, de coordination) de toutes les données statistiques agrégées, y compris les méta-données.

À l'extérieur de Statistics Netherlands, Statline offre déjà aux clients importants un accès direct à la multitude de données produites par l'organisme. La combinaison de la base de données à l'interface WITCH, en d'exploiant le concept de la page d'accueil dynamique, permettra à de nombreux autres clients d'accéder plus facilement à Statline grâce à Internet.

16. DIFFUSION : CONCLUSIONS

Étant donné la croissance spectaculaire de l'usage d'Internet à l'échelle mondiale, la publication électronique devient rapidement une réalité. Le réseau Internet, en particulier le WWW (World Wide Web), est certes très convivial et facilite considérablement l'accès à un immense univers d'information, mais il pose aussi de grands défis aux statisticiens. Devrions-nous simplement mémoriser nos publications sous forme électronique dans la base de données d'un serveur Web, en gardant la même présentation que pour la version imprimée? Ou bien, devrions-nous présenter nos données statistiques de façon plus structurée? Nous pensons que la technologie de la page d'accueil dite «dynamique» en tant qu'instrument frontal d'une base de données statistiques offre une meilleure solution que la page d'accueil statique qui, dans une certaine mesure, ne fait que reproduire la présentation sur papier.

De façon plus générale, une fois que l'information statistique sera disponible sous une forme structurée, exploitable par machine, comme c'est le cas dans Statline, il suffira de se servir d'interfaces telles que WITCH pour la présenter sous n'importe quelle autre forme. Ce type de base de données permettra de produire de façon complètement automatisée non seulement des pages d'accueil sur le vif, mais aussi des messages transmis par télécopie ou par courrier électronique, des communiqués de presse, des bases de données sur disque compact et même les désuettes publications imprimées. Évidemment, au même titre que les données, les méta-données devront être exploitables par machine, y compris la syntaxe (structure) et la sémantique (contenu) des données. Une fois ces étapes accomplies, l'échange de données entre statisticiens au moyen de présentations d'exportation normalisées, telles que GESMES d'Eurostat, deviendra chose aisée à l'exemple des exportations WordPerfect que nous effectuons actuellement à partir d'un document MS-Word. Il suffit pour cela qu'on établisse pour toutes les données statistiques une présentation de stockage structurée, exploitable par machine et bien documentée.

Toutefois, quand toutes nos données seront disponibles en direct, exploitables par machine et décrites par une grande quantité de méta-données, les utilisateurs commenceront de nouveau à se plaindre aussitôt que des discordances entre les publications électroniques apparaîtront. Nous devrons alors mettre en place des mécanismes de coordination et d'intégration statistiques, comme nous l'avons fait pour les comptes nationaux, mais à plus grande échelle. Un premier pas dans la bonne direction consisterait sans doute à essayer d'intégrer un aussi grand nombre que possible de publications dans un nombre restreint de compartiments. Pour y arriver, il faudra s'efforcer par tous les moyens de rationaliser les définitions et les classifications statistiques. Donc, en dernière analyse, les difficultés d'ordre conceptuel pourraient dépasser celles d'ordre technique.

Enfin, il convient de soulever la question intéressante des coûts et des prix dans l'environnement Internet. Présumant que les renseignements statistiques proprement dits sont du domaine public, les statisticiens établissent souvent le prix des données qu'ils diffusent

uniquement en fonction des coûts marginaux de reproduction et de distribution. Or, dans Internet, la reproduction et la diffusion sont virtuellement gratuites. Aussi est-il raisonnable de se demander quelle attitude il conviendra d'adopter une fois que nous pourrons introduire *tous* nos gigaoctets d'information statistique disponibles dans Internet. Ce service devrait-il être gratuit? La question a suscité des débats animés au sein de Statistics Netherlands. Donc, en plus des problèmes techniques et conceptuels, l'usage d'Internet soulève d'importantes questions stratégiques. Les choses ne seront jamais plus ce qu'elles étaient!

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

GESTION DES DONNÉES POUR LE RÉSEAU D'INFORMATION SUR LA SANTÉ DU CANADA: CRÉATION D'UN ENTREPÔT VIRTUEL D'INFORMATION GRÂCE À L'ÉTABLISSEMENT DE NORMES, DE LIENS DE COOPÉRATION ET DE PARTENARIATS

B. Bradley et J. Silins[1]

RÉSUMÉ

Les auteurs discutent des méthodes et des plans suivis en vue de créer un entrepôt virtuel d'information destiné à répondre aux besoins d'un grand nombre d'organismes interdépendants. En s'appuyant sur le modèle du SGDD/SID et de la Base de données sur la lutte contre les maladies de Santé Canada, ainsi que sur les enseignements tirés de ce modèle, ils mettent l'accent sur le rôle que jouent les métadonnées et les normes dans le partage, ainsi que l'échange transparent et l'intégration de ressources en apparence fragmentées englobant divers organismes, emplacements, mandats, territoires et domaines de compétence. Ils soulignent aussi l'importance de la création de réseaux sociaux interfonctionnels complémentaires, y compris l'élaboration de méthodes et de politiques communes de gestion de l'information, le partage de la propriété et de l'administration, et enfin, l'établissement coopératif, par consensus, de priorités et de plans.

MOTS CLÉS : Systèmes d'accès aux données; entreposage des données; métadonnées.

1. INTRODUCTION

La création du Réseau d'information sur la santé a pour objectif d'augmenter la capacité de surveillance de la santé de la population, des facteurs de risque et de la maladie au Canada. On le met sur pied afin de mieux faire le suivi de l'état de santé, des facteurs environnementaux et des facteurs socio-économiques connexes, d'augmenter la capacité d'enquête et de recherche, et d'assurer une diffusion plus efficace de l'information et des connaissances, au moment et là où cela est nécessaire, afin d'étayer par des faits les interventions relatives aux politiques et aux programmes.

2. RÔLE DE L'ENTREPOSAGE DES DONNÉES

L'«entreposage des données» est un concept très à la mode dans le milieu de la gestion de l'information centralisée. Si certains éléments de l'entreposage des données existent depuis de nombreuses années sous d'autres aspects - par exemple, à titre d'activités ou de programmes associés à «la gestion de l'information en tant que ressource collective», à l'élaboration «de recensements et d'archives de données», à la gestion et à la diffusion de données dans «des structures organisationnelles bidimensionnelles» pour «systèmes informatiques directoriaux», et, dans certains organismes, à titre de «bibliothèques de données», «d'archives de données», ou de fonctions liées à «la diffusion des données, aux services concernant les données ou à la commercialisation des données» - la notion même d'entreposage prête une force industrielle à des activités qui, jusqu'à présent, ont eu tendance à être fragmentées et parfois fragiles.

Le concept d'entreposage fait ressortir la valeur fondamentale des données en tant que matière première importante non seulement en regard des processus intégrés de gestion et d'orientation stratégique, mais aussi à titre de produit de base sur le marché naissant de l'information. Il évoque aussi l'importance d'opérations analogues à l'extraction, au raffinage, à la fabrication, à la mise en rayon, au contrôle des stocks, à l'expédition, à la commercialisation, à la vente et à la livraison.

---

[1] Bill Bradley et John Silins, Santé Canada, Immeuble LCDC, salle 35 Parc Tunney, Ottawa (Ontario), Canada, K1A 0L2.

Les possibilités d'accès généralisé aux données qu'offre la mise en place de réseaux internes reliant la plupart des secteurs de l'entreprise, conjuguées à l'obligation onmiprésente de réduire les effectifs, de restructurer et de mieux rentabiliser les données internes et les investissements dans l'information, semblent être à l'origine de cette nouvelle orientation. En outre, dans le climat technologique parfaitement à point créé par les nouveaux réseaux internes, Internet a frappé avec la rapidité de l'éclair.

Alors que de plus en plus d'employés et de gestionnaires s'aperçoivent que leur ordinateur de bureau leur permet de communiquer instantanément non seulement avec leurs collègues de toutes les divisions et les décisionnaires des bureaux directoriaux, mais aussi avec des particuliers, des sources d'information de toutes formes et des organismes partout dans le monde, le sentiment qu'un phénomène important est en train de se produire s'intensifie. En outre, à mesure que les organismes reconnaissent le pouvoir que leur donnent les nouvelles technologies de remodeler et d'améliorer la façon dont ils conduisent leurs affaires, l'intérêt ne cesse d'augmenter pour les données et pour l'information en tant que produits stratégiques et économiques indispensables, particulièrement les métadonnées.

Car les métadonnées - information *au sujet* des données et de l'information - sont l'élément clé permettant de représenter et de découvrir ce qui est nécessaire dans les réseaux. Les métadonnées sont aussi l'adhésif qui rassemble le tout de façon utile.

Sur la scène statistique, l'Australian Bureau of Statistics (ABS) travaille à un projet d'entreposage centralisé des données depuis 1993 (Colledge et Richter, 1994). Comme le laissent entendre plusieurs exposés faits durant le symposium (par exemple, Priest (1995), Grenier (1995), Boucher (1995)), Statistique Canada se livre actuellement à un certain nombre d'activités ayant pour objet la mise au point de systèmes similaires et a lancé une discussion à l'échelle de l'organisme afin de déterminer comment l'entreposage de données pourrait améliorer les opérations internes et augmenter les recettes.

Aux États-Unis, où le système statistique est relativement décentralisé et où les produits d'information sont diffusés plus librement, l'Executive Office et l'Office of Management and Budget (OMB) appuient vivement la centralisation et l'amélioration de l'accès électronique à l'information. Le Statistical Policy Office de l'OMB (1995) indique que 17 organismes fédéraux utilisent déjà Internet pour diffuser l'information statistique.

De plus en plus conscient de l'importance des métadonnées et des normes en ce qui concerne l'accès électronique, le Geographic Data Committee de l'OMB est en train d'élaborer une norme relative aux métadonnées culturelles et démographiques en vue de faciliter l'identification des données géospatiales, l'accès à ces dernières, leur coordination et leur échange (Federal Geographic Data Committee, 1995). Le Bureau of the Census et le Bureau of Labour Statistics travaillent de concert à la création d'un outil commun permettant de parcourir les métadonnées et d'avoir accès aux données (Capps,1995). Le Bureau of the Census s'efforce aussi de définir et de normaliser des éléments de métadonnées qui seraient applicables à tous ses secteurs (Gillman, 1994; Sundgren et coll., 1995), et collabore avec d'autres à un projet de l'American National Standards Institute (ANSI) et de l'Organisation internationale de normalisation (ISO) ayant pour objectif d'élaborer une norme de description et d'enregistrement des éléments de données statistiques (ANSI X3L8, 1995).

Les éléments clés de ces activités d'entreposage des données statistiques consistent à :

- se concentrer davantage sur les données et les produits d'information sous l'angle de l'organisme dans son ensemble;
- accorder plus d'attention à la saisie de métadonnées lisibles par machine et à l'accès à ces métadonnées;
- intégrer les données et les métadonnées dans un entrepôt collectif d'information, quand cela est possible, ou grâce à l'amélioration des systèmes d'accès;
- mettre davantage l'accent sur l'intégration statistique;
- mettre davantage l'accent sur la normalisation des concepts et des définitions des mesures;
- mieux faire connaître les produits aux clients externes et donner à ces derniers les moyens d'y avoir accès et de les obtenir plus facilement;
- mettre davantage l'accent sur les activités susceptibles de produire des recettes.

Les intervenants les plus importants sont les utilisateurs de données. Les bibliothèques de données des universités, les archives de données statistiques et d'autres organismes utilisateurs de données procèdent à des activités d'entreposage depuis de nombreuses années. Dès les années 60, ces collectivités se sont attaquées, énergiquement et efficacement, à la résolution de la question des métadonnées qui, jusqu'à récemment, étaient appelées «documentation sur les données».

L'Interuniversity Consortium for Political and Social Research (ICPSR) et les U.K Data Archive (UKDA), qui sont vraisemblablement les plus grands entrepôt de données statistiques du monde, sont les principaux acteurs. Il convient aussi de mentionner l'Association internationale pour les services et techniques d'information en science sociale (AISTISS), l'Association internationale des organisations de services de données en science sociale (AIOSD), l'Association of Public Data Users (APDU), la Canadian Association of Public Data Users (CAPDU), et le Council for European Social Science Data Archives (CESSDA). Les activités intensives d'entreposage et de préservation des données, et de création de métadonnées de ces organismes dépassent le cadre de la présente discussion. Le lecteur trouvera dans Bradley et coll. (1990, 1994) une introduction aux activités fondamentales de normalisation des métadonnées.

En Europe, un projet d'entreposage virtuel des données basé sur un catalogue intégré de données vient d'être lancé sous les auspices du CESSDA, (Musgrave, 1995). L'ICPSR et l'APDU, quant à eux, viennent d'entreprendre des travaux de normalisation des métadonnées statistiques. L'ICPSR a formé un comité international d'experts chargés d'élaborer une table de codage normalisée des microdonnées, sous forme de définition du type de document (DTD), utilisable avec les logiciels de traitement et d'édition de textes qui appliquent le langage standard généralisé de balisage (SGML) de l'ISO. Un groupe de travail de l'APDU est en train d'examiner le problème que posent la table de codage et la documentation relative au questionnaire dans le cas des instruments de sondage préparés selon les méthodes de collecte de données d'enquête assistée par ordinateur (Doyle, 1994).

3. DES DONNÉES À L'INFORMATION - ACQUISITION DE CONNAISSANCES DANS L'ENTREPÔT VIRTUEL D'INFORMATION

Comme l'a mentionné Hicks (1995) dans l'allocution d'ouverture du symposium, les technocrates doivent être en mesure de regrouper selon des modes variables, adaptés aux études ou aux séances d'information du moment, des données produites au fil du temps par un grand nombre de secteurs, d'enquêtes, d'organismes et de partenaires. Par exemple, dans le domaine du renseignement sur la santé de la population, il pourrait être nécessaire de montrer les tendances particulières de la mortalité ou de la morbidité en regard d'interventions liées à des politiques ou des programmes et en regard d'autres éléments, tels que les facteurs épidémiologiques, l'état de santé, le mode de vie et les conditions socio-économiques, de comparer des sous-populations, des régions, des collectivités ou des pays, ou d'évaluer les effets d'une gamme de régimes de santé et de sécurité sociale, de milieux culturels, de conditions environnementales et de stratégies de promotion et de prévention. En raison de l'importance accordée à l'adoption d'une politique gouvernementale axée sur la santé - autrement dit l'intervention de tous les ministères et secteurs compétents en ce qui concerne les déterminants de la santé connexes et, dans la mesure du possible, la réduction au minimum des risques pour la santé - l'accès à l'information et la synthèse doivent s'étendre bien au-delà des organismes de santé et des sources de données traditionnelles.

Hicks place aussi les données et l'information statistique dans le contexte plus large de l'acquisition de connaissances aux fins de l'élaboration de politiques sociales, et souligne l'ampleur et la profondeur des données et de l'information nécessaires.

Nous présentons à la figure 1 un modèle qui a été proposé pour faciliter la discussion et la compréhension détaillées des méthodes d'utilisation des données, pour appuyer l'élaboration des politiques et la prise de décisions dans le cadre des programmes, et pour planifier les systèmes de soutien (Alexander, 1990; Bradley et coll., 1994). Aux fins de la présente discussion, nous avons étoffé le diagramme pour donner une idée de l'envergure de l'entrepôt de données nécessaire et souligner la nature virtuelle de ce dernier. La figure 2 fournit la liste de certaines exigences fonctionnelles clés auxquelles doivent répondre les systèmes statistiques nationaux sous-entendus par le modèle, quand on associe ce dernier à des observations sur l'utilisation de l'information pour élaborer des politiques et mettre en place des programmes en matière de santé et de bien être social (Bradley et coll., 1994).

Comme l'indique l'illustration, l'entrepôt inclut non seulement des données, mais aussi l'information et les connaissances tirées de ces données. Selon nous, on ne peut limiter la portée de l'entrepôt aux données uniquement. En tout premier lieu, les chercheurs, les technocrates et les décideurs sont à la recherche de connaissances. Ce n'est que si les connaissances pertinentes font défaut ou sont introuvables qu'ils se tournent vers l'information existante en vue de créer le savoir. Creuser jusqu'aux données pour créer de nouveaux produits d'information est une tâche relativement technique, longue et coûteuse qui, du moins en théorie, ne devrait être effectuée qu'en dernier

ressort[2].

L'expression *entrepôt d'information* évoque donc une gamme de produits réutilisables plus étendue qu'une simple collection de données. Il existe manifestement un chevauchement entre cet entrepôt et les bibliothèques traditionnelles dont le rôle a toujours consisté à entreposer les produits du savoir.

Qui plus est, les différents types de produits que contient l'entrepôt - données, information et connaissances - devraient être complètement intégrés. Après avoir lu un chiffre dans un rapport, combien de fois les statisticiens, les chercheurs ou les décideurs ne se sont-ils pas demandé, ou n'ont-ils pas dû expliquer, d'où il provenait, comment il se comparait à des mesures similaires ou pourquoi il ne semblait pas concorder avec les conclusions émanant d'autres études? Pour répondre à de telles questions dans les délais ordinairement accordés, l'utilisateur doit pouvoir, en cliquant sur le chiffre, faire apparaître les renseignements sur les analyses et sur les processus de collecte et de transformation des données sous-jacents, ainsi que des renseignements similaires et comparables, avec les estimations connexes, en provenance d'autres sources pertinentes.

Similairement, un analyste à la recherche de microdonnées au niveau de la question ou de la variable devrait pouvoir fouiller dans les tableaux et dans les produits d'information et de connaissances dans lesquels la variable étudiée a été utilisée, afin de déterminer quelles analyses ont déjà été effectuées et de se servir des résultats. Les nouvelles technologies hypertextes permettent de relier, pour chaque variable sous-jacente, le niveau des connaissances aux métadonnées en passant par les produits d'information intermédiaires, et vice versa, tandis que le langage standard généralisé de balisage (SGML) permet d'effectuer la normalisation indispensable des produits et les enchaînements au sein de la pyramide.

Les lignes en pointillés tracées verticalement pour diviser la pyramide d'information en secteurs étroits illustre la nature virtuelle de l'entrepôt. Chaque secteur représente un élément de l'entrepôt, c'est-à-dire l'entrepôt d'un des nombreux organismes qui font partie du réseau.

Donc, un secteur pourrait représenter l'entrepôt de Santé Canada, un autre, celui de Développement des ressources humaines Canada, un troisième, celui d'Affaires indiennes et du Nord Canada, en quatrième, celui de Statistique Canada, un autre, celui de Santé Québec, un autre celui du ministère de la Santé de l'Ontario, un autre, celui de la Fondation de la recherche sur la toxicomanie, un autre, celui du Centre for Disease Control d'Atlanta, et ainsi de suite. La liste de partenaires qui échangent des données, de l'information et des connaissances pertinentes est, en principe, infinie.

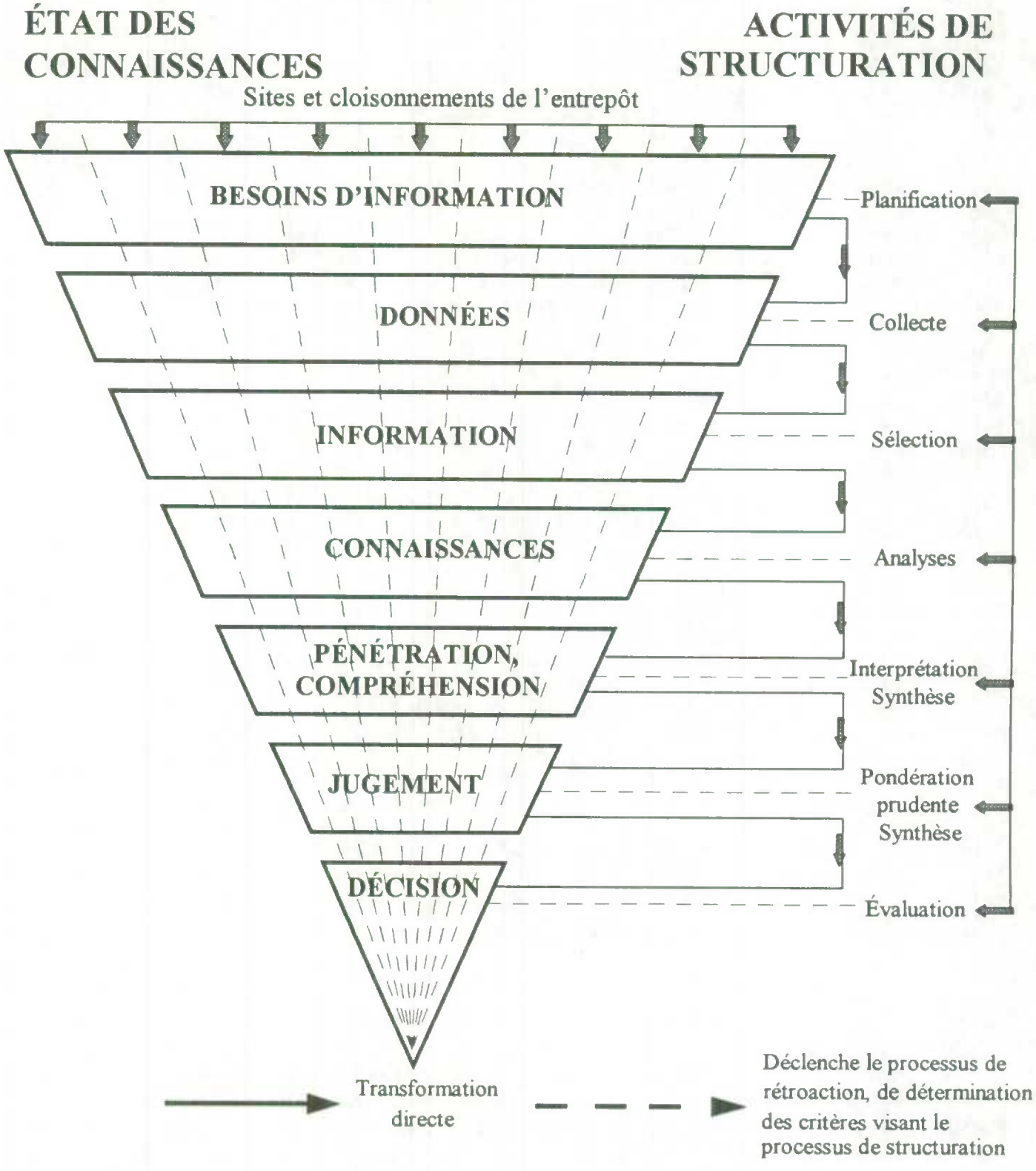
Tel que nous l'envisageons aujourd'hui, l'entrepôt illustré à la figure 1 permettra, grâce à la technologie connexe, l'inclusion ouverte d'éléments de ressource fournis par tous les participants au partage et à l'échange d'information. Les partenaires se joindront au Réseau s'ils ont l'impression de pouvoir tirer profit de cette décision. Ils contrôleront leur propre entrepôt et décideront quelle part du contenu peut ou ne peut pas être partagée. Ils seront co-propriétaires de l'entrepôt virtuel, participeront à son administration et accepteront de se conformer aux politiques, aux normes et aux protocoles d'échange de données et d'information élaborés par le groupe, de façon à en tirer des avantages tant collectifs que mutuels.

4. CLOISONNEMENT ET ENTREPÔT VIRTUEL D'INFORMATION

Priest (1995) fait allusion au problème des organismes «cloisonnés». Ces organismes, particulièrement aux niveaux inférieurs où sont produites les matières premières de l'information statistique, s'intéressent avant tout aux pressions et aux flux verticaux plutôt qu'horizontaux. Les organismes de ce type sont endémiques dans les structures organisationnelles à niveaux multiples et, tout naturellement, ont tendance à construire des systèmes

[2] Néamoins, la fourniture des microdonnées et des installations d'accès connexes demeure la plus haute priorité opérationnelle en ce qui concerne l'entrepôt particulièrement au tout premier stade. Cette situation tient à ce que les produits deviennent de plus en plus spécialisés à mesure qu'on évolue des données à l'information et aux connaissances, et que, par conséquent, il est vraisemblable que les agrégats, l'information et les connaissances ne satisferont pas les exigences existantes. Aux premiers stades de la création de l'entrepôt, il se pourrait simplement qu'il soit difficile de découvrir l'information et les connaissances déjà créées à partir des données, et impossible de les obtenir dans une présentation électronique utilisable, de sorte que, pour de nombreux analystes dont la tâche consiste à fournir des renseignements pour soutenir les décisions et à réagir rapidement, la solution la plus expéditive consiste à créer leurs propres produits sans se référer largement à ce qui a déjà été fait. Selon les auteurs, la saisie et la gestion de cette valeur ajoutée et fortement dispersée, et son intégration à l'entrepôt virtuel est une tâche de toute première importance.

ACQUISITION DE CONNAISSANCES DANS L'ENTREPÔT VIRTUEL D'INFORMATION



Adapté de : Alexander, Cynthia: Towards The Information Edge and Beyond: Enhancing the Value of Information in Public Agencies , Justice Canada, 1990.

Figure 2.
Exigences fonctionnelles auxquelles doivent répondre les systèmes nationaux d'information

Le modèle d'acquisition des connaissances (figure 1) et les observations des auteurs concernant l'utilisation de l'information par le gouvernement en vue d'élaborer des politiques et des programmes (Bradley et coll., 1994), donnent à penser que les systèmes nationaux d'information doivent répondre aux exigences fonctionnelles clés que voici.

i) Les systèmes doivent permettre aux entrepôts et aux entités cloisonnées figurant au haut de la pyramide d'avoir accès à la base la plus large possible de données d'intérêt général, et faciliter l'extraction, l'intégration, le traitement ultérieur et la diffusion de ces données sous la forme d'informations et de connaissances relativement personnalisées, présentant un intérêt pour tous les organismes qui figurent à la base de la pyramide.

ii) Les systèmes doivent être conçus pour permettre le mouvement du haut vers le bas de la pyramide, tant au sein des entrepôts qu'entre ces derniers, aussi rapidement que possible.

iii) Les systèmes doivent permettre la rétroaction et l'itération à chaque étape de progression à travers la pyramide.

iv) Les produits issus de tous les stades du processus - données, informations et connaissances - doivent être préservés, décrits et intégrés dans tous les entrepôts de façon normalisée, structurée, accessible et réutilisable.

v) Quand l'information ou les connaissances disponibles sont douteuses ou insuffisantes, ne satisfont pas les exigences d'acquisition des connaissances du moment, ou soulèvent de nouvelles questions, il doit être possible de creuser à travers tous les niveaux, dans tous les entrepôts, en partant des connaissances jusqu'aux données. Il doit être possible d'examiner et d'évaluer, pour toute application, si les ressources et les produits sous-jacents sont appropriés, et de créer de nouvelles informations et connaissances si cela est nécessaire ou souhaitable.

vi) Comme le caractère personnalisé, filtré et spécial des produits augmente à mesure qu'on évolue des données aux connaissances en passant par l'information, il est de la plus haute importance d'intégrer aux systèmes des capacités de retour en arrière fondées sur l'accès à des microdonnées bien décrites.

vii) Les systèmes doivent permettre d'avancer, de retourner en arrière, de faire des itérations et de synthétiser, autrement dit de se déplacer dans toutes les directions dans la pyramide, tant au sein des entrepôts que de l'un à l'autre, aussi rapidement que possible. La capacité de se rapprocher ou de s'éloigner rapidement du niveau des données est particulièrement importante.

viii) Si on repère des lacunes ou des défauts dans les sources de données existantes, le retour en arrière doit se faire jusqu'au stade de la planification et de la conception de la recherche pour s'assurer que les données pertinentes soient collectées ou acquises aussi rapidement que possible, que les éléments et les définitions soient harmonisées dans l'entrepôt et d'un secteur à l'autre, et que les chevauchements non intentionnels soient réduits au minimum.

ix) Les systèmes doivent être exploités en premier lieu grâce aux ressources et aux entrepôt locaux, mais sa portée doit être globale.

informatiques qui servent des perspectives et des objectifs spécialisés. Ces organismes accordent souvent une faible priorité à l'intégration horizontale.

L'entrepôt virtuel est un concept destiné à faciliter l'harmonisation des données et de l'information entre les diverses unités cloisonnées. On peut concevoir chaque secteur vertical de la figure 1 comme un vase clos issu du cloisonnement. Selon la perspective organisationnelle, un secteur donné peut, en fait, correspondre à une collection de vases clos. Par exemple, l'entrepôt de Statistique Canada pourrait comprendre les produits de données, d'information et de connaissances dérivés du Recensement de la population, du Programme des enquêtes post-censitaires, de l'Enquête sociale générale, de l'Enquête nationale sur la santé de la population, de l'Enquête sur la dynamique du travail et du revenu et des enquêtes de la Division des enquêtes spéciales, pour ne mentionner que quelques-

uns des programmes de l'organisme. L'entrepôt de Santé Canada pourrait contenir l'information et les connaissances produites par les diverses directions générales du Ministère, y compris la Base de données sur la lutte contre les maladies de la Direction générale de la protection de la santé, les nombreux projets financés par la Direction générale des programmes et de la promotion de la santé, les programmes de recherche intra et extramuraux, et diverses bases de données créées par la Direction générale des services médicaux. La tranche du secteur privé pourrait englober les produits du Sondage Santé Canada et de l'Environmental Monitor, ainsi que ceux fournis par l'Environics Research Group, l'Angus Reid Group, et d'autres firmes offrant des produits dignes d'intérêt.

Le modèle peut aussi inclure les organismes de santé publique, les ministères provinciaux et territoriaux, les ONG, tels que l'Institut canadien d'information sur la santé, le Canada Fitness and Lifestyle Research Institute, la Fondation de la recherche sur la toxicomanie et la Fondation des maladies du coeur, ainsi que les nombreux chercheurs universitaires, ministères, instituts, archives et réseaux dont les travaux contribuent au renseignement sur la santé de la population et à l'élaboration de politiques fondées sur des faits au Canada.

L'objectif consiste à offrir une seule fenêtre commune d'accès aux données et aux produits d'information à un aussi grand nombre que possible de représentants de ces divers programmes et services. Dans la mesure où cela est possible, les utilisateurs pourront :

- trouver, parcourir, visualiser et comparer des renseignements fournis par les nombreux serveurs où sont stockées des ressources pertinentes;
- réunir tous les éléments de donnée et points d'estimations comparables;
- effectuer des analyses personnalisées, et donner une plus-value aux données en créant de nouvelles informations et de nouvelles connaissances;
- collaborer en vue de combler les lacunes, d'éviter les chevauchements et d'améliorer la comparabilité et l'efficacité des futures collectes de données;
- saisir et partager la valeur qui est ajoutée à tous les sites et à tous les niveaux de la pyramide d'information.

De ces activités découlera une meilleure harmonisation des concepts et des définitions, grâce à un processus naturel que l'Australian Bureau of Statistics appelle «confrontation» (Colledge et Richter, 1994). Si on facilite l'accès des représentants de tous les programmes aux métadonnées, ceux qui planifient la collecte de renseignements dans divers organismes et secteurs seront en mesure d'observer, de consulter et d'évaluer les diverses définitions et méthodes appliquées. Ils négocieront certaines modifications des définitions en ayant à l'esprit l'intérêt et l'information de tous, et apporteront des corrections, au besoin, en fonction de leurs exigences en matière d'intégration des données, des objectifs de recherche et des besoins d'information des clients.

5. LA CLÉ – NORMALISATION[3] DES MÉTADONNÉES GRÂCE À UNE GESTION ET DES OUTILS D'ACCÈS COMMUNS

Comme nous l'avons exposé précédemment, les données et l'information pertinentes dans le cadre de l'analyse des politiques et des programmes proviennent rarement d'un organisme ou d'un programme statistique unique. Même au Canada, qu'on estime posséder un système statistique fortement centralisé, un bureau unique, tel que Statistique Canada, ne représente qu'un

---

[3] Par «normalisation des métadonnées», nous entendons la normalisation des éléments, des présentations lisibles par machines et des structures des dictionnaires de données, des tables de codage, des descripteurs d'études, des catalogues de données, du texte documentaire et des résultats de recherche. Au niveau des données, cette liste inclut des éléments tels que les étiquettes de variables, et de valeurs, les positions de zones, le texte des questions et les instructions «passez à», les descripteurs sommaires de l'univers géographique, la source des variables, le champ d'observation, les notes explicatives, les instructions données à l'intervieweur, les unités d'analyse, les populations cibles et le plan d'échantillonnage, ainsi que divers attributs destinés à l'identification et au contrôle des catalogues.
Nous *ne faisons pas* allusion à la normalisation des concepts et des définitions des mesures, tels les codes de la Classification internationale des maladies (CIM), ou les codes des industries et des professions, qui incombe à d'autres organismes. *Nous nous intéressons à la normalisation des conduits des systèmes et réseaux qui créent et fournissent les renseignements sur la santé, mais non de ce qui y circule.* Nous nous attaquons uniquement aux conteneurs lisibles par machine grâce auxquels lesdits codes, concepts et définitions sont transmis aux utilisateurs. Cependant, comme nous l'avons mentionné précédemment, nous sommes convaincus que la normalisation des conteneurs sous-jacents aura pour avantage d'accélérer considérablement le processus d'harmonisation des concepts et des définitions des divers programmes de statistique et d'information.

noeud dans le Réseau d'information sur la santé. De nombreuses autres sources existent, y compris les gouvernements provinciaux, les ONG, les universités, les organismes de santé publique et les organismes communautaires. Les données internationales comparables deviennent aussi de plus en plus importantes. La plus-value résultant de l'usage collectif s'inscrit partout, à tous les niveaux de la pyramide d'information.

Les analystes de programmes et de politiques de Santé Canada et d'autres ministères doivent souvent puiser à certaines de ces sources, ou à toutes, dans des délais très courts, pour comparer des questions d'enquête et d'autres paramètres méthodologiques importants, et pour faire rapidement une synthèse en vue de répondre aux exigences du moment en matière de séances d'information ou de consultation. L'accès par une fenêtre unique à *toutes* les sources, grâce à des métadonnées normalisées figurant dans les entrepôts des nombreux organismes concernés, est un facteur essentiel. L'absence d'une telle normalisation constitue un obstacle considérable qui empêche de fournir efficacement des renseignements aux fins de l'élaboration de politiques et de prises de décisions relatives aux programmes. Les utilisateurs finals des ministères et d'ailleurs ont tout intérêt à s'assurer que cette normalisation ait lieu et qu'elle réponde à leurs besoins et à leurs exigences.

Bradley et ses collaborateurs (1994) montrent la puissance des métadonnées normalisées en tant qu'outil d'accès aux ressources interfonctionnelles stockées à un site unique et de synthèse de ces dernières. Les nouvelles technologies de l'information permettent d'étendre la normalisation à tous les partenaires pertinents, qu'il s'agisse de programmes, de données ou d'information, de sorte qu'il devient possible de visionner, de manipuler, d'analyser et de synthétiser le contenu de tous les entrepôts de façon transparente, indépendamment de la source, de l'emplacement, de la plate-forme de traitement ou de l'autorité compétente.

Du point de vue technique, on effectue la normalisation en fournissant un ensemble commun d'outils experts de création, de gestion et de consultation de métadonnées et de développement de connaissances. Les outils sont fondés sur les meilleures normes existantes et sur les pratiques généralement acceptées.

Bradley et ses collaborateurs (1994) décrivent les exigences fonctionnelles précises auxquelles doivent répondre tant les métadonnées nécessaires que les outils connexes de gestion, de normalisation et de consultation. Nombre des principes et méthodes sous-jacents sont déjà approuvés en ce sens qu'ils ont été mis en oeuvre dans des logiciels que Santé Canada et la plupart de ses fournisseurs de données utilisent depuis pratiquement une décennie (Bradley et coll., 1990, 1991, 1992). En voici la liste.

- ***Documentation à la source :*** Les métadonnées devraient être créées en une fois, convenablement, sous forme normalisée, à la source. Ni les organismes fournissant des services aux utilisateurs de données ni leurs clients n'ont les ressources nécessaires pour préparer ou réviser les dictionnaires de données et la documentation sur ces dernières en vue de permettre la consultation, l'utilisation et la comparaison de toutes les données et métadonnées - les données sont tout simplement trop nombreuses et la préparation de la documentation est une tâche coûteuse, à haute intensité de main-d'oeuvre. Comme l'indique Sundgren (1993), les éléments de métadonnées ne devraient jamais être saisis plus d'une fois. La documentation doit être préparée convenablement, à la source, par les personnes chargées, au départ, de collecter les données[4].
- ***S'appuyer sur les meilleures normes existantes :*** S'appuyer sur les meilleures normes existantes et sur les pratiques généralement reconnues, particulièrement les meilleures pratiques des gros producteurs de données. Renforcer et appuyer l'élaboration systématique de normes pratiques, ainsi que la convergence des normes et des pratiques.
- ***Outils logiciels et métadonnées portables :*** Les outils logiciels destinés à la création, à la vérification et à la gestion des métadonnées ainsi qu'à l'ajout de valeur à ces dernières doivent être transférables à tous les producteurs et utilisateurs de données, et utilisés par ces derniers. Similairement,

---

[4] Nous estimons que le coût du chevauchement des tâches, sous forme de restructuration et de nouvelle saisie des dictionnaires de données aux fins de leur application à divers progiciels statistiques, pourrait s'élever jusqu'a 250 années-personnes par an au Canada, selon le nombre d'ensembles de données diffusés dans le domaine public et le nombre d'utilisateurs qui les installent (Bradley et coll., 1994). Cette estimation ne tient pas compte du temps que les chercheurs consacrent à la recherche de renseignements sur les sources de données qui les intéressent, à l'obtention de toute la documentation pertinente et à la production de nouveaux tableaux et produits d'information de base, ni de l'éventuelle production en double de métadonnées à divers stades par ceux qui collectent et traitent les données en premier lieu.

il faut établir l'accès universel aux outils permettant de rechercher des éléments de données et d'information pertinents, d'extraire et d'analyser des données et de créer, de gérer et de consulter les produits d'information et de connaissances connexes, et rendre l'utilisation de ces outils universelle.

- ***Ajouter de la valeur aux processus cloisonnés :*** Du point de vue des gestionnaires de programme d'enquête, les outils doivent augmenter la valeur des activités de diffusion des données et de l'information au sein de chaque unité cloisonnée, dans la mesure du possible. La normalisation entre programmes est effectuée à titre d'activité secondaire.
- ***Débuter avec les microdonnées :*** Commencer par créer des outils pour décrire les microdonnées[5]. Les agrégats, quant à eux, sont entièrement décrits par les métadonnées correspondant aux microdonnées sous-jacentes, conjuguées aux renseignements sur la ou les méthodes de calcul des agrégats. Il n'est pas possible de comprendre ce qu'est un agrégat sans se faire, préalablement, une bonne idée des unités d'analyse sous-jacentes, ainsi que des observations effectuées ou des mesures prises sur ces unités. Les métadonnées et les systèmes qui s'articulent uniquement sur les agrégats risquent d'être définis incorrectement si on ne les construit pas en s'appuyant sur ceux établis pour les microdonnées sous-jacentes.
- ***Dictionnaires génériques de données :*** S'appuyer sur des structures génériques de dictionnaire de données. Les dictionnaires de données sous-jacents devraient être conçus pour être exploités avec un aussi grand nombre que possible de progiciels d'analyse statistique et de traitement de l'information passés, présents et futurs. Dans la mesure où cela peut être accompli, la métabase résultante servira de plate-forme d'accès aux données et à l'information, indépendamment du nombre de progiciels distincts susceptibles d'être utilisés pour accomplir des activités individuelles de gestion des données et de développement de l'information.
- ***Table de codage élargie :*** Intégrer toute la documentation se rapportant à chaque variable figurant dans l'enregistrement de microdonnées aux éléments du dictionnaire générique de données correspondants à la variable en question. On obtient ainsi ce que Doyle (1994) appelle une «table de codage élargie».

  La table de codage élargie contient non seulement les libellés des collections de programmes et des renseignements sur les clichés d'enregistrement, sur les valeurs manquantes et sur les présentations qui peuvent être utilisés par toutes les plates-formes de traitement, c'est-à-dire les éléments des dictionnaires génériques de données, mais aussi les énoncés des questions et les notes concernant les sources, les fréquences d'échantillonnage et les fréquences pondérées pour toutes les catégories de réponse, les notes concernant la couverture des sous-populations, les notes explicatives, les instructions à l'intention des interviewers, les règles de vérification et d'imputation, les pointeurs logiques des opérations «passez à» et d'autres éléments documentaires dont un utilisateur pourrait avoir besoin pour se servir de la variable. Dans la mesure où cela est réalisé, on obtient, au niveau de la variable, un schéma documentaire utilisable par tous les producteurs d'information et de connaissances, dans tous les entrepôts d'information et sur toutes les plates-formes de traitement du monde.
- ***Catalogue et description d'étude génériques :*** Correspondent à la couche suivante d'éléments structurels de documentation qui se rapportent aux données prises comme un tout. Du point de vue des normes et de la structure, les plus critiques de ces éléments sont ceux que les spécialistes de la bibliothéconomie et des sciences de l'information qualifient de «bibliographiques», «catalogues» ou «descriptions d'étude». Ces éléments fournissent des renseignements d'identification et de contrôle bibliographique sur l'ensemble de données, et résument le contenu et les propriétés scientifiques essentielles de cet ensemble.

  La méthode adoptée par Santé Canada consiste à choisir un groupe de descripteurs génériques communs à tous les systèmes bibliographiques et de description d'étude importants en usage, de façon à pouvoir mettre les éléments en correspondance avec

---

[5] Par microdonnées, nous entendons les données se rapportant à la plus petite unité d'analyse ou, en termes statistiques fondamentaux, à ce qui est dénombré. Dans le cas d'une enquête, les microdonnées correspondent aux enregistrements contenant les observations sur chaque répondant. Par agrégat, nous entendons une mesure ou estimation sommaire, telle qu'un dénombrement, un total, une moyenne, une variance ou un pourcentage. Dans le contexte des microdonnées, certains voient les agrégats comme des «macrodonnées» (Creecy et coll., 1994).

tous ces systèmes (Ruus, 1992). Ces derniers incluent les systèmes Marc canadien, américain et britannique, le Standard Study Description (SSD) européen et les systèmes utilisés par la NASA et par l'Interuniversity Consortium for Political and Social Research (ICPSR).

- ***Création de liens avec les textes documentaires :*** Les éléments de dictionnaire, de table de codage et de catalogue susmentionnés sont communs à la plupart des microdonnées, constituent une trousse de documentation minimale, mais adéquate, pour bon nombre d'applications analytiques courantes, et peuvent être définis et structurés de façon relativement simple et normalisée. Toutefois, il existe aussi un grand nombre de textes documentaires, tels que les descriptions contextuelles des études ou des programmes, les études bibliographiques et les justifications, les articles traitant de questions méthodologiques, les instruments d'enquête, les études ou les rapports spéciaux, les résumés de traitement et les graphiques de cheminement, les raisons justifiant le choix des questions ou des échelles, les manuels de codage, ainsi que les fichiers et la correspondance connexes. Ce genre de documentation doit également être structuré, intégré au catalogue et à la table de codage, et produit sous forme électronique portable pour faciliter sa consultation au niveau de l'étude ou de l'ensemble de données.
- ***Structure des bases de données relationnelles et documentaires :*** Santé Canada a utilisé avec beaucoup de succès des structures xBASE pseudo-relationnelles portables pour les volets «dictionnaire», «table de codage» et «catalogue» de la structure. Les nouvelles technologies de traitement de textes, particulièrement celles basées sur le langage standard généralisé de balisage (SGML) et sur son sous-ensemble Hypertext Markup Language (HTML), facilitent l'intégration, la normalisation et la structuration de tous les éléments de documentation, l'amalgamation des produits d'information et de connaissances, et l'accès en direct à tous les entrepôts grâce à des navigateurs commercialisés à grande échelle.

6. INTÉGRATION DES PRODUITS D'INFORMATION ET DE CONNAISSANCES À LA MÉTABASE

Selon nous, en plus de fournir une documentation complète et structurée, normalisée pour tous les ensembles de données dans tous les entrepôts, ainsi que les outils connexes de création, de gestion, de consultation et d'extraction, il est très important d'intégrer les produits d'information et de connaissances à la métabase. Les utilisateurs recherchent en premier lieu des connaissances, plutôt que des données et des statistiques. Les métadonnées, telles que nous les concevons, sont formées non seulement d'information et de connaissances *au sujet* des données et de l'information, mais aussi d'information et de connaissances *tirées* des données et de l'information.

Cette clause de notre définition des métadonnées tient à des considérations purement pragmatiques - *ayant trait à ce qui est nécessaire et à ce qui permet de fournir ce nécessaire*. Voici nos arguments à l'appui de cette position.

Hicks (1995) insiste sur la nécessité de dégager des «tendances de longue durée» d'un grand nombre de sources de données distinctes pour étayer l'élaboration des politiques sociales. À titre d'exemple, examinons le produit de métadonnées illustré à la figure 3, qui montre une «tendance de longue durée» importante ayant trait à la gestion du risque dans le domaine de la politique en matière de santé. En paraphrasant Bradley et ses collaborateurs (1994, p. 30), nous disons que *ce produit est tiré entièrement de l'information qui figure dans ... la métabase de Santé Canada... Tous les éléments, sauf l'énoncé interprétatif correspondant au deuxième point vignette, ont été extraits de la métabase, analysés et synthétisés comme un tout cohérent pour leur donner une signification, puis restructurés aux fins de leur présentation.*

Les métadonnées - éléments de documentation des données - sont les composantes élémentaires de l'information et des connaissances.

La figure 3 est un exemple du type d'information que nous appelons «bribe d'information» - fiches signalétiques et autres produits autonomes, réutilisables, mettant l'accent sur des éléments d'information relativement distincts et simples. Une bribe d'information typique contient les réponses à une seule question d'enquête pour diverses sous-populations ou d'autres caractéristiques étudiées, ou les réponses à une même question posée à l'occasion d'un grand nombre d'enquêtes distinctes, mais comparables, à divers points dans le temps.

Retournant à l'article de 1994, ... *de tels produits peuvent être stockés électroniquement sous forme intégrée avec les métadonnées, parcourus par des moyens automatisés et extraits, examinés et imprimés sélectivement aux fins d'être intégrés, à titre de justification, aux notes de synthèse, aux résumés*

d'étude et aux présentations.

Les produits de connaissances résultants peuvent, à leur tour, être stockés dans la métabase, d'où on peut les extraire en vue de les adapter à de nouvelles exigences, ou d'examiner minutieusement l'information et les données de soutien.

En suivant le modèle de la figure 1, *il est alors possible de retourner en arrière, des connaissances à l'information de base, et de chaque point de données agrégées dans un produit d'information à la documentation des microdonnées sous-jacentes. ... il sera possible de cliquer sur un point de donnée ou sur un énoncé interprétatif pour faire apparaître toute l'information et la documentation connexes. Quand les connaissances ou l'information existantes laissent à désirer ou ne satisfont pas les exigences du moment, le retour en arrière peut se faire jusqu'aux données elles-mêmes afin de produire de nouveaux agrégats et produits d'information, ou d'examiner le comportement des agrégats existants à la lumière d'autres considérations théoriques et analytiques.*

La figure 3 illustre aussi une des propriétés les plus importantes des métadonnées normalisées... nommément faciliter l'intégration, en un tout cohérent, des données et de l'information provenant de sources très diverses. Chaque point de données... provient d'une enquête distincte, les diverses enquêtes ayant été effectuées par de nombreux organismes appartenant à différents niveaux d'administration, ainsi que par des ONG et des entreprises du secteur privé. Comme, à l'époque, certains de ces organismes avaient le sentiment d'être en concurrence avec d'autres et que, par ailleurs, certains n'étaient pour ainsi dire pas au courant des travaux des autres, les données résultantes présentent de nombreuses différences frustrantes tenant à la méthodologie et à l'énoncé des questions, différences qui rendent les comparaisons et l'intégration difficiles... Néanmoins, la documentation normalisée permet de placer et de manipuler les métadonnées et les agrégats dans un même contenant logique, et d'examiner les différences de façon systématique.

Donc, le fait qu'elle fournisse des outils qui facilitent la méta-analyse recoupant les divers sites d'entreposage est peut-être l'avantage le plus important de la normalisation des métadonnées. Comme l'ont montré Bradley et ses collaborateurs(1994), de tels outils augmentent considérablement la capacité qu'ont les spécialistes de l'information de fabriquer et de gérer des richesses qui prennent la forme de nouveaux produits d'information servant répétitivement à un large éventail d'applications de développement d'information et de connaissances. Il est par conséquent fort sensé d'intégrer l'information et les connaissances aux données sources sous-jacentes au moyen de métadonnées, et de concevoir ces dernières en conséquence.

Beaucoup reste à faire pour parachever notre compréhension des processus de développement des connaissances, mais, d'emblée, nous distinguons trois types fondamentaux d'éléments de métadonnées dans les catégories «information» et «connaissances» :

- ***Agrégats totalisés types :*** De nouveau, nous paraphrasons l'article de 1994. *Les points de données de la* figure 3 *sont des estimations simples, univariées. ... Il est également nécessaire de ventiler la consommation de tabac au cours du temps selon le sexe et selon l'âge et, compte tenu des modifications récentes de la politique de taxation, tout spécialement selon la province. ... l'ajout d'agrégats totalisés types selon l'âge, le sexe, la région, la province, le revenu, le niveau de scolarité et ainsi de suite représente une extension aisée de la métabase. Il en est de même de la capacité d'**automatiser** la production de ces estimations et d'autres à partir des bases de données distinctes d'où il faut calculer chaque point de données, indépendamment de la source de la base de données ou du milieu de traitement dans lequel elle est mise en oeuvre.*

- ***Bribes d'information :*** L'expérience acquise par Santé Canada après avoir préparé et diffusé une collection d'environ 1 000 bribes d'information donne à penser que ces produits très simples, réutilisables, semblables à des fiches signalétiques, ainsi que les outils logiciels destinés à les créer, à les gérer, à les stocker, à les manipuler et à les extraire, seront parmi les composantes les plus demandées et les plus précieuses de l'entrepôt. Notre expérience laisse en outre entendre que les bribes d'information seront produites dans des conditions quasi industrielles, à partir d'autres éléments de métabase, et sélectionnées par les clients de façon automatisée aux fins de leur utilisation dans des résumés d'étude, des notes de synthèse et d'autres applications de la production de connaissances. Désormais, on peut aussi envisager de se servir de clips audio et vidéo, comme supports normalisés des bribes d'information.

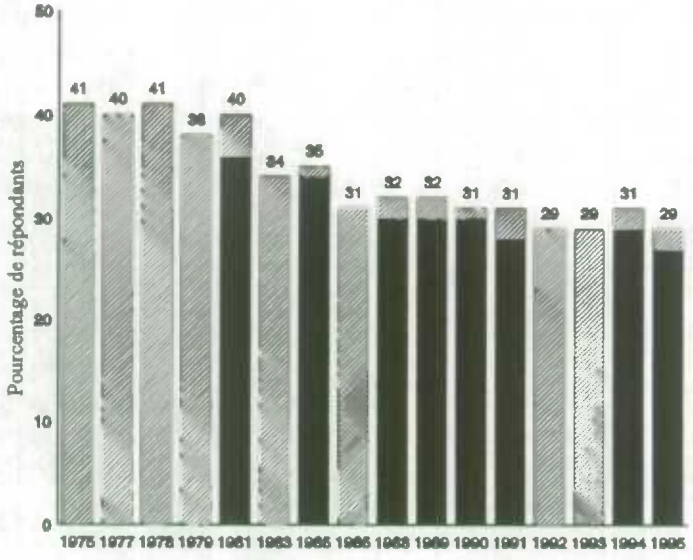
Figure 3. FUMEURS DE CIGARETTE

- *Des questions sur la consommation de cigarettes ont été posées dans le cadre des Enquêtes sur les habitudes de fumer de 1975, 1977, 1979, 1981, 1983 et 1986 de l'Enquête sur la population active, de l'Enquête santé Canada de 1978, de l'Enquête condition physique Canada de 1981, de l'Enquête sociale générale de 1985 et de 1991, de l'Enquête promotion santé Canada de 1985 et de 1990, de l'Enquête Campbell sur le mieux-être au Canada de 1988, de l'Enquête nationale sur la consommation d'alcool et de drogues de 1989, du Sondage Santé Canada de 1988-1995 et de l'Enquête sur le tabagisme au Canada (cycles 1 et 4) de 1994-1995. On a, par exemple, posé la question :*

  À l'heure actuelle, fumez-vous des cigarettes?

- *Les résultats de toutes ces enquêtes donnent à penser que la proportion de fumeurs de cigarette a diminué considérablement au Canada depuis 1978. Le mouvement de baisse semble s'être stabilisé depuis le milieu des années 80.*

À l'heure actuelle, fumez-vous des cigarettes?
1975-1995



Les pourcentages représentent les membres de la population de 15 ans et plus qui ont répondu «oui» à la question posée. Ils incluent les fumeurs de cigarette réguliers et occasionnels.

Enquêtes effectuées la même année : 40% ECPC81, 36% EHF81; 35% ESG85, 34% EPS85; 32% SSC88, 30% ECMEC88; 31% SSCE90, 30% EPS90; 31% ESG91, 28% SSCE91; 29% SSCE94, 31% ETC94C1; 29% SSCE95, 27% ETC94C4.

Source : Métabase du SGDD : EHF75, EHF77, ESC78, EHF79, ECPC81, EHF81, EHF83, ESG85, EPH85, EHF86, SCC88, ECMEC88, SSCH89, ENCAD89, SSCE90, EPS90, SSCH91, ESG91, SSCE92, SSCE93, ETC94C1, SSCE94, ETC94C4, SSCE95.

- ***Produits d'information et de connaissances de haut niveau :*** D'après le modèle de la figure 1, les produits «de haut niveau» ou produits «de connaissances», englobent les publications imprimées traditionnelles de toutes sortes, les notes de synthèse, les documents de recherche, les cours, les discours et les présentations. Les produits de haut niveau se distinguent des produits de «bas niveau», ou «produits d'information», par le nombre de faits ou «bribes d'informations» qui sont synthétisés, par la portée et la profondeur des thèmes étudiés, par le niveau d'analyse, et par la présentation et la structuration sélectives de l'information, conformément à des fondements théoriques ou en vue de se concentrer sur une exigence, une tâche ou une application particulière. Il s'agit ordinairement de produits finals à part entière, souvent non réutilisables comme un tout à d'autres fins[6].

Formés principalement de texte et de graphiques, les produits de connaissances traditionnels peuvent être intégrés assez facilement à la métabase grâce aux nouveaux outils hypertextes de structuration et de normalisation, particulièrement SGML, HTML et les éditeurs et les navigateurs connexes. Pour le moment, on se concentrera de toute évidence sur les publications imprimées traditionnelles, mais le fait qu'il devient possible d'utiliser les technologies audio et vidéo à tous les niveaux de la pyramide d'information laisse entendre qu'on pourra continuer à améliorer le rendement et les délais de production, et à faire baisser les coûts.

Les figures 1 et 2 donnent à penser qu'au plan de la structure, on devrait avant tout s'assurer qu'il soit possible de creuser dans les niveaux inférieurs d'informations et d'estimations jusqu'aux métadonnées et aux données sous-jacentes, et de remonter du niveau des variables jusqu'à ceux des produits d'information et de connaissances élaborés à partir de ces variables.

La capacité de creuser permettra aux spécialistes de l'information d'évaluer les hypothèses et les faits sur lesquels se fondent les rapports et les recommandations, ouvrira des voies qui les mèneront à des sources et à des fournisseurs pertinents d'information et de données en regard d'exigences connexes, mais différentes, et leur permettra d'effectuer rapidement de nouvelles analyses tenant compte d'autres approches analytiques et théoriques.

La capacité de remonter des variables aux connaissances permettra aux utilisateurs de données de visualiser des relations et des résultats analytiques connus, et de découvrir des faits et des connaissances préexistants présentant un intérêt pour leurs clients ou leurs collègues. La capacité de remonter à travers la structure permettra aussi aux planificateurs des données de se faire une idée plus complète qu'à l'heure actuelle de la signification et de l'exactitude des mesures ainsi que des méthodes de mesure, et de mieux évaluer les lacunes, les priorités quant aux mesures et les possibilités de mieux harmoniser les divers programmes de collecte de données et les diverses études.

7. DONNÉES ET INFORMATION PROVENANT DE SOURCES ADMINISTRATIVES

L'expérience de Santé Canada en ce qui concerne les métadonnées émane principalement de la gestion des microdonnées provenant de la collecte de données d'enquête. Toutefois, les données dérivées des processus administratifs fournissent aussi des renseignements importants. En ce qui concerne la mise sur pied du Réseau d'information sur la santé, une des priorités consiste à adapter et à appliquer aux données administratives des méthodes similaires de gestion des métadonnées, afin de pouvoir consulter et manipuler les données sur les résultats, en utilisant la fenêtre qui donne aussi accès aux données sur les facteurs de risque, les déterminants de la santé et les problèmes de santé autodéclarés.

Une de nos sources les plus importantes de données administratives est la Base de données sur la lutte contre les maladies (BDLCM). La BDLCM offre aux analystes de Santé Canada une source primaire de renseignements

---

[6] Le produit de connaissances pris dans son ensemble ne convient ordinairement pas à d'autres fins, mais les faits et les interprétations qui le composent sont fréquemment réutilisés dans d'autres contextes, avec des citations appropriées, s'ils répondent à d'autres exigences de développement de connaissances. D'où la nécessité de restructurer ces composantes en tant que «bribes d'information» distinctes et réutilisables. À l'heure moderne de l'électronique, il n'y a peut-être pas de plus grand gaspillage qu'un gros recueil imprimé de tableaux statistiques ni de situation plus frustrante que d'être obligé de taper à nouveau ou de copier par un autre moyen un tableau et ses énoncés interprétatifs afin de les utiliser dans un nouveau contexte, si on a eu la chance, avant toute chose, de découvrir un tel trésor et d'y avoir eu accès.

sur la mortalité, sur le cancer et sur la morbidité hospitalière. Elle a pour élément central une collection gigantesque de microdonnées extraites des fiches de radiation des hôpitaux et des certificats provinciaux d'enregistrement de décès par les gouvernements provinciaux, par Statistique Canada et par l'Institut canadien d'information sur la santé (ICIS). Le volume important de microdonnées sous-jacentes constitue un recensement des décès et des maladies survenus au Canada au cours d'un très grand nombre d'années. À l'instar des méthodes d'accès aux données d'enquête, le système de la BDLCM permet aux utilisateurs de données d'effectuer des extractions et des totalisations spéciales à partir des microdonnées. Cependant, le nombre très élevé d'enregistrements et relativement faible de variables que contient cette base de données favorisent un modèle d'accès légèrement différent. Au lieu d'effectuer des passages fréquents de microdonnées sur machine pour produire des totalisations spéciales, la stratégie modale d'accès consiste à intégrer des ensembles d'estimations prétotalisées à la base de données, en se servant de groupements prédéterminés de données démographiques, de périodes et de maladies. Les utilisateurs peuvent alors afficher et manipuler les estimations grâce au langage d'interrogation structurée (SQL).

Conformément aux principes et aux méthodes susmentionnés, nous proposons d'intégrer la Base de données sur le contrôle des maladies et les données d'enquête grâce à des structures communes de métadonnées. On peut, en effet, décrire les microdonnées sur la mortalité et sur la morbidité, et les intégrer au logiciel existant de consultation des données d'enquête en appliquant des méthodes courantes. Au moment de l'ajout de descripteurs et d'estimations totalisés types à la base de données d'enquête, nous veillerons à ce qu'il soit également possible d'appliquer systématiquement les structures aux tableaux de données tirées de sources administratives.

Les structures des dictionnaires génériques de données déjà établies dans le cadre de la méthode existante de gestion des métadonnées permettront de produire des énoncés en langage d'interrogation structurée (SQL) afin d'automatiser l'accès à chaque composante des fichiers bidimensionnels des structures relationnelles. Nous proposons d'élargir le dictionnaire normalisé de données afin d'y ajouter des descripteurs génériques de schéma pour les structures de données complexes. Cette dernière mesure sera bientôt nécessaire non seulement pour décrire les bases de données administratives telles que la BDLCM, mais aussi la nouvelle vague de données complexes provenant des enquêtes longitudinales effectuées au Canada. Les premières données longitudinales tirées de l'Enquête nationale sur la santé de la population (ENSP), de l'Enquête sur la dynamique du travail et du revenu (EDTR) et de l'Enquête longitudinale nationale sur les enfants (ELNE) apparaîtront en 1997 et en 1998.

En procédant comme nous le proposons, on élargira la portée de l'entrepôt virtuel qui passera des simples fichiers de microdonnées bidimensionnels issus des processus administratifs, lesquels peuvent être mis en oeuvre pour ainsi dire immédiatement, aux fichiers et aux agrégats complexes provenant tant des enquêtes que des sources administratives.

8. PLAN DE MISE EN OEUVRE : DÉVELOPPEMENT GRADUEL BASÉ SUR LES STRUCTURES EXISTANTES

Comme l'ont exposé Bradley et ses collaborateurs(1990 - 1994), depuis de nombreuses années, Santé Canada et ses principaux fournisseurs de données d'enquête ont mis au point des métadonnées normalisées, conformément aux principes et aux exigences résumés plus haut. Santé Canada a élaboré les premiers dictionnaires de données lisibles par machine au début des années 70 et les a mis en oeuvre de façon normalisée dans des réseaux nationaux exploités en temps partagé en 1975. La conception de la présente activité a eu lieu au début des années 80, à l'époque où s'est amorcée la révolution engendrée par les ordinateurs de bureau, et le démarrage a eu lieu en 1985, quand il est devenu manifeste que le traitement décentralisé des données et l'intégration grâce à des réseaux locaux et de grands réseaux seraient la voie de l'avenir.

Au milieu des années 80, Santé Canada a élaboré un outil de création, de gestion et de restructuration des métadonnées appelé SGDD, pour *Système de gestion de la documentation des données*, que les employés ont appliqué aux ensembles de données d'enquête existants à l'époque. Pour satisfaire le critère de portabilité, le SGDD a d'abord été développé au moyen de l'application dBASE du pseudo-système de gestion de base de données relationnelles (SGBDR), omniprésent à l'époque, puis converti au système CLIPPER '86 qui offrait des modules de chargement portables. Dès le début des années 90, les exigences de l'application ayant étiré la technologie de CLIPPER '86 à la limite de ses capacités, il a fallu créer un système supplémentaire pour ajouter le texte des questions et les descripteurs de catalogue à la métabase.

La métabase actuelle décrit environ 250 ensembles

de données d'enquête englobant la plupart des enquêtes importantes effectuées au Canada depuis le début des années 70 sur la santé, les facteurs sociaux, les facteurs démographiques, le revenu et les dépenses. Le logiciel contient une table de codage élargie conformément aux meilleures normes établies durant les années 80 et au début des années 90, est capable de restructurer ses dictionnaires génériques sous forme de programmes de définition des données applicables à la plupart des progiciels statistiques, est utilisé à la source par bon nombre de fournisseurs de données de Santé Canada appartenant au secteur aussi bien public que privé, et a facilité les activités de diffusion de données des universités du Canada sous les auspices de consortiums d'achat de données parrainés par l'Association des bibliothèques de recherche du Canada (ABRC) et par l'Initiative de démocratisation des données que vient de mettre sur pied le Conseil de recherche en sciences humaines (CRSH).

La Base de données sur le contrôle des maladies est un produit achevé, ayant ORACLE pour plate-forme. À l'heure actuelle, on planifie l'intégration d'autres ensembles nationaux de données.

9. BASE DE DONNÉES DE RENSEIGNEMENT SUR LA SANTÉ DE LA POPULATION

La création d'une base de données virtuelle de renseignement sur la santé de la population englobant des données d'enquête aussi bien que des données administratives et des renseignements destinés à Santé Canada et à ses partenaires, est un des objectifs prioritaires de l'équipe chargée de la mise en place du Réseau d'information sur la santé. Les travaux se dérouleront en deux phases.

Durant la première, qui sera exécutée immédiatement, le SGDD et le SID, pour *Système d'information et de données*, existants seront installés dans le réseau d'entreprises de Santé Canada afin que tous les analystes puissent les utiliser aux fins de production. Ces nouvelles versions des systèmes profiteront aux employés du ministère qui sont chargés des opérations internes et qui doivent prendre des décisions fondées sur des faits, ainsi qu'à de nombreux fournisseurs de données.

Durant la deuxième phase, qui débutera en 1996, la normalisation, l'accès virtuel, l'information et les agrégats intégrés, les outils de développement de connaissances et les méthodes d'échange de données enchâssés dans le modèle SGDD/SID seront appliqués et intégrés à la Base de données sur le contrôle des maladies dans une nouvelle exécution Windows, qui permettra de travailler aussi dans les univers Unix et World Wide Web en vue de partager un entrepôt virtuel avec des partenaires dans Internet.

10. CONCLUSION - FLUX DE DONNÉES, UTOPIE ET LE RÉSEAU DE TRAMWAYS ÉLECTRIQUES D'OTTAWA

Si le présent symposium avait eu lieu cinquante ans plus tôt ou avant, la plupart des participants ne résidant pas à Ottawa seraient vraisemblablement arrivés par chemin de fer et auraient loué une chambre au vénérable hôtel Château Laurier, situé à côté de la colline du Parlement, en face de l'ancienne gare d'Ottawa. Pour se déplacer de l'hôtel au lieu du symposium, ils auraient probablement emprunté le réseau de tramways électriques d'Ottawa, formé de voitures électriques appelées tramways circulant sur rails, qui constituait le principal moyen de transport public à Ottawa des années 1890 aux années 1950. Avant l'électrification du réseau, les voitures étaient tirées par des attelages de chevaux.

Un certain nombre des participants au symposium d'aujourd'hui logent aussi à l'hôtel Château Laurier. Essayez d'imaginer ce que cela aurait représenté pour eux de venir à la réunion d'aujourd'hui et d'en repartir si chaque bloc du réseau de transport public était contrôlé par un propriétaire distinct et relativement autonome, chacun de ces propriétaires ayant une vision légèrement différente des choses ou des personnes qui devraient être transportées, où et pourquoi, ou de la valeur du service, et s'il existait peu de normes, voire aucune, concernant la dimension des rails et le matériel roulant et peu de routes, de cartes, d'horaires et de tarifs communs.

Dans de telles circonstances, il faudrait vraisemblablement changer de tramway au moins une douzaine de fois pour emprunter un autre réseau. À l'entrée du territoire exploité par chaque nouveau transporteur, les passagers devraient probablement se familiariser avec les services offerts et avec les routes et les horaires locaux. Puis, ils devraient sans doute décider d'une route, faire la file et attendre la prochaine voiture, en espérant que la route choisie les mènera au prochain transporteur, à un point où existent des rails et des véhicules d'une sorte ou l'autre qui leur permettront de progresser dans la direction générale souhaitée.

Nous ne tolérerions pas longtemps une telle situation ni ne jugerions utopique un réseau harmonisé de transport à l'échelle de la ville, complètement interexploitable avec des transporteurs régionaux,

nationaux et internationaux. Fort heureusement, même à l'époque où les voitures étaient tirées par des chevaux, le réseau de transport public sur rail d'Ottawa ne souffrait pas d'un tel manque de normes et de normalisation. Alors, plus d'un siècle plus tard, pourquoi continuons-nous à tolérer de telles conditions en ce qui concerne nos systèmes nationaux et internationaux de renseignements statistiques?

En soulignant et en démontrant certains avantages qu'offre la normalisation de la structure et du contenu des métadonnées, nous décrivons une situation jugée ordinaire dans des domaines où la normalisation a déjà été réalisée pour des raisons d'ordre pratique évidentes et pour le bien commun. Nous proposons simplement de placer toutes les données statistiques et la valeur qui y est ajoutée sur une même voie, pour pouvoir les harmoniser et les transmettre plus efficacement et intelligemment aux clients, et pour que toutes les personnes concernées bénéficient rapidement des avantages qu'offre le passage des voitures tirées par des chevaux à l'autoroute électronique de l'information.

L'entrepôt virtuel d'information est techniquement réalisable, à peu de frais. Nous n'avons donné qu'un bref aperçu des avantages énormes que nous avons tous à tirer d'une collaboration en vue d'appuyer et de réaliser la normalisation requise. Comme dans le cas des machines de traitement de textes, des tabulateurs et des progiciels de traitement des données statistiques à grande diffusion, le syndrome du «pas inventé ici» n'est désormais plus acceptable.

De nombreux travaux restent à accomplir au plan technique, mais les questions essentielles n'appartiennent plus à ce domaine. Désormais, on se soucie surtout de savoir si les intervenants de tous les secteurs auront la capacité de créer et de partager des infrastructures communes et d'élaborer des politiques et des procédures viables pour gérer, évaluer et échanger les produits et les ressources sous-jacents. En ce qui concerne la normalisation requise, rien ne donnera peut-être un meilleur élan initial que la nécessité d'améliorer la surveillance et la protection de la santé et du bien-être de l'homme, mais, une fois cet élan donné, la poursuite du projet dépendra de l'appui et de la collaboration soutenue des partenaires.

Selon notre expérience, le partage de la propriété et de l'administration des ressources et des activités entre les partenaires, et la mise en place de politiques d'information et de structures administratives durables sont des ingrédients indispensables. Nous invitons tous les intervenants à se joindre à l'effort.

BIBLIOGRAPHIE

Alexander, C. J. (1990). Towards the information edge and beyond: enhancing the value of information in public agencies, Rapport soumis à Justice Canada, Department of Political Science, University of Western Ontario, London, Ontario.

ANSI X3L8 (1995). Voir, par exemple, Gillman, D.V. (Ed), ISO/IEC STANDARD 11179 Specification and Standardization of Data Elements, (1995).

Balcer, M. (1992). Group facilitator presentation to final plenary, National Summit on Information Policy, Ottawa, Canada.

Beck, N. (1992). Shifting gears: thriving in the new economy, Harper Collins, Toronto.

Berk, K., et Ryan, T.(1992). Report from the Share File Committee, Statistical Computing and Graphics Newsletter, *American Statistical Association*.

Boucher, L. (1995). Connaître les produits et services de Statistique Canada: la manière IPS, *Recueil du Symposium 95: Des données à l'information - méthodes et systèmes*, 217-220.

Bradley, W. J., Diguer, J., et Ellis, R. J. E. (1990). Methods for producing interchangeable data dictionaries and documentation, papier préparé pour la réunion International Association for Social Science Information Service and Technology (IASSIST), Poughkeepsie, N.Y.

Bradley, W. J., Diguer, J., Ellis, R. K., et Ruus, L. (1991). DDMS: A PC-Based package for managing social science data dictionaries and documentation - introduction and basic functions, 7th Draft Edition, Social Environment Information, Information Systems Directorate, Santé et Bien-être Canada.

Bradley, W. J. (1992). Developing information for policy decision making, présenté pour la réunion International Association for Social Science Information Service and Technology (IASSIST), Edmonton, Alberta.

Bradley, W. J., Diguer, J., Ellis, R. K., et Ruus, L. (1992). DDMS: A PC-Based package for managing social science data dictionaries and documentation - document de référence, 12ième Edition ébauche, Social Environment Information, Policy, Planning and Information Branch, Santé et Bien-être Canada.

Bradley, W. J., Diguer, J., et Touckley, L. (1992). DDMSS: data dictionary management system supplemental - user's guide, 1st Draft Edition, Social Environment Information, Policy, Planning and Information Branch, Santé et Bien-être Canada.

Bradley, W. J., Hum, J., et Khosla, P. (1994). Metadata matters: standardizing metadata for improved management and delivery in national information systems - Parts 1 - 3, Bureau of Surveillance and Field Epidemiology, Laboratory Centre for Disease Control, Santé Canada.

Capps, C. (1995). FERRET Overview - A Federal electronic research and review extraction tool for data access and dissemination of micro and macro data, Survey Modernization Programming Branch, Demographic Surveys Division, Bureau of the Census.

Cohen, D. (1993). No small change: succeeding in Canada's new economy, Macmillan Canada, Toronto.

Colledge, M., et Richter, W. (1994). La gestion des données et la banque d'informations: pour une infrastructure de restructuration, *Recueil du Symposium '94: Restructuration pour les organismes de statistique*, 197-212.

Creecy, R.H., Gillman, D. W., et Appel, M.V. ta, (1994). Metadata, statistical software and information systems at the U.S. Census Bureau: current practice and future plans, U.S. Bureau of the Census, Washington D.C.

Deecker, G., Murray, T. S., et Ellison, J. (1993). On providing client support for machine readable data files, Household Surveys Division, Statistics Canada, Ottawa.

Dodd, S.A. (1982). Cataloguing machine-readable data files: an interpretive manual, American Library Association, Chicago.

Doyle, P. (1994). Modernizing data documentation, Agency for Health Care Policy, Presented at IASSIST '94, San Francisco.

Doyle, P. (1994). User documentation for CASIC systems, agency for health care policy and research, presented to APDU '94, Washington D.C.

Fierheller, G. (1992). Growing the infratechnology, Theme Presentation to the National Summit on Information Policy, Ottawa, Canada.

Gillman, D.W. (1994). Developing a metadata database at the Census Bureau, Statistical Research Division, Bureau of the Census, Washington D.C.

Gorman, B., Silverman, A., et McLure, J. (1992). Council for administrative renewal, présenté à l' Expo Innovation, Government of Canada Canadian Centre for Management Development, Ottawa, Canada.

Grenier, R. (1995). Meilleur service électronique à la clientèle: le projet StatCan en Direct, *Recueil du Symposium '95: Des données à l'information - méthodes et systèmes*, 235-236.

Hammer, M., et Champy, J. (1993). Reengineering the corporation: a manifesto for business revolution, Harper Collins, New York.

Hicks, P. (1992). Proposal for new data on skills and learning: social statistics for prosperity, economic union and fairness, policy, Planning and Information Branch, Health and Welfare Canada.

Hicks, P. (1995). Le rôle des statistiques dans l'élaboration des politiques sociales, *Recueil du Symposium '95: Des données à l'information - méthodes et systèmes*, 7-12.

Ludley, J. H. (1993). The use of meta data in the UK Central Statistical Office, Information Systems, Central Statistical Office, Great George Street, London SW1P 3AQ.

MacDonald, A. (1994). Blueprint for renewal in the public service of Canada, Treasury Board Secretariat, Ottawa.

Massé, M. (1993). Partners in the management of Canada: the changing roles of government and the public service, The 1993 John L. Manion Lecture, Canadian Centre for Management Development, Ottawa.

METIS90 (1990). Users' guide to meta-information systems in statistical offices, United Nations, Geneva.

Nordbotten, S. (1993). Statistical meta-knowledge and data, Invited opening lecture, Conference Proceedings, Statistical Meta Information Systems Workshop, EUROSTAT.

Priest, G. (1995). Intégration des données: le point de vue de ceux qu'on relègue à l'arrière de l'autobus, *Recueil du Symposium '95: Des données à l'information - méthodes et systèmes*, 15-22.

Roistacher, R. C. (1976). The data interchange file: A first report, Document No. 207, Center for Advanced Computation, University of Illinois, Urbana, Illinois 61801.

Roistacher, R. C. (1978). A style manual for machine-readable data files and their documentation, Draft2, Center for Advanced Computation, University of Illinois, Urbana, IL 61801.

Smith, S. (1992-93). Concluding plenary address and summary, in national information summit stimulates communication and proceedings: National Summit on Information Policy, Canadian Library Association and Association pour l'avancement des sciences et des techniques de la documentation, Ottawa, Canada.

Subcommittee on Cultural and Demographic Data, (1992) Content Standard for Cultural and Demographic Data Metadata, Federal Geographic Data Committee Secretariat. USGS MS 590 National Centre, 12201 Sunrise Valley Drive, Reston VA 22092.

Subcommittee on Electronic Dissemination of Statistical Data, Electronic dissemination of statistical data, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, Washington D.C., novembre 1995.

Sundgren, B. (1973). An infological approach to databases, Urval 7, Statistics Sweden, Stockholm.

Sundgren, B. (1993). Statistical metainformation systems - pragmatics, semantics, syntactics, Statistics Sweden, S-11581 Stockholm.

Sundgren, B., Gillman, D.W., and Appell, M.V. (1995). Towards a unified data and metadata system at the U.S. Bureau of the Census (BOC) (Preliminary Draft), United States Department of Commerce, Bureau of the Census, Washington D.C.

Tapscott, D., et Caston, A. (1993). Paradigm shift: the new promise of information technology, McGraw-Hill.

Zeisset, P. T. (1993). Meta-information for summary statistics: the EXTRACT experience, U.S Bureau of the Census, Economic Census Staff, Washington.

Silverman, A. (1993). Administrative renewal in the federal public service: the next generation, *Optimum*, 24-1.

Wilk, M. B. (1991). Health information for Canada 1991: Report of the National Task Force on Health Information, The National Health Information Council, Canadian Centre for Health Information, Statistics Canada.

National Task Force on Health Information project team on health policy information requirements, (1991). Health Policy Information Requirements, Canadian Centre for Health Information, Statistics Canada.

National Task Force on Health Information project team on the template, (1991). Development of a Structural Model (Template), Canadian Centre for Health Information, Statistics Canada.

National Task Force on Health Information project team health information analysis, (1991). *Health Information Analysis: Potentials and Impediments*, Canadian Centre for Health Information, Statistics Canada.

Wolfson, M. (1992). New electronic data products - experience in Statistics Canada, paper presented at 22nd General Conference of the International Association for Research in Income and Wealth, Films, Switzerland, Analytical Studies Branch, Statistics Canada.

Office of Technology Assessment (1989). U.S. Congress, Statistical Needs for a Changing U.S. Economy, Background Paper OTA-BP-58, Washington D.C.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

CONNAÎTRE LES PRODUITS ET SERVICES DE STATISTIQUE CANADA : LA MANIÈRE IPS

L. Boucher[1]

RÉSUMÉ

L'IPS (Information sur nos Produits et Services) est une application Windows de recherche et de récupération qui permet aux employés de Statistique Canada d'identifier l'information courante sur nos produits et services et de répondre aux demandes de nos clients. Conçu originalement pour le personnel des services consultatifs et de la bibliothèque de Statistique Canada, l'IPS fournit un accès unique à plusieurs sources de métadonnées et permet une recherche organisée et efficace. L'IPS est un outil innovateur qui facilite la tâche toujours plus exigeante de retrouver l'information pertinente pour notre clientèle à partir d'une gamme de produits et services en croissance.

L'IPS est un guichet unique : plus besoin de chercher à plusieurs endroits pour trouver quels produits ou services sont disponibles sur un sujet particulier. Tous les produits et services «enregistrés» s'y retrouvent. Les usagers peuvent interroger l'IPS à partir de mots, de titres, de sujets et d'auteurs, entre autres possibilités, et ce à travers de milliers de documents descriptifs. Suite à une recherche, l'IPS affiche une liste de produits et services dont l'usager peut utiliser l'information pour remplir la demande, l'acheminer au client ou même assister à la création de mini-catalogues.

MOTS CLÉS: Métadonnées; application de recherche et de récupération; enregistrement de produits; catalogue électronique.

1. INTRODUCTION

Au delà d'un demi-million de requêtes d'information sont traitées chaque année par les services consultatifs de Statistique Canada, et ce sans compter un nombre impressionnant de demandes adressées directement aux différents spécialistes en la matière. Cette tâche est de plus en plus contraignante et difficile. Les usagers et les clients de Statistique Canada ont des besoins qui vont des données spécifiques aux analyses socio-économiques génériques alors que la gamme de produits et services offerts est en constante évolution.

On pourrait s'imaginer une journée typique au sein de services consultatifs de Statistique Canada comme suit. On reçoit une demande par courrier, on consulte les différents catalogues et la documentation disponible sous la main, cherche l'information à partir des index pour ensuite localiser et reproduire les métadonnées appropriées. Il est probable que les catalogues ou la documentation ne fournisse toute l'information nécessaire à l'exécution de la demande. À ce moment on localise la publication pour en tirer la table des matières ou on consulte une personne-ressource. Le client peut alors recevoir une réponse à sa demande. Les demandes plutôt simples pourraient être réglées directement au téléphone alors que les plus complexes seraient alors traitées comme illustré ci-haut.

Le volume de demandes que Statistique doit traiter demandait que l'on organise les métadonnées sur nos produits et services à partir d'une base de données centralisée et dont chacun des intervenants pourrait faire des recherches et mieux servir notre clientèle. Nous avons élaborer l'IPS pour ainsi fournir la gestion et l'accès aux métadonnées centralisées et organisées. Nos employés n'ont plus à maintenir leur propre centre de référence et Statistique Canada peut assurer un niveau de service plus constant à sa clientèle.

2. IPS : LE PARTENARIAT APPLIQUÉ

A première vue, L'IPS pourrait être perçu comme l'incarnation électronique du catalogue traditionnel et des catalogues spécialisés devenus trop

[1] Louis Boucher, Division de la diffusion, Statistique Canada, Ottawa (Ontario) K1A 0T6.

nombreux, trop volumineux, qui demandaient beaucoup de travail et devenaient désuets le jour même de leur impression. Sans contredit, nous avions besoin de catalogues plus intelligents. En fait, l'IPS devait plutôt originalement servir à appuyer une politique administrative interne.

En mai 1994, la gestion supérieure de Statistique Canada approuvait la «Politique sur l'enregistrement des produits et services». Avec la croissance de la gamme de produits et services offerts par divers secteurs de l'organisation, la gestion sentait le besoin d'une certaine coordination et la consolidation dans ses activités de diffusion. En plus de certaines directives quant à la diffusion, la politique stipulait que «*... tous les produits et services seront enregistrés dans une base de données corporative des produits et services...*». Suite à son adoption, une telle base de données était élaborée par la Division de la diffusion et le processus d'enregistrement mis en place sous la gouverne de la Division du marketing pour l'ensemble de Statistique Canada.

Nous avons réalisé rapidement le potentiel de ces métadonnées pour nos activités de diffusion et l'impact possible pour mieux servir notre clientèle. Avec l'aide de nos services consultatifs, nous avons d'abord élaboré une application, à partir de la base de métadonnées, qui pourrait faciliter leur tâche. Par la suite, les services de la bibliothèque se sont joints à l'initiative IPS avec des données additionnelles et un besoin en termes de catalogue. De par le passé, les services de la bibliothèque ont toujours été les maîtres-d'oeuvre du catalogue de Statistique Canada. Plusieurs autres divisions de Statistique Canada se sont joints au projet depuis et y ont grandement contribué.

En bref, l'IPS n'est pas une solution voulue mais plutôt le fruit des opportunités reliées à la mise en place d'une base de métadonnées corporative. L'IPS est devenu un outil de première ligne pour nos services consultatifs, l'assise de la production de notre catalogue et des catalogues spécialisés et une référence générale pour tout le personnel oeuvrant auprès de la clientèle de Statistique Canada.

3. DES METADONNÉES AU BOUT DES DOIGTS

Dans ce chapitre, nous allons brièvement illustrer l'étendue de l'information disponible via l'IPS et élaborer sur les sources de métadonnées. La plupart des métadonnées existaient avant l'IPS, elles étaient toutefois isolées et d'accès des plutôt limité. L'apport majeur de l'IPS a été de rassembler toutes ces métadonnées et d'y aménager des outils de recherche efficaces. Sans avoir à déterrer les métadonnées, l'IPS fournit aux usagers toute l'information disponible et mise à jour à travers l'organisation. Peut importe qui fournit de l'information, le même outil nous permet de mieux servir notre clientèle.

Tel que mentionné auparavant, nous avons d'abord créé une base de données corporative qui par la suite a déclenché la conception d'un application de recherche couramment désigné IPS. La base de données corporative des produits et services contient les métadonnées des produits et services, les métadonnées sur les articles publiés en plus de contenir les identificateurs d'enquêtes et de séries chronologiques. Ces identificateurs permettent aux usagers de mieux comprendre et utiliser nos données en connaissant les caractéristiques des enquêtes et leur disponibilité en séries chronologiques.

Les métadonnées sur les produits et services de Statistique Canada sont la principale composante de l'IPS. Tel que décrit auparavant, le processus d'enregistrement assure la maintenance continue d'un répertoire organisé de tous les produits et services pour ainsi faciliter nos activités de diffusion. Par l'entremise de l'enregistrement, les usagers retrouvent une intégrité de présentation et de prix à travers la gamme de produits et services de Statistique Canada. Les métadonnées facilitent l'intégration de produits et services en aidant les auteurs et les spécialistes en la matière à localiser et à consulter les produits et services existants ou en cours de développement. Nous pouvons ainsi mieux coopérer et minimiser nos efforts et nos offres de produits et services à notre clientèle.

Nos services consultatifs reçoivent un nombre considérable de demandes d'information quant aux articles publiés par Statistique Canada. Il a toujours été difficile de localiser et de consulter la documentation de ces articles avant la mise en place de l'IPS. Au cours du développement de l'IPS, nous avons fait un effort particulier pour regrouper ou développer toutes les métadonnées nécessaires. Nos clients recherchent des métadonnées qui vont au-delà de la documentation des produits et services, ils veulent en connaître les sources.

Afin de fournir un accès unique et ponctuel, l'IPS regroupe plusieurs sources de métadonnées connexes aux produits et services. Les sources majeures de métadonnées disponibles par L'IPS proviennent des bases de données qui supportent le programme de publication et d'enregistrement et des fichiers de la bibliothèque. Les métadonnées sur les enquêtes proviennent du «Système de Documentation des

Données Statistiques (SDDS)» alors que le registre de CANSIM fournit celles sur les séries chronologiques.

4. L'INFRASTRUCTURE DERRIÈRE L'IPS

Après avoir brièvement décrit le contenu et les sources de l'IPS dans le chapitre précédent, nous allons maintenant discuter son côté technique.

L'IPS est une application de type client-serveur qui vise à exploiter la puissance et la fonctionnalité des ordinateurs personnels fonctionnant avec Windows. En plus, IPS repose sur les avantages de maintenir un répertoire centralisé de metadonnées sur les produits et services de Statistique Canada. Le système IPS compte trois composantes, les fichiers sur le poste de travail, les fichiers sur le serveur de réseau local et les metadonnées qui se trouvent sur le serveur central.

La composante sur le poste de travail

Votre poste de travail contient la partie «client» de la technologie SearchServer de Fulcrum en plus des fichiers ODBC et quelques fichiers systèmes Windows. Le montant d'espace de disque dur requis est d'environ 1.54 megaoctets.

La composante sur le réseau local

Votre réseau local contient les fichiers utilisés pour supporter la partie «client» de l'application elle-même. Ceci inclut les fichiers d'aide en ligne, le dictionnaire du recensement, les fichiers de la CTI-F, les fichiers de support aux roues de mots ainsi que le fichier d'exécution lui-même. Ces fichiers (environ 4.68 megaoctets) doivent maintenues dans un répertoire dont tous les usagers de l'IPS ont accès (pour un maximum de protection, ces fichiers sont en mode lecture seulement).

La composante serveur

Le serveur central contient l'information qui apparaît à votre écran à la suite de l'exécution d'une recherche avec IPS. Il contient aussi les fichiers "serveur" de SearchServer et tous les fichiers d'indexation qui permettent les activités de recherche et de récupération rapide.

La seule autre composante du système consiste en des lignes de communications. Ce système a été conçu afin de minimiser la circulation sur le réseau.

5. LES MÉTADONNÉES EN PLEINE ACTION

Nous avons déjà beaucoup discuté de ce qui se passe derrière le rideau et dont les usagers de l'IPS sont peu conscients alors qu'ils fournissent des services de consultation à leurs clients. Prenons maintenant quelques instants pour scruter l'écran et explorer certaines fonctions disponibles sur l'IPS. En bref, L'IPS fournit un accès unique à plusieurs sources de métadonnées de Statistique Canada et permet aux usagers de chercher et compiler des listes détaillées sur nos produits et services afin de satisfaire la demande. Pour faciliter de telles tâches, de nombreux outils et facilitateurs ont été incorporé à l'IPS. L'IPS comprend trois écrans de base, à savoir l'écran principal, la liste de résultats et l'écran des métadonnées. A travers ces écrans, des menus déroulants et des boîtes de dialogue fournissent tous les outils et les facilitateurs nécessaires.

L'écran principal de l'IPS est constitué de quatre composantes, la barre de menu standard, la barre d'outils, le panneau de recherche et une carte de résultats. La barre de menu standard fonctionne comme la plupart des applications Windows et présente des menus déroulants pour les commandes et fonctions disponibles. Une fonction particulièrement intéressante consiste en la possibilité de passer d'une la langue de la base de données à l'autre (français-anglais) sans avoir à démarrer une nouvelle session de recherche. Dans la mesure où il n'est pas possible de savoir à l'avance la langue du client qui nous appelle, cette fonction est des plus utiles afin de servir notre clientèle dans la langue de leur choix.

La barre d'outils facilite l'accès à de nombreuses commandes et fonctions et se retrouve sur tous les autres écrans. Il est possible de gérer l'ensemble de nos recherches en les sauvegardant, les récupérant ou en ayant accès aux plus récentes recherches directement. Des outils de référence comme le dictionnaire du recensement et la classification type des industries sont présentement disponibles pour consultation interactive à l'intérieur même de l'IPS et nous songeons à en ajouter plusieurs autres sous peu. La barre d'outils donne accès à l'aide en ligne via le bouton «Guide à l'utilisation de l'IPS».

Le panneau de recherche permet aux usagers de faire des recherches des plus complexes en utilisant des opérateurs booléens. Vous pouvez diriger votre recherche à travers les mots partout, les titres, les sujets, les auteurs et le médium recherché, parmi plusieurs autres alternatives. L'usager peut entrer directement ses critères de recherche ou les sélectionner à partir de listes déroulantes (roues de mots). Par exemple, on peut faire

une recherche sur «emploi» et «chômage» dans les mots partout en association avec «scolarité» comme sujet et seulement sur «CD-ROM». Les produits et services présentés seront ceux qui correspondent à tous ces critères de recherche.

Nous retrouvons finalement sur l'écran principal une carte de résultats. À la suite de chaque recherche, l'IPS fournit le résultat préliminaire de la recherche en affichant le nombre de produits et services qui satisfont cette recherche. Ce pointage a plusieurs objectifs. Premièrement, vous savez si votre recherche a été vaine ou non et vous indique si vous devez rajuster vos critères. Deuxièmement, on minimise la circulation entre votre poste de travail et le serveur principal en n'effectuant que le compte plutôt que la recherche en elle-même. Les résultats sont présentés par médium, à savoir combien de produits imprimés, d'articles, de séries chronologiques, de produits électroniques et ainsi de suite correspondent à votre recherche.

Vous accédez au deuxième écran, celui de la liste de résultats, en sélectionnant une ou toutes les catégories de la carte de résultats. C'est là où vous retrouvez une liste organisée de produits et services pour votre recherche. Pour chaque item, on y retrouve le titre, l'identificateur et le rang (pointage déterminé à partir du nombre d'occasions où le ou les critères de recherche sont présents dans les métadonnées pour cet item). Cette liste peut être réorganisée par rang, titre ou identificateur au besoin. Les critères de recherche sont sur lignés dans le titre. Vous pouvez sélectionner un ou plusieurs items à partir de cette liste.

Jusqu'à présent dans l'exemple décrit, l'IPS a utilisé les métadonnées en arrière-plan sans trop les présenter à l'usager. C'est en sélectionnant un ou plusieurs items de la liste de résultats que l'IPS reproduit les métadonnées sur l'écran des métadonnées. Pour chaque item sélectionné de la liste, l'IPS affiche toute l'information disponible sur ce produit ou service. Vous pouvez vous déplacer à travers le document (terme suivant ou précédent) ou aller à un autre document, si vous en avez sélectionné plus d'un de la liste. Vous pouvez aussi copier le texte ou une partie si nécessaire et l'utiliser dans votre logiciel de traitement de textes favori. De plus, il est possible de poursuivre avec une recherche intuitive à partir d'un document en y choisissant un mot ou groupe de mots et ainsi démarrer une recherche plus approfondie.

L'écran des métadonnées donne aussi accès aux autres métadonnées connexes au document présenté. Par exemple, lorsqu'approprié, les métadonnées sur les enquêtes, les versions électroniques ou les séries chronologiques sont disponibles par l'entremise des boutons de produits connexes.

L'IPS permet de générer plusieurs formats de sortie pour les métadonnées à partir des différents écrans de base. Il est possible d'imprimer directement, de créer des fichiers texte ou de générer des fichiers en SGML. La sélection des items se fait soit en choisissant le document courant, soit en marquant certains documents ou en sur lignant des parties du document. Des formats de sortie pour les descriptions courtes et longues des produits et services ainsi que pour les listes sont prédéfénis. Il est de plus possible de générer des mini-catalogues ou fournir toute l'information nécessaire au catalogue général.

6. DÉVELOPPEMENTS À VENIR

L'IPS est actuellement utilisé par le personnel de Statistique Canada mais sera sous peu disponible au public via notre site Internet et vraisemblablement reproduit sur CD-ROM. L'IPS en est à ses premiers pas et saura bénéficier de nombreux développements à venir. Parmi ces développements, certain aspects retiennent notre attention.

L'IPS bénéficiera d'un apport de nouvelles métadonnées pour assurer la pleine couverture des table de matières et des articles, entre autres. Nous sommes déjà à développer un thésaurus pour faciliter la recherche par l'utilisation de terminologie normalisée. Un accès plus répandu à l'intérieur de Statistique Canada et au public général via Internet et CD-ROM permettra au personnel et à la clientèle d'avoir un meilleur aperçu des métadonnées disponibles. Nous espérons que l'accessibilité accrue à ces métadonnées incitera nos usagers et clients à mieux comprendre et utiliser nos produits et services ainsi que nos données.

Dans un avenir rapproché, Statistique Canada verra à rapprocher ses initiatives de métadonnées et d'entreposage de données, dont l'IPS et l'outil de recherche thématique. Une telle interaction entre ces initiatives bénéficiera à l'ensemble de Statistique Canada et encore plus important, verra à fournir à nos usagers et clients une meilleure compréhension de nos produits et services. L'IPS est un exemple encourageant de coopération et de synergie entre les nombreuses divisions de Statistique Canada et saura sûrement continuer à générer une approche créative et des objectifs partagés.

SESSION 8

Diffusion électronique des données

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

COMMENT RÉUSSIR SUR LES MARCHÉS SANS BOULE DE CRISTAL

U. de Stricker[1]

RÉSUMÉ

Les organismes statistiques, peu importe comment ils perçoivent leurs rôles, font partie du secteur de l'information, parce qu'ils détiennent des données de grande valeur. Le secteur de l'information se distingue par une évolution technologique rapide et importante, dont les diverses conséquences sont déterminantes : les clients exigent des produits d'information de plus en plus commodes et flexibles, et les nouvelles technologies de collecte et de distribution de données augmentent la concurrence. Or, une façon de prospérer dans une branche d'activité risquée consiste à miser sur les partenariats avec d'autres intervenants dans le but d'exploiter et de bonifier les compétences et les données disponibles.

MOTS CLÉS : Concurrence sur les marchés; évolution technologique; attentes des clients.

1. LE SECTEUR DE L'INFORMATION

Fin octobre, l'*Information Industry Association* — un groupe industriel influent établi à Washington en 1968 — a tenu sa conférence annuelle à Toronto. Plusieurs conférenciers se sont réjouis du fait qu'à la fin de la décennie le secteur de l'information constituerait un segment de l'économie représentant 20 % du PNB des États-Unis.

Deux choses me frappent : primo, c'est que 20 % équivaut à un bon morceau du PIB; secundo, je suis ici pour vous en parler. Je ne suis par certaine qu'on m'aurait invitée à présenter un exposé au Symposium il y a cinq ans, sur le sujet de la commercialisation de l'information. Mais le monde a changé.

Il n'y a pas si longtemps, les organismes publics du domaine de la statistique diffusaient principalement leurs données à l'intention des intérêts pour le compte desquels celles-ci avaient été recueillies. Mais ce n'est plus le cas. Aujourd'hui, les organismes statistiques font partie du secteur de l'information — que cela les enchante ou les rebute.

Cette branche d'activité toute récente impose de nouvelles règles et de nouveaux défis, offre de nouveaux débouchés et attire des conférenciers comme moi, qui viennent exposer leurs vues à l'occasion d'événements comme le symposium qui me sert de tribune aujourd'hui.

Évidemment, le monde de l'information n'est pas de tout repos. Il nous oblige à harmoniser des impératifs apparemment contradictoires comme la nécessité d'encaisser des recettes et d'assurer à la fois le bien public, de préserver l'intégrité de l'organisation et de tirer parti d'actifs de grande valeur.

Par ailleurs, si l'ampleur des défis à relever vous intimide, vous avez su reconnaître, au cours de ce symposium, l'objectif fondamental du secteur de l'information, qui consiste à répondre aux besoins et aux attentes de la clientèle. Or, si votre MISSION consiste à diffuser, de la manière et au moment opportuns, des produits et services conformes aux exigences de vos clients, vous devrez, POUR RÉUSSIR, assurer la « vendabilité » des produits et services offerts. Je n'ai pas de recettes magiques, mais j'ai néanmoins consulté ma boule de cristal et j'espère que les messages que j'en ai tirés vous aideront à garantir, en permanence, la valeur marchande maximale de vos produits et services. Nous oeuvrons dans un secteur difficile, qui se distingue par des changements plus rapides et plus profonds que tout ce qu'on a pu voir jusqu'à présent.

2. PREMIER MESSAGE DE LA BOULE DE CRISTAL

Ma boule de cristal m'a transmis deux messages. Voici le premier : ATTENTION AUX CLIENTS et, parallèlement, MÉFIEZ-VOUS DES VACHES

[1] Ulla de Stricker, de Stricker & Associates, 3934 Selkirk Place, Mississauga (Ontario), L5L 3L5, tél: (905)820-4525.

SACRÉES. Il concerne l'incidence de la technologie sur les attentes que nourrissent les consommateurs de produits d'information, et les relations avec la clientèle.

Je constate que l'appétit pour l'information s'accroît de manière exponentielle à mesure que le volume d'information disponible augmente. Or, ce n'est pas de renseignements quelconques, présentés dans d'emballages quelconques qu'on veut, mais de flexibilité, de variété et de commodité, ainsi que d'une longueur d'avance sur la concurrence. Les clients veulent recevoir automatiquement l'information correspondant à leur profil d'intérêt; ils veulent pouvoir manipuler les données exactement comme ils l'entendent, différemment selon les exigences du moment. Ils refusent de perdre du temps et exigent, au besoin, un solide soutien technique. Parfois, ils veulent des services d'interprétation et d'analyse; à d'autres moments, ils préfèrent analyser eux-mêmes les données de base. Il se trouve qu'ils disposent des outils logiciels nécessaires pour exploiter les données qui les intéressent et qu'ils en obtiendront de nouveaux, à mesure que les sociétés de services et d'ingénierie en informatique pourront satisfaire leurs exigences en la matière.

Ces traits généraux des utilisateurs de données sont présents, à divers degrés, chez l'ensemble de la clientèle, dont le spectre s'étend des écoliers aux professeurs d'universités et aux professionnels, en passant par le «grand public» (indépendamment de ce qu'on entend par cette vague notion). Le stratagème consistera, bien sûr, à offrir à chaque type de client la possibilité de trouver la combinaison précise des attributs qui lui conviennent le mieux en termes de contenu, de mode de diffusion, de facilité de manipulation et de prix. Les services d'information généraux ne conviendront pas aux clients de l'avenir, et les difficultés inhérentes au développement d'une infinie variété de services personnalisés soulignent la nécessité d'apporter un soin extrême à la stratégie en matière de diffusion.

Les clients ne tolèrent plus ce que j'appelle les « vaches sacrées », situation découlant du gonflement sans précédent des exigences des utilisateurs de données. En dehors d'un certain niveau d'exactitude, de fiabilité et de nouveauté, les clients n'ont que faire des règles et des idéaux méthodologiques des organismes statistiques à l'égard de leurs données; curieusement, ils sont seulement désireux d'obtenir les renseignements dont ils ont besoin pour faire leur travail. Mes collègues aiment dire « laissons le client décider de ce qui est assez bon pour lui — assez bon pour lui permettre d'atteindre les objectifs qu'il s'est donnés».

Dans cette perspective, nous, les professionnels de l'information, utilisons souvent comme exemple ce qu'on appelle le « continuum d'information ». Les clients qui valorisent les renseignements immédiats (p. ex. le budget fédéral quelques secondes après sa présentation, les nouvelles en temps réel) ne s'intéressent pas au mode de présentation, aux erreurs typographiques, à l'absence ou à la présence de raffinements éditoriaux; par contre, les clients qui valorisent l'analyse, la structure du message et les commentaires réfléchis sont généralement disposés à attendre quelques minutes, quelques heures, voire des mois pour obtenir ce qu'ils veulent. En règle générale, les éditeurs s'aperçoivent qu'ils peuvent livrer la marchandise et gagner de l'argent à vendre de l'information aux différents points du continuum, qui va des « données immédiates mais brutes » aux « informations tardives mais soignées ». Le fait est que les clients veulent choisir eux-mêmes en sachant parfaitement ce qu'ils obtiennent (p. ex., «ces données proviennent d'un petit échantillon tiré il y a trois ans et...»). L'acceptabilité des données dépend de leur finalité.

Alors, que veut dire pour vous ce premier message? À mon avis, il signifie que votre avenir dépend non seulement de la connaissance intime des besoins de vos clients, mais également du maintien de communications continues avec eux, afin que vous ne perdiez jamais de vue les modes qu'ils privilégient en matière de diffusion de données. Autrement dit : à l'heure actuelle, vos clients sont en mesure d'apprécier vos produits au moment où ceux-ci sortent de vos bureaux ou de vos usines, après une période de développement ardue d'une durée plus ou moins longue — parfois des mois ou des années. Mais à l'avenir, vos clients (ou leurs agents qui sont plus proches d'eux que vous ne l'êtes) voudront participer aux efforts de votre entreprise et tenter de déterminer avec vous les points du continuum « qui se prêtent le mieux » à l'extraction des données et, le cas échéant, les compromis que vous devrez faire concernant les conventions de votre entreprise.

3. DEUXIÈME MESSAGE DE LA BOULE DE CRISTAL

Le second message est le suivant : SOYEZ À L'AFFÛT DE LA CONCURRENCE IMPRÉVUE et, parallèlement, TROUVEZ-VOUS DES PARTENAIRES. Ce message concerne l'impact de la technologie sur le positionnement des organismes statistiques ou des ministères dans le marché de l'information.

Vous ne serez pas surpris d'apprendre que tout sentiment de sécurité à l'égard du positionnement de

votre organisation vis-à-vis de la concurrence ne peut être qu'un leurre. La technologie accroît de manière exponentielle les possibilités qui, à leur tour, créent de nouveaux débouchés pour vous et de nouveaux choix pour vos clients. Ainsi les occasions favorables qui s'offrent à vous sur le marché de l'information s'offrent également à d'autres organisations — dont certaines peuvent mieux que vous s'adapter à la compression des calendriers de production qui se mesurent maintenant en semaines, alors qu'ils s'étiraient naguère sur des mois ou des années.

J'ai dégagé CINQ aspects du message de ma boule de cristal sur les effets de la technologie :

Premièrement, la collecte de données et la constitution de bases de données ne sont plus l'apanage exclusif d'un petit nombre d'organisations privilégiées. La technologie permet aujourd'hui de recueillir facilement des données à faible coût, et parfois celles-ci sont tout simplement des produits dérivés d'autres activités. Par exemple, tout comme l'invention du code à barres des supermarchés pour faciliter la tâche des caissiers a permis d'obtenir des renseignements précieux sur les habitudes de consommation, on peut facilement imaginer que la technologie des satellites utilisée pour suivre les objets mobiles pourrait servir au contrôle de la circulation. De la même manière, la technologie interactive de l'Internet peut exécuter, aisément et à bon marché, des tâches visant à regrouper des informations. Les progrès de la technologie de la câblodistribution et de la téléphonie accroîtront les possibilités au chapitre de la collecte de données sur les transactions des ménages et des entreprises; et ainsi de suite. En l'absence de données bon marché et facilement accessibles auprès des sources évidentes, le marché se chargera de créer les sources de remplacement dont il a besoin.

À propos, permettez-moi de mettre en relief un phénomène particulier que mon collègue américain, Stephen Arnold, appelle la « Network Publishing » (édition par réseau) dans un livre à paraître traitant des nouveaux médias d'information. Rendue possible par l'infrastructure de réseau dont l'Internet est un bon exemple, la « Network Publishing » ou, si vous préférez, la création de contenus en collaboration, signifie que la diversité des noeuds d'un réseau peut contribuer au développement d'une ressource de données commune. Nombre de ressources d'information disponibles sur l'Internet illustrent le type de structure dont il s'agit; par exemple, des informations sont recueillies ou créées dans de nombreux sites, après quoi le réseau rassemble le tout en une entité virtuelle apparaissant à l'utilisateur comme une ressource unique. On voit facilement comment un réseau de participants répartis aux quatre coins du pays pourrait rassembler une base de données commune contre une certaine forme de compensation ou tout autre facteur de motivation (p. ex., une visibilité accrue, facteur d'incitation d'extrême importance de nos jours). Je ne propose rien de révolutionnaire; mais ma proposition rend possible le déploiement d'efforts à grande échelle (des efforts équivalant à l'accomplissement d'une tâche quasi-impossible comme de convaincre l'ensemble des livreurs de journaux du pays de vous dresser la liste des marques et modèles des voitures garées sur les voies d'accès de leur parcours, de vous décrire le contenu des boîtes bleues, le nombre et l'âge des enfants, la race ou l'espèce des animaux domestiques, etc.).

L'accroissement fabuleux du nombre d'« agents logiciels » qui envahissent l'Internet à la recherche de toutes sortes de choses est un autre facteur d'incitation. Comme le contenu du Réseau W3 double environ tous les 24 jours, la quantité de données accessibles aux recherches de ces agents n'est pas à dédaigner...

Autre élément à considérer : les éditeurs de bases de données traditionnelles cherchent à déléguer aux machines un travail jusqu'ici dévolu aux humains, de manière à pouvoir exploiter des volumes toujours croissants de données et à maximiser la productivité en réservant les facultés du cerveau humain à des tâches d'analyse et de créativité, plutôt qu'à des corvées ingrates et fastidieuses comme l'indexation et la classification. Il se peut que les textes indexés par une machine ne soient pas aussi bien indexés qu'ils le seraient par un spécialiste de l'édition, mais ce dernier travaille beaucoup plus rapidement lorsqu'il s'attaque à un texte après que la machine y a jeté un premier coup d'oeil. Vous avez été témoins, au cours des ans, d'une évolution similaire des bases de données statistiques, et vous ne serez pas étonnés d'apprendre qu'un des prochains grands secteurs d'intérêt est l'application des logiciels à des tâches qu'on effectue encore manuellement, ce qui nous laissera du temps pour nous occuper des défis que posent les exigences de la clientèle.

Le second aspect du message sur la technologie est que la manipulation et la distribution de données ne relèvent plus du domaine exclusif d'organisations vouées à l'exploitation de bases de données, offrant des services directement accessibles par ligne commutée; l'explosion des applications que l'on retrouve sur l'Internet et la puissance grandissante des machines des utilisateurs sont responsables de cette évolution. L'architecture de réseau TCP/IP a tranquillement distancé les technologies d'IBM et de Novell; un logiciel (Mosaic/Netscape) créé en laboratoire à Genève, à l'été 1993, est devenu le

«moteur » de l'Internet que nous connaissons aujourd'hui, transformant un environnement axé sur la connectivité en un environnement d'applications. Quiconque dispose d'un ordinateur le moindrement performant et de logiciels faciles à utiliser peut se lancer dans l'édition et la collecte de données sur le Net, et diffuser de l'information et des documents dont l'apparence haut de gamme rivalise avec tout ce que peut faire une firme spécialisée en graphisme; et cela, en une fraction du temps que mettraient des professionnels à sortir le produit.

L'incidence économique des transformations précitées est que, les impératifs financiers des nouveaux « éditeurs » étant différents de ceux des grandes entreprises, on constate une pression à la baisse sur les prix et une adaptation conséquente des attentes de la clientèle à cet égard. Avec moins de finesse, je dirais que les clients ont l'esprit critique et s'attendent à obtenir ce qu'il y a de mieux pour leur argent, même pour très peu d'argent. Macrorecettes pourraient bien un jour rimer avec microprix.

Le **troisième** aspect du message sur la technologie est que les logiciels favorisent une intimité accrue entre les vendeurs et les acheteurs. Par exemple, les sociétés émettrices de cartes bancaires et de cartes de crédit et leur clientèle peuvent se relier beaucoup plus étroitement si l'ordinateur des clients peut parler directement avec l'ordinateur de la banque (de la compagnie aérienne ou de l'entreprise de location d'automobiles). L'impact de la fidélisation de la clientèle est évident, comme le sont les nouvelles possibilités offertes en matière de collecte de données.

La capacité de communication avec la clientèle se perfectionne constamment, et cette donnée est le **quatrième** aspect du message sur la technologie. Par exemple, les brochures coûteuses en couleur sont devenues pur gaspillage et sont remplacées par des brochures électroniques mises à jour en temps réel; l'exactitude des méthodes de ciblage s'accroît constamment à mesure que les méthodes de marketing permettent d'atteindre la communauté virtuelle des clients, par-delà l'espace physique.

Le **cinquième** aspect est lié à la transition que l'on constate entre une « technologie simple, fondée sur la connectivité » et la nouvelle « technologie axée sur les applications ». Il n'est pas étonnant que l'Internet soit un lieu illustrant au mieux cette évolution et, à mon avis, l'Internet démarrera vraiment le jour où de nouveaux progiciels comme Java et Blackbird seront largement répandus. Par exemple, des intérêts suisses offrent actuellement sur l'Internet un fichier d'archivage concernant le change sur les devises, qui possède des capacités de calcul. Des outils comme Java permettent aux utilisateurs de demander un service d'alerte qui se déclenche dans certaines conditions, d'examiner des cotes en temps réel, de faire des offres, etc.

Les nouveaux outils créeront une « plate-forme commerciale » caractérisée par des « sites fonctionnant comme des aimants », capables de regrouper tous les participants intéressés à un domaine virtuel particulier. Permettez-moi d'illustrer cette situation par un exemple neutre pris dans le domaine du jazz, que vous pourrez facilement transposer dans votre propre domaine : admettons qu'une ressource contenant des feuilles de musique et des enregistrements de jazz soit mise à votre disposition sous forme numérique dans un site de l'Internet. Disons qu'un commanditaire organise un forum « d'échanges » où les collectionneurs de jazz peuvent acheter et vendre, et généralement faire connaître à leurs collègues ce qu'ils recherchent. On fera l'hypothèse qu'il y a également, pour les étiquettes des enregistrements, un comptoir relié au site. Supposons même qu'un studio d'enregistrement participe à l'événement, offrant, sur demande, des enregistrements qui seraient autrement inaccessibles, accompagnés de fichiers types audios. Supposons encore que des historiens y ajoutent photos et archives vidéographiques d'interprétations ou de spectacles. Ajoutons que des musicologues et des responsables de département de musique d'universités du monde entier ont encouragé leurs professeurs et leurs étudiants diplômés à verser les résultats de leurs recherches aux fichiers d'archivage, toujours croissants, de contenu original (sommaires gratuits, texte intégral moyennant des frais). Disons encore qu'un fabricant d'instruments de musique décide également de participer. Que dirait-on d'un programme de concert, d'un programme de visites guidées et d'un service de réservation pour des groupes spécialisés en jazz? D'un catalogue spécialisé d'ouvrages et d'autre matériel de recherche? De la liste des musées de la Nouvelle-Orléans? D'une coopération en vue de composer de la musique? Et si l'animateur de visites organisait des visites spécialisées sur des sites historiques pour les mordus du jazz? Et pourquoi pas une coopération en vue de composer de la nouvelle musique de jazz? Etc., etc. Je pense que vous pouvez maintenant vous faire une idée des débouchés possibles en ce qui a trait aux environnements informatifs portant sur le genre de données que vous diffusez — c'est la notion de supermarché raffinée jusqu'à devenir de l'art... Rappelez-vous que toute cette activité, tant scolaire que commerciale, a lieu dans un espace virtuel, grâce aux outils disponibles et exécutables sur l'Internet.

4. LES BONNES NOUVELLES

Tout ceci pour dire qu'il faut absolument guetter la concurrence là où l'on n'aurait jamais pensé regarder. Une foule d'organisations que vous n'auriez jamais rêvé de retrouver parmi vos concurrents sont soudain capables de rassembler et de manipuler les données que vous aviez crues votre domaine exclusif et de diffuser des renseignements et des solutions à des gens que vous avez toujours pris pour vos clients. IBM, Wang, Wordstar et même Microsoft ont appris que le positionnement sur les marchés n'est pas durable sans une constante vigilance.

Y a-t-il du « positif » à notre nouvelle situation? Certainement. Tout le monde a probablement compris pourquoi le prochain panel traite de partenariats. Mon expérience du secteur — mettons de côté la boule de cristal — me dit que, d'une façon ou d'une autre, pour prospérer dans le secteur de l'information, l'intervenant doit trouver des partenaires et des alliés qui l'aideront à garder une longueur d'avance sur la progression irrésistible des technologies qui, autrement, risqueraient de le mettre rapidement hors jeu; qui l'aideront aussi à exploiter ses informations et ses compétences, sans compter que partenaires et clients tireront également profit de l'accroissement des débouchés qui s'ensuivra.

La technologie n'est pas le fruit du hasard. Elle s'enrichit de l'énergie des gens qui l'utilisent et la développent pour leurs propres fins. Pour bâtir votre réussite sur la technologie plutôt que de voir cette dernière comme un problème, il vous faut tirer parti de tout ce que les autres ont fait pour l'exploiter avantageusement et apporter une contribution qui augmentera encore la valeur de ce qu'elle peut donner : je parle évidemment de vos données et du capital intellectuel que vous aurez investi en elles.

C'est drôle : à la conférence de l'*Information Industry Association*, j'ai remarqué que les cadres supérieurs des entreprises du secteur de l'information étaient du même avis que moi.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

LANDDATA BC : LE PASSÉ ET L'AVENIR

G. Sawayama, H.A. Kucera et E. Kenk[1]

RÉSUMÉ

Les internautes ne tarissent pas d'éloges sur les vertus démocratiques de l'autoroute de l'information mais ils oublient que le commun des mortels est dérouté par la complexité de l'échange d'une poignée de mains sur ordinateur, via le téléphone. Les nouveaux systèmes d'exploitation qu'on voit poindre à l'horizon promettent de simplifier le rituel des branchements point à point. Mais ensuite? Comment trouver ce que l'on cherche? Commander et recevoir les données pourrait bien devenir banal, mais tirer quelque chose de sensé de ce qu'on finit par télécharger sûrement pas. LandData BC est une infrastructure de services d'information (un système de contrôle de la circulation, et non un poste de péage) mis au point par Macdonald Dettwiler pour le gouvernement de la Colombie-Britannique. Un prototype de ce système qui veille à la consultation, à la diffusion et à l'intégration des données spatiales (géographiques et textuelles) est en usage depuis mai 1993. La mise au point et l'exploitation du prototype nous ont donné quelques leçons utiles qui se refléteront dans notre approche technologique et administrative, à mesure que la version commerciale du produit se concrétise.

MOTS CLÉS : Données spatiales; LandData BC; infrastructures de services d'information.

1. INFRASTRUCTURES DE SERVICES D'INFORMATION

Si les émetteurs hyperfréquence, les câbles à fibres optiques et les systèmes de commutation complexes définissent l'infrastructure *matérielle* popularisée sous l'expression «autoroute de l'information», on pourrait considérer LandData BC[2] comme une infrastructure de *services* permettant d'accéder à des *programmes* (données, applications et représentations) ou de relayer ceux-ci par l'autoroute de l'information, en Colombie-Britannique.

L'implantation d'une telle infrastructure de services suppose deux choses : convaincre les institutions qui disposent d'importantes banques de données à adhérer au programme et amener les abonnés qui désirent s'en servir à en faire autant. Vers le milieu des années 70, Willis Roberts et Angus Hamilton[3] ont parlé de courtage de l'information pour décrire cette manière classique de jumeler l'offre et la demande. En réalité, cependant, il existe un troisième aspect, à savoir l'épanouissement des services à valeur ajoutée qui font le pont entre l'offre et la demande. Le défi pour les infrastructures de services d'information consiste à synchroniser l'élaboration de leurs composantes, parallèlement aux investissements dans l'infrastructure physique. Bill Gates et ses homologues se trouvent aujourd'hui au même point que les grandes chaînes de télévision au début des années 50.

Les infrastructures de services d'information peuvent jouer deux rôles : proposer des répertoires structurés de fournisseurs de données et de leurs produits, et acheminer les données du fournisseur à l'utilisateur. Une infrastructure bien conçue offrirait des *rayons* auxquels des tierces parties pourraient venir se fixer afin d'incorporer leurs services à valeur ajoutée à l'infrastructure. Des experts (humains ou mécaniques) s'occuperaient du tri, de la transformation, de l'intégration, de la visualisation et, dans certains cas, de la prise de décisions afin de dispenser aux consommateurs des services extérieurs aux données.

Les infrastructures de services d'information exigeront un investissement qui viendra de trois sources: les fournisseurs de données qui souhaitent commercialiser leurs produits, les abonnés qui désirent accéder aux données et les promoteurs de l'infrastructure matérielle qui recherchent des programmes à véhiculer.

2. LE GOUVERNEMENT EN TANT QU'INVESTISSEUR

On pourrait se demander si les forces du marché justifieront à elles seules l'injection rapide de sommes

---

[1] Gary Sawayama, Henry A. Kucera, et Evert Kenk, Geographic Data BC, British Columbia Ministry of Environment, Lands & Parks, 1802 Douglas Street, 4th floor, Victoria, (Colombie-Britannique) Canada, V8V 1X4.

élevées dans l'infrastructure matérielle et l'infrastructure de services. L'administration actuelle des États-Unis[4] saisit très bien la symbiose qui existe entre ces deux éléments et l'importance stratégique de chacun d'eux. Elle l'a démontré par l'intérêt qu'elle porte à la National Spatial Data Infrastructure.

En 1989, le gouvernement de la Colombie-Britannique s'est engagé à créer une infrastructure de services d'information après avoir compris l'existence d'un besoin opérationnel qui, si on parvenait à le satisfaire, présenterait un potentiel stratégique sur le marché. La dépendance de l'économie provinciale à l'égard des ressources naturelles, les débats acrimonieux qu'engendre l'exploitation soutenable des terres d'un bassin hydrographique à l'autre et les revendications territoriales des Autochtones ont tous fait couler beaucoup d'encre.[5,6,7,8]

Les méthodes spéciales ayant pour but de recueillir les données qui devaient étayer les décisions étaient inadéquates. C'est pourquoi l'administration de la Colombie-Britannique s'est efforcé d'élaborer une infrastructure de services d'information sur les terres qui couvrirait toutes les bases de données pertinentes du gouvernement provincial. Des centaines de services publics de la Colombie-Britannique investissent chaque année 50 millions de dollars pour recueillir des données spatiales; en outre, cette somme ne tient pas compte des analyses ni de l'utilisation des données. L'infrastructure de services d'information sur les terres, baptisée depuis LandData BC, a vu le jour bien avant l'adoption de la loi provinciale sur l'accès à l'information et la protection de la vie privée. Elle poursuivait l'objectif modeste (mais extraordinaire) d'aider d'autres ministères à accéder à l'information gouvernementale afin d'assurer un meilleur service à la population.

LandData BC mettrait un terme à la répétition inutile des efforts de collecte et de gestion des données spatiales. Il aiguillerait les décideurs du gouvernement vers des sources normalisées d'information interne, ce qui garantirait une plus grande cohérence des décisions au sein d'une administration décentralisée. Ainsi, on résoudrait le problème des licences de déversement des déchets délivrées par un service du ministère de l'Environnement, des Terres et des Parcs, dans l'ignorance des permis d'exploitation hydraulique, en aval accordés par un autre service du même ministère.[9]

LandData BC ne se borne pas à garder et à acheminer les données comme le suppose le rôle des infrastructures de services d'information décrit précédemment. En réalité, l'organisme propose une gamme de services experts (sans écarter la possibilité que des tierces parties procurent d'autres services du même genre) afin d'offrir une vue intégrée des données issues de maints fournisseurs. LandData BC épouse le concept des fédérations de données créées par des groupes d'intérêt s'efforçant d'instaurer des normes qui assureront l'interopérabilité des données.

3. LANDDATA BC : LE PROTOTYPE

Macdonald Dettwiler, notre expert-conseil sur LandData BC, a commencé à se renseigner sur les besoins des utilisateurs en décembre 1990, en testant une série de prototypes rapides simulant la fonctionnalité du produit pour l'utilisateur final. Des spécifications ont suivi, puis on est passé à la conception avancée en décembre 1991. Après perfectionnement, développement et essai, le prototype a été livré en mai 1993. Au départ, celui-ci devait démontrer la faisabilité technique du projet, et non servir à vérifier les hypothèses d'une commercialisation éventuelle. Les possibilités du prototype ont été élargies par adjonction d'un module de comptabilité, ce qui en permet l'utilisation comme infrastructure de services provisoire. En tout, le prototype a coûté un peu moins de trois millions de dollars et a vu le jour sur une période de 30 mois.

LandData BC s'appuie sur un jeu de politiques, de méthodes et de normes en trois volumes baptisé *Land Information Management Framework* (LIMF). Les ministères de la Colombie-Britannique sont tenus de se conformer au LIMF dont les principes sont **théoriquement** intégrés au cycle qui conduit à l'approbation des budgets et des dépenses. Le LIMF n'englobe pas des spécifications comme le système de référence NAD 83 et le *Spatial Archive Interchange Format* (SAIF) mais il renvoie à ceux-ci. LandData BC a été conçu d'emblée selon l'hypothèse que les données ne seraient pas centralisées dans des banques physiques, mais continueraient d'être gérées par leurs conservateurs respectifs. Le LIMF maintient la cohésion de cette infrastructure virtuelle.

3.1 État du prototype actuel

Le prototype actuel utilise le langage TPC/IP. La banque de données et le module de comptabilité fonctionnent sur un serveur libre Sybase raccordé à un ordinateur Unix situé à Victoria. Le logiciel User Access, qui permet à l'utilisateur d'accéder aux données, a été élaboré en environnement Windows au moyen d'un SIG bon marché appelé Terraview. Le logiciel a été installé sur une douzaine de plateformes d'ordinateurs personnels existants, à Victoria et dans un bureau

régional, à 750 km de là. À Victoria, la connexion s'établit à la vitesse de 100 mégaoctets par seconde (en bandes segmentées d'une largeur efficace de 10 mégaoctets) pour le réseau métropolitain, ce qui est conforme aux vitesses Ethernet entre clients et serveurs. Le mois dernier, des dispositions ont été prises pour autoriser l'accès par modem à la vitesse de 28 800 bauds. Le logiciel User Access permet de fureter dans la base de données, d'envoyer des commandes électroniques et d'établir la couverture géographique des données spatiales grâce à une interface géographique. Le logiciel fait aussi office de visionneuse pour les produits de données en ligne dont il est question à la partie 3.1.1. La trousse User Access, qui comprend les licences Sybase et Terraview ainsi que PC-NFS pour la communication, coûte 1000$ mais un rabais de 50% est consenti pour les achats dépassant la centaine. La sélection d'autres produits pour la version commerciale du projet pourrait faire baisser le coût de la trousse à 100$, étant donné le nombre de branchements visé.

Le prototype gère simultanément jusqu'à 12 interrogations (limite imposée par la licence Sybase mais qui pourrait être relevée) à partir d'un ordinateur personnel utilisant le logiciel User Access. La banque énumère 35 sources d'information totalisant au-delà de 200 produits de données sur les terres qui vont de cartes à des rapports et à des données brutes. On peut commander directement un des produits énumérés d'un site User Access. L'entente officielle conclue avec les conservateurs de données exige que ces derniers accusent réception de la commande dans les trois jours et y répondent dans les 30 jours. Selon la présentation matérielle des données (à savoir, publication, etc.), le produit est acheminé à bon port par messager ou par un moyen de télécommunication. C'est ce qu'on appelle les commandes hors ligne.

3.1.1 Produits de données en ligne

Tel qu'indiqué précédemment, le prototype n'est pas simplement un dépôt de données et ne sert pas qu'à la livraison hors ligne des produits de données par leurs propriétaires. Trois produits de données sont aussi accessibles en direct pour des recherches à partir d'un site User Access : topographie, cartes cadastrales et registre des terres publiques.

Les recherches précises sur LandData BC peuvent être entreprises à partir de n'importe quel site User Access pour un des trois produits de données précités. Ces derniers résident sur des plateformes différentes à des endroits distincts et sont gérés par leur propre conservateur à Victoria. Les demandes formulées en SQL sont traduites dans le langage de recherche original par un serveur Data Access, puis appliquées aux données appropriées. Chacune des trois bases de données dispose d'un serveur Data Access spécial qui traite les demandes par l'entremise de l'interface reliant les systèmes d'information public et privé.

Les données de géométrie spatiale sont renvoyées au serveur Data Access qui les convertit en format SAIF/XDR (pour l'instant, les attributs textuels sont transférés directement en langage SQL) avant de les acheminer au site User Access. Le logiciel User Access traduit les données SAIF/XDR en Terraview afin que l'utilisateur puisse les consulter. Selon la source des données et la nature de la demande, le temps nécessaire pour recevoir l'information désirée varie de moins de dix minutes (pour la topographie) à une nuit (demandes complexes sur les cartes cadastrales ou les terres publiques).

Fait plus important, les données des trois sources sont intégrées au moyen d'un datum commun et rationalisées les unes par rapport aux autres, car en réalité, on procède à la conversion d'un modèle de données à un autre.

4. LA VERSION COMMERCIALE DE LANDDATA BC (AOÛT 1995 À MARS 1997)

La version commerciale de LandData BC portera le nombre d'interrogations simultanées de 12 à 500 grâce au logiciel User Access qu'on pourra installer sur un ordinateur personnel en environnement Windows ou NT, sur un ordinateur Macintosh et sur une plateforme Unix. Le serveur de recherche de la version commerciale pourra analyser et acheminer une simple interrogation vers une quantité de produits de données. Par exemple, une simple interrogation sur la propriété sera acheminée le registre des terres publiques, les titres miniers, les ententes d'exploitation forestière et les réserves de terres agricoles sans obliger l'utilisateur à effectuer un grand nombre de choix. On ne sait toujours pas si les demandes de renseignement spéciales seront acceptées. Les commentaires des utilisateurs donnent à penser que l'investissement nécessaire pour cela ne profiterait qu'à une poignée de *puissants* utilisateurs et pourrait soulever des préoccupations auprès des conservateurs de données quant à la prise de mesures de protection supplémentaires.

Une autre douzaine de produits de données «internes» pourront être examinés en ligne. Pour cela, ces produits devront cependant se conformer davantage aux normes et présenter une plus grande spécificité afin de se plier aux exigences rigoureuses du serveur Data Access.

En juin 1995, nous avons obtenu des fonds de 7,4 millions de dollars pour mettre au point la version commerciale de LandData BC avec Macdonald Dettwiler (MDA). Cette somme était nettement inférieure aux 25 millions requis pour la période de trois à cinq ans prévue, mais la situation n'est pas aussi désastreuse qu'on pourrait le croire. En effet, 60% des fonds réclamés devaient aider les conservateurs des données susceptibles d'être consultées en direct à adapter celles-ci au format SAIF, à effectuer une mise à niveau en fonction des exigences minimales du système (à savoir NAD 83) et à effectuer des essais d'assurance de la qualité pour vérifier la cohérence interne des données et l'intégrité des références, de façon à garantir la capacité d'intégration et l'interopérabilité pour l'utilisateur final. Dans les circonstances, les conservateurs de données devront simplement entreprendre ce travail à leurs frais et, en fin de compte, il ne faudra qu'un peu plus de temps avant qu'on puisse accéder aux données en leur possession pour une gamme de produits particulière.

Le plan de mise en oeuvre en deux temps de LandData BC (gouvernement d'abord, population ensuite) sera modifié en fonction de la loi provinciale sur l'accès à l'information et des considérations d'ordre pratique concernant l'échange d'information entre le gouvernement et l'industrie forestière aux termes du nouveau code de pratiques forestières.[10] Pour l'instant, c'est le gouvernement qui développe, donc exploitera LandData BC, mais on espère que les recettes réalisées rendront le projet autonome. Nous avons préféré une stratégie d'expansion en spirale plutôt qu'une gestation de 30 mois aboutissant à un projet clés en main.

La conception du prototype remontant déjà à plus de deux ans, Macdonald Dettwiler passera la technologie en revue pour nous. Avant de retenir Sybase, la dernière fois, nous avions envisagé un gestionnaire de base de données avec orientation objets ou un hybride, mais en fin de compte, les programmes de ce genre semblaient présenter trop de risques dans un environnement presque opérationnel. Aujourd'hui, non seulement souhaitons-nous réfléchir à nos décisions antérieures, mais la technologie ayant progressé, il convient aussi de déterminer quels développements sont survenus dans l'intervalle et d'envisager les futures architectures que laissent présager les systèmes d'information géographique avec objets, l'intégration d'objets, le langage SQL3 et le reste.

5. L'AVENIR

5.1 Évolution d'internet

L'examen de la technologie[11] révèle que les services articulés sur Internet comme les fureteurs World Wide Web et les serveurs FTP se sont affinés, ce qui réduira les frais de développement de la version commerciale de LandData BC.

Parallèlement, les services Internet n'ont pas montré qu'ils pouvaient soutenir le visionnement et la recherche des données spatiales, et encore moins certains aspects commerciaux comme la comptabilité et la sécurité. On s'inquiète aussi du peu d'uniformité des services Internet.

5.2 Nouveau rôle des serveurs de bases de données spatiales

Il fut un temps où nous pensions que les SIG constituaient la technologie idéale pour les questions relatives aux données spatiales, qu'il s'agisse de leur saisie, de leur analyse ou de leur gestion. À mesure que la taille et la complexité des bases de données augmentent cependant, beaucoup abandonnent maintenant les SIG et se tournent vers des gestionnaires de base de données au plein sens du terme pour gérer leurs données. En effet, les SIG manquent de finesse pour une gestion formelle des données spatiotemporelles. Les grands systèmes de gestion de base de données ne cessent de s'améliorer afin d'accepter le traitement des données spatiales.

Ainsi, un de nos conservateurs gère au-delà de 5 200 fichiers de données topographiques numériques sur la Colombie-Britannique occupant 20 gigaoctets de mémoire. Cette base de données devrait passer à 7 000 fichiers et 25 gigaoctets au terme du projet, l'an prochain. La base de données est exploitée au moyen d'un SGBD relationnel. Malgré cela, l'intéressé a éprouvé des:

- problèmes de gestion en raison des instruments rudimentaires utilisés pour effectuer des mises à jour chirurgicales et des sauvegardes

- des problèmes de rendement.

Les programmes axés sur la recherche d'objets pourraient atténuer le second problème dans l'avenir.

TABLEAU 1

Produit de données	Type d'archivage	Plateforme ordinateur/logiciel
Topographie (graphique) Cartes cadastrales (graphique) Terres publiques (texte)	SAIF/XDR Arc/Info Base de données	VAX / RdB SUN/ESRI IBM 4300/SQLDS

6. UN REGARD SUR LE PASSÉ

6.1 Application du SAIF sur XDR

En recourant au SAIF, la province souhaitait définir d'emblée un format de transfert sans lien avec le vendeur, de façon à assurer la portabilité maximale des données spatiales entre technologies exclusives. À un moment quelconque en 1992, dans notre empressement à rendre nos données topographiques accessibles afin de démontrer l'utilité du prototype de LandData BC, nous avons effectué un détour et essayé d'adapter le SAIF sur XDR, un logiciel SUN du domaine public. Les problèmes ont surgi lorsque nous avons élargi le logiciel XDR en vue d'en accroître l'efficacité aux fins envisagées. En nous écartant de la norme XDR, nous avons renoncé à la capacité à long terme de recourir aux outils normalisés.

Les problèmes de soutien n'ont pas manqué de suivre cette décision qui a compliqué exagérément la tâche de ceux qui devaient mettre au point les traducteurs SAIF pour les SIG exclusifs.

L'an dernier, nous sommes revenus sur notre décision antérieure concernant XDR et avons choisi une simple application OSN reposant sur le programme PK-ZIP. Ce choix se reflète dans la version 3.2 de la spécification SAIF (URL - http://www.env.gov.bc.ca/gdbc/saif32/toc.html). En septembre 1994, nous avons reçu une trousse d'utilitaires SAIF (interface API en environnement C ou C++) destinée à ceux qui élaborent des traducteurs pour SIG. D'ici février 1996, nous devrions disposer de traducteurs SAIF bidirectionnels, articulés sur le jeu d'utilitaires précité, ce qui permettra la traduction des deux principaux produits de SIG largement utilisés dans la province dans les deux sens.

Le SAIF sur XDR avait été appliqué aux serveurs Data Access pour les trois produits de données accessibles en direct au moyen du prototype. Ces produits ont dû être codés de nouveau. En effet, les données topographiques étaient les seules à être archivées en SAIF/XDR et 3 600 fichiers ont dû être convertis en SAIF/OSN.

6.2 Dépôts

Les dépôts jouent un rôle opérationnel (gestion des différentes versions) et stratégique (analyse des divergences en vue de la planification des programmes) au niveau de la gestion des données. Le prototype de dépôt de LandData BC répertorie les sources et les produits de données pour les utilisateurs de l'infrastructure de services. LandData BC ne garde aucun produit. Le dépôt n'est en réalité qu'un *menu* permettant à l'utilisateur en bout de ligne d'effectuer une commande. Le dépôt de LandData BC comprend aussi des tables schématiques donnant accès en direct aux produits de données.

D'après l'expérience acquise, une infrastructure de services d'information ne proposant que la consultation en direct d'un index des sources de données peut présenter une certaine utilité pour les personnes à la recherche de données, au départ. Cependant, pareil service perd rapidement sa valeur une fois que l'utilisateur final acquiert de l'expérience et découvre comment circonscrire ses recherches. À cet égard, un dépôt s'avère moins utile, et de loin, qu'un annuaire téléphonique. Pour survivre, l'infrastructure de services doit permettre l'acquisition de données, un peu comme un service de commandes postales.

On pourrait pousser les choses un peu plus loin en supposant que l'utilisateur final qui possède le matériel et les connaissances requis pour entreprendre une recherche en direct s'attend à obtenir rapidement une réponse. Il s'ensuit que la version commerciale de LandData BC devra offrir plus de produits de données en ligne. Lors du développement du prototype de l'infrastructure de services, je pense que nous avons pris une sage décision en proposant les services d'archivage, de commande, de livraison en direct et d'intégration des données. Les clients potentiels ont ainsi l'impression d'accéder à un guichet unique.[12]

Avec le prototype, les métadonnées étaient colligées par les conservateurs de données, qui devaient compléter des formulaires et les renvoyer à LandData BC. La saisie dans la base de données Sybase était effectuée par le personnel administratif. Les mises à jour suivaient un processus identique.

La version commerciale de LandData BC doit comprendre les outils nécessaires à une décentralisation et (ou) à une semi-automatisation de la gestion des métadonnées, des diagrammes et des modèles de données par les conservateurs de données. L'idéal serait que LandData BC aiguille l'utilisateur vers le dépôt original. À longue échéance, la tâche du personnel de LandData BC consistera à veiller au respect des spécifications et des politiques relatives aux dépôts.

6.3 Politiques

Qui y a-t-il de plus naturel pour un technocrate que des politiques? En voici un exemple.

Depuis 1992, le gouvernement de la Colombie-Britannique poursuit une politique fixant le prix des données.[13,14] Cette politique part du principe fondamental que les données constituent une ressource précieuse, au même titre que les fonds et les années-personnes, pour la prestation des services gouvernementaux. Pour le souligner, on a autorisé les ministères à fixer un prix pour leurs données, en fonction des critères suivants:

- coût de la colligation des données;
- estimation de l'investissement nécessaire pour garder les données à jour;
- mesure dans laquelle les données sont colligées pour appuyer les activités d'une organisation ou, inversement, estimation de la valeur résiduelle des données pour autrui;
- estimation du coût marginal de la diffusion des données.

Cette politique est valable pour tous les clients extra-ministériels du gouvernement provincial ou d'ailleurs et entraîne un transfert de revenu des Recettes générales au conservateur de données, ce qui compense en partie le coût du maintien des données. Ainsi, on freine la dévalorisation des données et encourage les conservateurs à maximiser l'usage résiduel des données.

Il s'agit d'une politique importante, car elle reflète la valeur des données pour l'utilisateur final et des efforts déployés par celui qui a la garde des données. En outre, la valeur réelle des données échangées démontre en soi le bien-fondé d'une infrastructure de services.

La politique de fixation des prix illustre parfaitement la multitude de décisions en matière de politiques auxquelles il faudra constamment veiller et donne un aperçu des décisions qui attendent encore. Dans certains cas, on se bornera à suivre d'un oeil intéressé d'autres tables rondes sur les politiques. Au Canada, le secteur des télécommunications, actuellement très réglementé, fait l'objet d'une **déréglementation**. La période de transition plonge les infrastructures de services dans l'incertitude à l'égard des tarifs et de la télétransmission des données. Comme cela se produit aux États-Unis, les personnes responsables de la déréglementation sont en train de jeter les bases de la concurrence au niveau des infrastructures matérielles.

6.4 Conception descendante, mise en oeuvre ascendante

Les infrastructures de services sont des projets de longue haleine qui exigent un travail considérable et éventuellement, des années s'écouleront entre les *esquisses* et la réalisation. À l'instar de tout autre projet d'envergure, le nôtre reste vulnérable à l'évolution de la technologie et au déplacement de la cible, attribuable aux changements que traversent les activités commerciales de la clientèle et le milieu technologique.

Une infrastructure de services qui procure l'accès à des sources de données décentralisées doit pouvoir s'adapter à la variabilité de ses fournisseurs. Si on voit dans l'imposition de normes la panacée aux problèmes, il n'en demeure pas moins que le respect des normes ou la capacité à s'y conformer varient, à moins qu'il ne s'agisse de normes minimales.

Il ne faut pas perdre de vue qu'au moment précis où il bâtit une infrastructure de services selon un modèle de données particulier, l'architecte du projet doit déjà y envisager des modifications. Pour bon nombre de conservateurs de données sur les terres, les modèles contemporains ne sont en réalité que le squelette des bases de données de l'avenir qui attendent d'être modélilsées. Les modèles de données homogènes pour un ensemble de données particulier constitueront plus l'exception que la règle.

Un développement en spirale permettra la prestation de services ou d'un niveau de services durant la vie entière du projet. On pourra donc réaliser des bénéfices partiels assez tôt. Cela implique aussi un examen ou un déplacement des services offerts à l'aube du projet, face à ceux proposés à sa conclusion. Pour donner de bons résultats, pareille stratégie exige un cadre théorique très rigoureux.

Il ne faut pas oublier que la production d'une version commerciale de LandData BC n'aboutira jamais à un produit fini, prêt à être utilisé. Il ne s'agira en l'occurrence que du point de départ d'extensions et de modifications constantes. Dans un contexte évolutif comme celui-là, la pérennité de l'infrastructure de

services dépendra de l'adhésion à des normes d'architecture reconnues. On peut suivre les progrès de LandData BC au site http://www.lii.crl.gov.bc.ca.

6.5 Des données, encore des données, toujours des données

Pour paraphraser le cri de ralliement des courtiers en immeubles au sujet de l'emplacement d'une propriété, la clé du succès réside dans les données. Avant tout et surtout, cela signifie gagner le soutien des conservateurs de données. Au gouvernement, ceux qui inventorient les données et ceux qui érigent les infrastructures de services obtiennent leurs fonds de la même source. Ainsi, le prototype de LandData BC a reçu sa sanction officielle en 1990, à l'époque où l'économie provinciale avait atteint un pic, et bénéficiait d'un fonds spécial. Bref, il ne livrait pas concurrence à nos fournisseurs potentiels de données en ce qui concerne le financement. Quand est venu le temps d'obtenir l'autorisation nécessaire pour produire la version commerciale, les perspectives économiques de la province avaient changé, le fonds spécial s'était évanoui et nos fournisseurs critiquaient ouvertement tout projet ayant pour but d'ériger une infrastructure à leurs dépens, au moment où ils étaient eux-mêmes ensevelis sous les problèmes créés par l'exploitation des terres et les revendications territoriales des Autochtones. Certains ont admis qu'une telle infrastructure s'avérerait nécessaire au bout de trois à cinq ans, mais pas dans l'immédiat. Notre position reposait sur le temps que requérait l'implantation d'une telle infrastructure au bout de trois ans et sur le fait que les données saisies par un organe du gouvernement aujourd'hui devaient pouvoir être réutilisées par d'autres subséquemment. Finalement, des fonds ont été débloqués pour ces deux volets. Toutefois, il reste encore du travail à faire pour mériter la confiance des fournisseurs de données.

L'eldorado que chacun convoite est la portabilité et l'interopérabilité des données - c'est-à-dire pouvoir accéder de partout aux données gardées n'importe où, pouvoir les traiter partout et pouvoir les intégrer à d'autres données. Les gens ont tendance à voir dans les systèmes ouverts des normes applicables à l'industrie du matériel et du logiciel. Si c'était le cas, le problème paraîtrait soluble, à la limite. Le principal obstacle qui nous empêche d'atteindre cet eldorado correspond aux données et aux normes qui s'y rapportent. Or, cet aspect se trouve sous le contrôle des institutions qui rassemblent, gèrent et traitent d'énormes quantités de données spatiales.

Les traducteurs SAIF ne règlent qu'une des contraintes relatives à un échange et à une intégration des données à toute épreuve, en tant que logiciel «encadrant» les technologies de SIG exclusives. Ils exigent l'installation d'un modèle de données aux deux sites avant que l'échange puisse débuter. La conversion d'un modèle à l'autre constitue une étape cruciale, qui n'est pas spécifique aux environnements multi-vendeurs, et le SAIF propose des méthodes normalisées pour la modélisation des données spatiales et non spatiales.

Comme on peut le constater dans le tableau qui précède, les traducteurs et les modèles de données n'élimineront pas les problèmes que pose le respect des spécifications. Si les données d'entrée ne se conforment

Aspect du problème	Nature du problème	Approche/solution
Interopérabilité du matériel et du logiciel	Technologie exclusive	Systèmes ouverts; traducteurs SAIF et modélisation des données
Normes de la discipline (p. ex., inventaire des forêts)	Multiplicité des définitions des données chez les organismes qui échangent les données	Consensus des spécialistes de la discipline; fédérations
Respect des spécifications	Manque d'uniformité des données d'un organisme particulier par rapport aux spécifications publiées	Respect des spécifications relatives aux données (y compris assurance de la qualité); procédés formels de gestion des données permettant d'associer les données aux diverses versions des spécifications

pas aux spécifications, il se pourrait qu'on n'obtienne pas les résultats escomptés, mentionnés dans la publicité, au site de l'utilisateur. Ce n'est qu'une fois les données transmises à d'autres utilisateurs qu'on peut vraiment en jauger la qualité. Aussi étrange que cela paraisse, avec les données, les erreurs que signalent les utilisateurs pourraient être perçues comme un prolongement du programme d'assurance de la qualité de l'organisation. D'après nos observations, le respect des spécifications demeure un objectif louable, réalisable à long terme.

BIBLIOGRAPHIE

[2] Building the GIS Infrastructure in British Columbia, Sawayama, G., *Proceedings of the Canadian Conference on GIS*, mars 1992, 1014-1024.

[3] Infrastructure Information Requirements in the Maritime Provinces: An Analysis, Hamilton, A.C., Department of Surveying Engineering, U.N.B., juin 1976, 3-6.

[4] Creating a Government that Works Better and Costs Less, Report of the National Performance Review, Vice-président Al Gore, 168 pp, page 116.

[5] British Columbia Employment Dependencies, rapport final rédigé pour la British Columbia Forest Resources Commission, Horne, G. , et Penner, C., février 1992, 39 pages.

[6] The Future of Our Forests, Forest Resources Commission, avril 1991, 97 pp.

[7] State of the Environment (1993). Report for British Columbia, 127 pp.

[8] Land Claims reference.

[9] Ministry of Environment Geographic Information System, Rapport final: Description of Physical Prototype, DMR Group Inc., mai 1990, 150 pp., pages 51-72.

[10] British Columbia Forest Practises Code, Standards with Revised Field Guide References (1994), 216 pp.

[11] LandData BC Technology Review Draft, Macdonald Dettwiler, septembre 1994.

[12] LandData BC Prototype System Review Draft, Evert Kenk, Surveys and Resource Mapping Branch, BC Ministry of Environment, Lands and Parks, 59 pp., page 41.

[13] Document de travail : Pricing and Distribution of Digital Land Information, Forum Consulting Group for BC Ministry of Crown Lands, avril 1990, 33 pp.

[14] Ministry of Lands and Parks et Ministry of Forests Digital Data Marketing Procedures, octobre 1991, Forum Consulting Group for Ministry of Lands and Parks, 69 pp.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

MEILLEUR SERVICE ÉLECTRONIQUE À LA CLIENTÈLE : LE PROJET STATCAN EN DIRECT

R. Grenier[1]

INTRODUCTION

StatCan en direct est, comme son nom l'indique, un service en direct qui est contrôlé et géré par Statistique Canada, de concert avec une entreprise du secteur privé spécialisée en technologie.

Aujourd'hui, StatCan en direct est accessible dans tous les pays du monde qui disposent d'un réseau informatique public.

En 1996, il sera possible d'accéder à StatCan en direct à partir d'Internet. D'ici là, nous souhaitons que nombre des problèmes actuels liés à l'Internet auront été résolus, notamment ceux de sécurité, de fiabilité et d'accessibilité.

OBJECTIFS

Les objectifs établis à l'origine pour ce projet expliquent les choix que nous avons faits en matière de conception technique, de données et de métadonnées, de marketing et de logistique.

Ces objectifs serviront de fil conducteur quant aux buts et aux défis permanents de la logistique, étant donné que certains ne seront jamais pleinement réalisés.

1. Améliorer l'accessibilité, la pertinence et la facilité d'utilisation des données de SC.

    - Accessibilité - 24 heures par jour, sept jours sur sept à partir de votre bureau, de votre foyer ou de votre chambre d'hôtel.
    - Utilisation facile des données - non seulement l'interface logiciel, mais une compréhension plus globale de l'information que nous produisons, grâce à des métadonnées complètes et à des services créatifs d'intégration et d'interprétation, ainsi qu'au recours à des spécialistes de SC.

2. Augmenter l'étendue et le détail des informations disponibles.

    - Nous ne pouvons nous permettre de publier, sur papier, qu'une fraction de nos fonds documentaires. Cette situation s'aggrave au fur et à mesure que les coûts augmentent, particulièrement ceux du papier. Toutefois,avec la diffusion en direct le niveau de détail n'est limité que par les contraintes de confidentialité.

    - On continuera toujours de publier des données en différé, les sources et la façon de faire étant décrites dans le service en direct.

3. Réduire le coût à l'unité pour les clients.

    - Il s'agit non seulement du coût réel de l'information, c'est-à-dire qu'un abonné ne paierait que pour les données dont il aurait besoin, et non pas pour un ensemble complet, mais aussi d'économies pour les organismes clients grâce à un emploi plus efficace du temps de leur chercheurs.

4. Améliorer notre connaissance des besoins et des intérêts des clients.

    - et par conséquent la *pertinence* des données que nous recueillons et des *produits* et *services* que nous offrons.

    - Ces éléments reposent sur un système d'information de gestion (SIG) très complet

[1] Ross Grenier, Directeur, Projet StatCan en direct, Statistique Canada, Ottawa (Ontario), K1A 0T6.

ainsi que sur un système de messagerie électronique grâce auquel les clients peuvent communiquer avec nous et réciproquement. Ils nécessitent en outre un personnel des ventes bien informé, un service d'aide efficace ainsi que des services de soutien spécialisé et technique.

5. Améliorer le rapport coûts/recettes comparativement à celui des méthodes actuelles de diffusion.

   - Par une réduction des coûts et la production de recettes.

6. Concevoir un service dynamique et solide, qui peut être élargi et amélioré facilement.

   - Pour être concurrentiel, il faut avoir la capacité d'adopter les nouveaux développements aux systèmes électroniques d'information.

   - Multi-média, gamme variée de plate-formes GUI, interconnexion, communications sans fil, tarification au détail, systèmes d'information géographique, etc.

   - Parallèlement, le contenu et les aspects fonctionnels doivent aussi évoluer pour répondre aux besoins des utilisateurs.

7. Mener à bien les objectifs énoncés précédemment sans compromettre la qualité des données.

MARKETING

Stratégie de marketing

- **Études de marché**

  - 9 groupes de discussion
  - Enquête téléphonique
  - Essai sur le terrain avec des données sur le commerce (questionnaire d'évaluation)
  - Test de marketing avant le lancement (170 abonnés pour un an)
  - Essai sur le terrain - CANSIM (questionnaire d'évaluation)

- **Analyse des besoins**

  - Marchés visés - les besoins et non les produits
  - SIG - identifier qui achète quoi, en quelles quantités, quand, en association avec quoi, etc.
  - Rétroaction du client sur support électronique, grâce au service d'aide, aux chargés de comptes.

Stratégie d'établissement des prix

- Aucun frais de raccordement au Canada
- Le client ne paie que pour les données qu'il reçoit
- Accès gratuit au *Quotidien*
- Un mois gratuit pour les abonnements annuels, autrement : 25 $ par mois.
- Accès spécial pour les médias sans frais d'abonnement
- Remise pour gros utilisateurs de données
- Remise pour les établissements d'enseignement

DÉMONSTRATION

On a procédé à une démonstration de StatCan en direct dans le cadre de laquelle on a tiré trois séries chronologiques de la base de données CANSIM.

SESSION 9

Panel

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

ÉVOLUTION DES PARTENARIATS DANS LE SECTEUR DE L'INFORMATION

D. Desjardins[1]

Nous disposons d'un sujet de choix pour la séance finale du Symposium. Nous avons aussi la chance de pouvoir compter sur un groupe d'experts réputés, qui nous feront part de leur expérience et de leur point de vue sur le sujet.

L'évolution des partenariats dans le secteur de l'information, c'est-à-dire le phénomène de la constitution de partenariats, n'est pas un concept nouveau. Il ne s'agit pas non plus d'une tendance passagère. Il faut plutôt y voir une nouvelle façon de faire qui est là pour rester.

Un bureau national de la statistique peut facilement être perçu comme un service d'information. Son mandat premier est de fournir des renseignements sur la situation sociale et économique des citoyens qu'il sert. La réalisation de ce mandat n'exclut toutefois pas la participation d'autres intervenants. Il est possible de tirer partie de l'utilité du fonds documentaire d'un organisme de nombreuses façons, profitables pour d'autres, par exemple, quant à la façon dont les données sont fournies, présentées ou transformées en d'autres données.

Aujourd'hui, les experts renommés de notre groupe parleront de ce phénomène et examineront certaines de ses caractéristiques. Ils nous feront part des progrès réalisés dans le secteur privé, dans le secteur public ainsi qu'aux niveaux provinciaux et municipaux. Ils nous entretiendront enfin des progrès réalisés au niveau international.

Ces partenariats comportent des avantages et des inconvénients. Parmi les avantages, on note une utilisation plus large de l'information et la contribution à l'établissement d'une industrie prospère de l'information au Canada et, enfin, un meilleur service à la clientèle.

Parmi les aspects négatifs figurent l'impression que ces partenariats constituent une utilisation non autorisée de données obtenues aux frais du public. On peut penser que l'octroi de licences au secteur privé pour utiliser les fonds documentaires publics peut donner lieu à une intrusion dans la vie privée, comme dans le cas du marketing direct.

On peut se poser la question suivante : quel est le prix juste de l'utilisation commerciale des données recueillies aux frais du public? Certains penseront qu'elle devrait être gratuite, étant donné que l'on a déjà payé pour ces données. D'autres prétendront qu'un bureau de la statistique n'a tout simplement pas sa place sur le marché.

Même si ces arguments se contredisent, ils sont légitimes. Ce matin, vous entendrez toute une gamme de points de vue exprimés par un groupe d'experts renommés qui, ensemble, comptent une centaine d'années d'expérience dans le secteur de l'information, expérience qui prend diverses formes.

---

[1] Denis Desjardins, Statistique Canada, Immeuble 10-A R. H. Coats, Parc Tunney, Ottawa (Ontario), Canada, K1A 0T6, internet desjden@statcan.ca.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

PANÉLISTE

J. Kestle[1]

Tout d'abord, j'aimerais vous dire combien j'apprécie d'avoir été conviée à me trouver parmi vous, aujourd'hui. Denis a parlé tout à l'heure de ma longue carrière dans la statistique au gouvernement de l'Ontario et des très nombreuses journées passées avec beaucoup d'entre vous dans cette salle à l'occasion de réunions fédérales-provinciales sur les statistiques. J'apprécie indubitablement tout ce que mon association avec Statistique Canada à l'époque m'a apporté d'enseignements, et me retrouver ici aujourd'hui réveille quelques-uns de ces souvenirs. Comme l'a dit Denis, je vais vous parler du regard de ma société en particulier sur les partenariats avec Statistique Canada.

Tout d'abord, je vais vous expliquer ce que fait Compusearch. Prenez cet exposé comme une étude de cas destinée à vous faire comprendre comment une entreprise du secteur privé utilise des renseignements, dont certains fournis par Statistique Canada.

Donc, quelle est la nature de notre entreprise ? Eh bien, nous fournissons des réponses par l'analyse de marché. Certaines personnes pensent que nous vendons des données, mais en fait, ce que nous faisons pour l'essentiel, c'est travailler de concert avec des services de marketing pour les aider à répondre aux questions suivantes : qui sont leurs clients, où habitent-ils et comment en apprendre plus sur eux ? Fournir des données brutes au marché représente, en réalité, une part plus petite de nos activités, et la plupart de ces données que nous fournissons sont destinées aux ensembles cartographiques du Système d'information géographique (SIG).

Juste pour vous donner un peu plus de détails à ce sujet, je vais vous présenter un échantillonnage très rapide du type de travail qu'une entreprise pourrait nous demander d'effectuer. Voici que ce qu'on appelle le graphique en barres d'une grappe. Ce que nous faisons, c'est classer la population en catégories. Vous auriez des tas d'ennuis si vous le faisiez comme nous, mais nous le faisons pour des entreprises privées en utilisant des statistiques. Ce que nous faisons, donc, c'est apposer une étiquette, qui correspond à l'un des soixante types uniques que nous classons, sur toutes les localités du Canada. Puis nous utilisons ce classement pour aider nos clients à comprendre qui sont leurs clients. C'est ce que nous appelons le profil caractéristique pour un produit assez haut de gamme. Je vous expliquerai plus loin ce dont il s'agit. Et voici une représentation graphique destinée à vous montrer que, sur les soixante segments de la population canadienne, certains sont surreprésentés sur le marché pour ce produit en particulier tandis que d'autres sont sous-représentés.

Pour arriver à cette information, nous utilisons des données tirées de recensements, mais nous nous servons également de renseignements que nous achetons au ministère des Transports. Nous employons des renseignements que certains de nos gros clients, les éditeurs par exemple, nous fournissent sur leurs clients. Donc, nous rassemblons différentes données et nous procédons à ce que vous connaissez bien, c'est-à-dire à une analyse typologique. Nous découvrons ainsi que les sous-groupes de population qui sont surreprésentés pour tel produit sont les technocrates et les bureaucrates, la petite élite urbaine, les érudits vieillissants, les classes maternelles en pleine expansion et ainsi de suite. Donc, nous avons ces groupements qui reposent sur le comportement d'achat, sur les facteurs démographiques et sur diverses autres statistiques. Nous aidons nos clients à cerner leur clientèle par rapport à ces critères.

Nous pouvons ensuite relier ces données à d'autres données qui nous permettent de dire que les gens qui achètent ce produit sont des gens qui, par exemple, reçoivent probablement chez eux deux ou trois fois par mois, qui jouent au golf, dînent dans des clubs privés, sont membres de clubs de santé, et qu'en tout cas, ils ne jouent pas aux quilles et ne tricotent pas. Ces gens aiment conduire des Jeep Cherokee, parce que leur profile correspond à ce type de voiture. Ils conduisent des Cherokee ou des Pathfinder, et ce ne sont pas vraiment de gros clients pour les Hyundai et Geo.

---

[1] Jan Kestle, Compusearch, Compusearch, 230 rue Front West, Suite 1100, Toronto (Ontario) M5V 3B7.

Il s'agit juste d'un tout petit exemple, pour vous donner une idée de ce que nous faisons. Nous ne vendons pas de données brutes. Nous prenons des données que nous combinons avec un tas d'autres renseignements afin de répondre à ces questions pour nos clients. Et, enfin, à partir de ce profil, nous pouvons transposer les données sur une carte.

Nous clients sont les banques, les sociétés d'assurance, les magasins de vente au détail, les entreprises du secteur des produits emballés, les entreprises médiatiques, les services de télécommunications, les organismes sans but lucratif, les constructeurs d'automobiles et les fabricants de biens de consommation durables. Nous renseignons également des fournisseurs de logiciels spécialisés, des experts conseils, des groupes qui intègrent des données. Plus de 4 500 clients nous achètent ces produits spécialisés.

La société a été fondée en 1974 par un entrepreneur. Nous avons parfois l'impression de faire figure de grand méchant loup dans ce secteur de l'industrie de l'information à cause de notre taille. Mais je tiens à préciser que nous avons commencé tout petits. Un certain Bill Goldstein, que quelques-uns d'entre vous connaissent, a démarré cette entreprise à l'arrière d'un camion, en 1974, et à l'époque, il avait une clientèle de détaillants. Il concevait des logiciels qui permettaient de trouver des repères. Ces programmes pouvaient réunir des données de recensement et dire aux gens quel type de population vivait dans leur secteur commercial. Il a déplacé son entreprise à Toronto et il est devenu le premier revendeur du recensement canadien. Le Blackburn Group de London (Ontario), qui est un important fournisseur d'information, a racheté l'entreprise en 1985, avant de fusionner, en 1994, avec la R.L. Polk Company, qui produit des statistiques sur l'automobile dans le monde entier. Nos activités d'analyse de marché en Amérique du Nord se chiffrent actuellement à une trentaine de millions de dollars. Ce qui ne signifie pas que nous vendons pour 30 millions de dollars de données de Statistique Canada. En fait, d'après nos derniers calculs, 2 % environ de cette somme peuvent être attribués à la redistribution directe de données brutes de Statistique Canada.

Nous participons à de nombreux partenariats importants et, juste pour vous donner une idée, nous obtenons d'un très grand nombre d'organismes des données qui entrent dans la création de ces outils dont je vous ai donné un aperçu. Cependant, notre principal partenaire reste Statistique Canada.

La nature de notre partenariat avec Statistique Canada est, à notre sens, la suivante : Statistique Canada produit des données de la meilleure qualité qui soit. Vous ne le dites pas assez souvent. Il vous arrive de le dire, mais nous répétons dans toutes les conférences internationales que nous avons beaucoup de chance au Canada d'avoir un fournisseur de données tel que Statistique Canada. Nous avons affaire à des entreprises américaines et européennes qui n'ont pas accès à des données du type et de la qualité de celles que vous nous fournissez. Dans ce partenariat, Statistique Canada soutient ces produits. Il favorise un échange d'informations professionnel. Compusearch emploie beaucoup de démographes et de statisticiens, c'est-à-dire beaucoup pour nous, soit 35 à 40 personnes qui sont des spécialistes de domaines techniques. Des personnes qui sont en communication sur des questions de statistique et de données avec la Division du recensement, la Division de la démographie et la Division des enquêtes-ménages, avec les fournisseurs de données commerciales et avec les fournisseurs de statistiques sur l'équité en matière d'emploi, pour n'en nommer que quelques-uns. Nos relations avec Statistique Canada, reposant sur le contenu de l'information et les besoins de nos clients, sont variées et profondes. Et nous apprécions la valeur que Statistique Canada leur accorde, ainsi que le professionnalisme dont il fait preuve dans l'échange d'informations avec l'entreprise privée que nous sommes.

Enfin, Statistique Canada est un excellent fournisseur, et vous savez que nous avons beaucoup de fournisseurs. Il apprécie notre clientèle, et nous considérons cette relation comme étant manifestement le partenariat le plus important que nous ayons.

Ce que nous apportons dans le partenariat, c'est une valeur ajoutée aux données de Statistique Canada. Et j'entends par là différentes choses. Cela signifie prendre ces données, les ajouter à d'autres recueillies dans le secteur privé ou dans d'autres ministères, et créer des produits statistiques et d'analyse de marché dont nos clients ont besoin. Mais cela signifie aussi découper les données. Vous savez, contrairement à bien des gens, la somme de travail qu'il faut juste pour fournir à quelqu'un, en une demi-heure, des données sur les origines ethniques dans les secteurs de recensement. Ou, ce travail fait, pour mettre les résultats sur une disquette, sous le format voulu, quand un client préfère les fichiers numériques des limites au format ASCII ou demande un format compatible avec Mapinfo. Il s'agit là d'une tâche à valeur ajoutée qu'il faut accomplir pour envoyer les données. Il faut voir les instructions que l'on nous donne pour l'envoi des données : nous devons les fournir en toutes sortes de formats. Nous en arrivons même parfois à la situation paradoxale où il nous revient à moins cher d'envoyer plus de données que moins de données. Vous connaissez le problème, mais bien des gens ne le

comprennent pas. Cela fait partie de la valeur ajoutée qu'une société comme Compusearch peut apporter à ce partenariat.

Nous recommandons également l'utilisation des données et nous nous efforçons d'éduquer les utilisateurs. Nous pouvons contribuer à définir les besoins du marché et communiquer ces données à Statistique Canada. Dans notre partenariat, il y a évidemment le fait que nous devons payer. Et nous versons à Statistique Canada et à nos autres fournisseurs de données des sommes importantes. En résumé, telle est à nos yeux la nature du partenariat et de ce que chaque partie lui apporte.

Pour une entreprise comme la nôtre, ce partenariat présente l'avantage suivant : nous pouvons offrir un guichet unique aux clients qui ont besoin de données. Je m'attarderai un peu plus ce matin sur le fait que nos clients deviennent de plus en plus exigeants en matière de service. Comme les entreprises compriment leurs dépenses, leurs services de marketing suivent le mouvement et la bonne vieille recherche autonome de données, ainsi que la préparation qui l'accompagne, disparaît. Les gens n'achètent plus de données à la tonne et n'embauchent plus de personnel pour les intégrer et les rassembler. Ils veulent que quelqu'un d'autre le fasse pour eux. Ils veulent que cela soit fait par des spécialistes formés parce que les gens qui font ce genre de choses coûtent cher. Il nous faut donc réunir des données provenant de sources variées.

Ce qu'il y a de formidable, entre autres, dans l'intégration de données, c'est qu'elle permet de garantir l'intégrité de ces dernières. Parce que, lorsque l'on réunit des données provenant de différentes sources, cela facilite le processus de contrôle de la qualité. Nous pensons que la présence d'entreprises du secteur privé dans ces partenariats contribue à élargir le bassin de compétences. Lorsque les gens achètent des données et ne savent pas quoi en faire, les données prennent mauvaise réputation parce que la recherche tourne mal. Donc, le rôle du partenariat est en partie d'éduquer les clients, ce qui signifie, et n'y voyez aucun irrespect à l'égard de quiconque, qu'un homme ou une femme qui met au point un logiciel dans son sous-sol et qui prend un paquet de données de recensement ou de données de Compusearch et les rassemble à l'aide de ce logiciel n'est peut-être pas la personne la mieux placée pour aider une entreprise à concevoir son programme d'analyse de marché. Et ceci, parce que si l'on utilise mal les données et que l'on obtient des résultats erronés le bruit court sur le marché que les données de Statistique Canada ou de Compusearch ne valent rien.

Les entreprises qui vont manipuler des données doivent s'engager à former et à soutenir leurs clients en ce qui concerne l'utilisation de données statistiques et les problèmes qui entourent cette utilisation.

Le secteur privé a aussi quelque latitude par rapport aux statistiques. Nous pouvons créer des données approximatives ou à 70 % exactes, et des mercaticiens s'en contenteront, tandis qu'un bureau de la statistique s'efforce de produire des données aussi correctes que possible sur le plan statistique. Voilà qui entre dans le rôle d'une entreprise qui s'occupe de distribution de données. Nous répondons aux besoins du marché, et nous sommes capables de collaborer avec Statistique Canada afin de comprendre ce dont le marché a besoin et d'accroître la demande globale d'un produit.

Nous sommes confrontés à des problèmes dans ce partenariat, et il y a des choses dont nous allons parler aujourd'hui, je l'espère. À mon sens, le premier problème auquel nous sommes confrontés en tant qu'entreprise privée partenaire d'un organisme public, c'est l'impression que les données sont trop chères. Et quand mes clients me disent qu'elles sont trop chères, je leur demande toujours combien elles devraient coûter à leur avis. J'ai compris pourquoi les entreprises canadiennes que nous avons comme clients trouvent nos données, nos données respectives, à Statistique Canada et à nous-mêmes, trop chères. C'est à cause des prix qu'elles paient aux États-Unis ou de ce qu'elles en entendent dire. Nous avons une filiale aux États-Unis qui achète toutes les données de recensement et tous les produits géographiques aux États-Unis, et nous payons 70 fois le prix demandé par le U.S. Bureau of the Census pour dix fois moins de données. Il me semble que les Américains vont devoir relever leurs prix parce que savons tous ce que coûtent des données, pas seulement à fabriquer mais aussi à conditionner et à vendre. Cependant, une fois que vous comprenez la situation, vous comprenez le tollé que soulève le prix des données.

D'aucuns pensent que les gouvernements ne devraient pas facturer les données et, comme le rappelait Denis, certains estiment que s'ils les ont déjà recueillies, il n'y a pas de raison de les faire payer encore une fois. Nous savons tous quoi répondre à ce type de raisonnement. Nous repayons parce que conditionner et redistribuer des données coûtent de l'argent. En outre, les gouvernements recueillent des données pour certaines fins et les entreprises veulent les acheter pour d'autres fins, et chaque utilisation suppose tout un traitement préalable. Je pense donc que le recouvrement des coûts est une politique gouvernementale. C'est une réalité dont tous les aspects ne nous plaisent sans doute pas, mais le Canada a fait ce choix, et nous savons, de par nos partenariats statistiques dans le monde entier que de plus

en plus de gouvernements considèrent le Canada comme un exemple à suivre. Quiconque dans le secteur privé estime que ce système va changer, que l'on va abandonner les politiques de recouvrement des frais risque fort, à mon sens, de faire fausse route.

Nous avons un problème à résoudre dans le partenariat, à savoir l'impression qu'il y a concurrence entre les partenaires. Je ne pense pas que nous soyons en concurrence avec Statistique Canada. La part de marché qui nous intéresse tous deux est minime et, chacun respectant l'autre, c'est le meilleur qui l'emportera. On me l'a déjà entendu dire. Les employés de Statistique Canada qui essaient de vendre des données de recensement à une entreprise sont inquiets et nous avons entendu dire que Compusearch a soufflé un contrat à Statistique Canada. Ce que j'en pense, c'est que nous vendons des données à ces clients depuis 1974 et que bien des gens continueront de s'adresser à nous parce qu'ils veulent tout acheter à un même « guichet ». Pour ce qui est de l'ampleur que peut prendre notre collaboration, de la répartition du marché, c'est avec plaisir que nous coopérons dans ce sens avec Statistique Canada.

Ces partenariats posent un autre problème important que je qualifie de problème de détournement de données. Nous devons veiller à l'application des accords. Personne n'a le droit d'obtenir des données gratuitement, mais beaucoup de gens sont d'avis contraire et estiment qu'il n'y a rien de mal à prendre des données et à les copier sur plusieurs machines. Ils pensent qu'il n'y a rien de mal à prendre des données pour les passer au voisin d'à côté. Nous devons donc nous pencher sur les politiques qui régissent les licences, qui garantissent que les gens qui achètent des données aux fins d'un projet ne les communiquent pas à quelqu'un d'autre. Tels sont les problèmes auxquels les partenariats sont confrontés.

Voyons, pour terminer, les tendances futures. Au risque de me répéter, laissez-moi vous dire que nos clients deviennent de plus en plus exigeants. Ils réclament plus de valeur ajoutée, pas moins. Or, le service à la clientèle nous coûte très cher, et vous connaissez aussi la question. Comme je l'ai déjà dit, si la recherche est mal faite, c'est tout le secteur de l'information dont la réputation est ternie. Il nous incombe donc à tous de trouver des mécanismes de distribution qui ne nous mettent pas en porte-à-faux par rapport aux clients.

On ne peut fournir de l'information en laissant aux clients le soin de se conduire en acheteurs avertis parce que ce type de situation finit par se retourner contre le fournisseur. Donc, à l'avenir, il nous faudra tous faire plus pour nous assurer que ces données seront bien utilisées. À cause de la prolifération des ordinateurs personnels, tout le monde fait de l'analyse de données. La plupart des gens ne connaissent rien à la suppression de données. On nous appelle sans arrêt pour nous poser des questions comme celle-ci, qui touche à la protection de la vie privée : « Comment se fait-il que je n'arrive pas à obtenir de données sur le revenu pour ce secteur de dénombrement en particulier ? » Vous savez et nous savons qu'il y a de bonnes raisons à cela. Dans ce pays, nous fournissons les données de manière à protéger la vie privée des Canadiens. Mais si l'on tire du recensement des données sur le revenu et un numéro de ménage, que l'on rassemble les SD d'un secteur commercial et que l'on calcule le revenu moyen, on arrive à de drôles de chiffres si l'on ne comprend rien aux suppressions de données. Cet exemple montre que la prolifération de données signifie que nous devons trouver des moyens d'aider davantage les gens pour nous assurer qu'ils utilisent correctement les données.

Il existe quantité de sociétés virtuelles sur le marché, ce qui signifie que nombre de personnes ont perdu leur emploi, qu'elles ont un esprit d'entreprise, qu'elles créent leur propre affaire et que tout le monde se trouve un créneau. Donc, chacun veut s'associer avec quelqu'un d'autre. Des entreprises s'associent autour d'un projet, c'est un concept de service commercial en quelque sorte, et cela signifie toutes sortes de partenariats compliqués entre fournisseurs de données. Les contrats de distribution, les concessions de licence, deviendront donc plus compliqués, pas moins.

J'ai déjà parlé de l'informatisation généralisée dans l'entreprise, de l'idée que des données sont copiées sur toutes sortes de machines. L'industrie des logiciels réussit assez bien à combattre le piratage, alors que nous n'arrivons pas à arrêter le piratage des données. À franchement parler, je ne sais même pas, en tant que fournisseur de données, si nous allons pouvoir résoudre ce problème et si nous devons le résoudre. Je suis d'avis que les clients devraient acheter des licences de copies multiples ou de mise sur réseau d'entreprise. Je puis vous assurer que tout cela est très difficile à vérifier.

Enfin, pour ce qui est des tendances futures, nous voyons le prix de l'amélioration des données qui ne cesse de gonfler. Comme je le disais plus tôt, les gens ne veulent pas savoir pourquoi ils ne peuvent pas additionner ces chiffres pour calculer le revenu moyen. Ils veulent que quelqu'un y réfléchisse bien, en donne une idée, pas forcément exacte à 100 %, mais que ce chiffre soit la meilleure approximation d'un méthodologiste ou d'un statisticien. Or, pour expliquer, corroborer, ou mettre en garde, et examiner les résultats de la recherche, il faut dépenser beaucoup d'argent en élaboration des données. Et je crois que j'en resterai là.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

PANÉLISTE

D. Roy[1]

Je remercie les organisateurs du Symposium de m'avoir invité à participer à cette discussion. La formation de partenariats est une stratégie clé qu'adoptent tous les intervenants du secteur de l'information.

Dans mes remarques, je décrirai très brièvement :

- l'expérience de Statistique Canada;
- l'idée que se fait Statistique Canada du marché de l'information par suite des récents sondages auprès des clients et des distributeurs;
- notre «VISION» de la diffusion des données de Statistique Canada et du rôle des partenariats;
- les types de partenariats qu'il nous faudrait établir; et enfin
- les prochains pas de Statistique Canada dans ce «meilleur des mondes».

PARTENARIATS - L'EXPÉRIENCE DE STATISTIQUE CANADA

Aujourd'hui, les données de Statistique Canada sont diffusées à grande échelle par des partenaires des secteurs public et privé.

Cette stratégie date de la création de la base de données chronologiques CANSIM qui, depuis 1976, est diffusée par des distributeurs commerciaux. À l'époque, le Bureau était un pionnier de la diffusion en direct.

Au départ, pour Statistique Canada, il s'agissait essentiellement d'ententes non commerciales. Le Bureau ne facturait alors aux distributeurs que les frais de la mise à jour quotidienne. Toutefois, à partir de 1985, à la suite de l'adoption de la politique de recouvrement des coûts, le Bureau a commencé à percevoir une redevance pour l'utilisation des données.

À l'occasion du Recensement de 1986, les ententes commerciales ont été étendues à un petit nombre d'organismes qui revendent les produits ou mettent au point des produits «à valeur ajoutée» - notamment CompuSearch.

Par suite de cet apprentissage et en vue de l'établissement d'une stratégie pour le Recensement de 1991, nous avons sondé le secteur sur les pratiques en cours, consulté les distributeurs actuels et potentiels, puis élaboré la politique de concession de licences et de signature d'ententes avec les utilisateurs finals en vigueur aujourd'hui. À cet égard, nous avons essayé de nous adapter aussi bien aux demandes des gros que des petits revendeurs et «producteurs de valeur ajoutée». (On peut se procurer des copies de ces ententes auprès de la Division du marketing.)

ATTENTES DES CLIENTS ET TENDANCES DU SECTEUR

Comment ces mesures s'accordent-elles avec les attentes des clients et avec les tendances du secteur?

- Aujourd'hui, le secteur est poussé par l'évolution rapide de la technologie. Cette évolution s'articule en grande partie autour d'INTERNET, qui est devenu la porte d'accès courante à des services en direct offrant un contenu dynamique ou en rapide évolution. Durant les rencontres récentes de l'IIA qui ont eu lieu à Toronto, les conférenciers ont tous mentionné :
  - la commercialisation rapide de l'univers;
  - l'élaboration de stratégies entre-croisées/complémentaires pour le Web et les produits imprimés ou sur disque compact;

[1] David Roy, Statistique Canada, Division du marketing, 9-A Immeuble R.H. Coats, Parc Tunney, Ottawa (Ontario), Canada, K1A 0T6.

- un gain, plutôt qu'une perte, de clients attribuable à INTERNET.

Les utilisateurs s'attendent à pouvoir obtenir l'information aussitôt qu'ils la demandent. Des outils de recherche puissants leur donnent accès à l'information présentée dans plus d'un million de pages Web. En outre, la création et le stockage d'informations deviennent de plus en plus faciles.

La valeur de l'information évolue aussi. Les données brutes perdent la faveur des utilisateurs, tandis qu'à l'autre extrémité de l'échelle, on est prêt à payer le prix fort pour des services entièrement personnalisés.

Les clients s'attendent à obtenir auprès d'un fournisseur unique des données provenant de sources multiples. Ils recherchent des fournisseurs qui comprennent leur type d'entreprise, et veulent obtenir des produits et des services en harmonie avec leurs cycles de planification et de production de rapports. Le *Wall Street Journal* publie maintenant un «Journal personnel», dont le contenu est défini par le lecteur. Ce journal contient des résumés de reportages qui peuvent être étoffés de façon à les relier au contexte pertinent - selon le choix du lecteur.

Le service à la clientèle est désormais indispensable pour établir des liens permanents avec les clients et pour accroître les recettes grâce à l'offre de services.

(Reportage de l'agence Reuters - Le prix élevé de la création de nouveaux clients)

Le secteur compte un grand nombre de nouveaux intervenants. Beaucoup sont des petites entreprises qui se spécialisent dans l'intégration de l'information sur une branche d'activité et ajoutent une valeur aux produits sous forme d'analyse des tendances, de prévisions et de «recommandations».

Les partenariats et les alliances constituent des stratégies essentielles pour que les organismes puissent se concentrer sur leur domaine de spécialité. Les partenaires bénéficient de la valeur ajoutée grâce à la mise en commun de leurs points forts - contenu, technologie, diffusion ou force de vente.

Que veulent nos clients? Quel que soit le sujet ou le produit étudié, un même commentaire revient constamment. Ils veulent que Statistique Canada leur offre des données intégrées qui soient accessibles à partir d'un point unique. Ils veulent éviter de communiquer avec 3, 5 ou 8 divisions. Ils s'attendent à une harmonisation des concepts et des normes, y compris dans le domaine de la géographie.

Selon notre étude des modèles et des profils sectoriels établis d'après les données de CANSIM, les clients veulent en général non pas une présentation type, mais des données adaptées à leurs besoins.

Nos clients, en particulier ceux qui intègrent les données et font des prévisions, indiquent aussi de plus en plus fréquemment que Statistique Canada n'est qu'une source parmi tant d'autres. Bon nombre d'organismes ont réduit l'effectif de groupes qui achetaient des données et les analysaient, par exemple, les groupes affectés à la recherche et aux études économiques. Ces services sont désormais achetés selon les besoins - situation qui a fait apparaître sur le marché de nouveaux «intégrateurs» et «producteurs de valeur ajoutée».

Récemment, nous avons interviewé environ 40 distributeurs actuels et potentiels dans le cadre d'une étude de la diffusion. Les répondants ont indiqué que nos pratiques actuelles pourraient limiter l'accès à l'information, qu'il existe de nombreuses possibilités inexploitées et que nos pratiques commerciales ne sont plus en harmonie avec celles du secteur.

UNE VISION DE LA DIFFUSION

Compte tenu des tendances du marché et du secteur, comment allons-nous diffuser nos données dans l'avenir et quelle place allons-nous réserver aux partenariats?

Le graphique que voici présente notre vision d'une stratégie de diffusion. Au centre figure un entrepôt ou base de données centrale contenant

- des données sommaires non publiées (Banque interne de données statistiques),

- des données publiées (Banque externe de données statistiques), qui permettra d'intégrer les données et l'information produites par toutes les divisions du Bureau.

Le deuxième élément important de la vision est StatCan en direct, une interface qui donne accès en direct à l'entrepôt, et contient des métadonnées et des instruments de recherche très poussés. Cette interface est la clé qui nous permettra d'exploiter pleinement nos archives de données, c'est-à-dire d'informer précisément les utilisateurs des renseignements disponibles sur un sujet particulier.

L'entrepôt centralisé donnera naissance à quatre flux de produits et services destinés à servir les utilisateurs finals.

Produits types - Dans l'avenir prévisible, ces produits seront en demande sous forme imprimée, sur CD, sous forme de télécopie et sous d'autres formes.

Accès en direct - L'accès à la base de données centrale pourra avoir lieu pour des raisons aussi bien d'intérêt public que commerciales. En effet, nous compterons toujours des utilisateurs finals «non spécialistes» qui maîtrisent bien nos concepts et la technologie, et dont le nombre ne cessera d'augmenter.

En outre, les «portes d'accès» offertes par d'autres fournisseurs de services en direct élargiront le marché en y faisant entrer les utilisateurs occasionnels. En ce qui concerne l'accès en direct, nos métadonnées seront un outil de recherche essentiel qui permettra de repérer l'information précisée par le client, puis de l'extraire à la demande de ce dernier.

Services personnalisés - À mesure que vont s'améliorer la gamme de données stockées dans l'entrepôt et les métadonnées, nous serons mieux équipés pour offrir des extractions intégrées et personnalisées de données publiées et non publiées. Il s'agira d'une «compétence fondamentale», «hautement valorisée», que ne pourront offrir les autres sources.

Partenariats - Il se créera un réseau élargi entre les fournisseurs de données. Ces partenariats permettront à Statistique Canada de se concentrer sur ses compétences fondamentales et d'en rehausser d'autres en vue :

i) d'intégrer les données à l'information provenant d'autres sources;

ii) d'ajouter une valeur aux produits par des analyses, des prévisions, des présentations ou des affichages;

iii) d'avoir accès à des marchés qui, autrement, ne seraient pas desservis;

iv) de tirer parti de la compétence de tiers en matière de développement et de présentation de produits.

PROCHAINES ÉTAPES

Au nombre des grands projets qui seront entrepris dans les mois et les années à venir en vue de concrétiser cette vision, mentionnons :

- le lancement de StatCan en direct offrant l'accès à CANSIM, à la Base de données sur le commerce et au *Quotidien*;

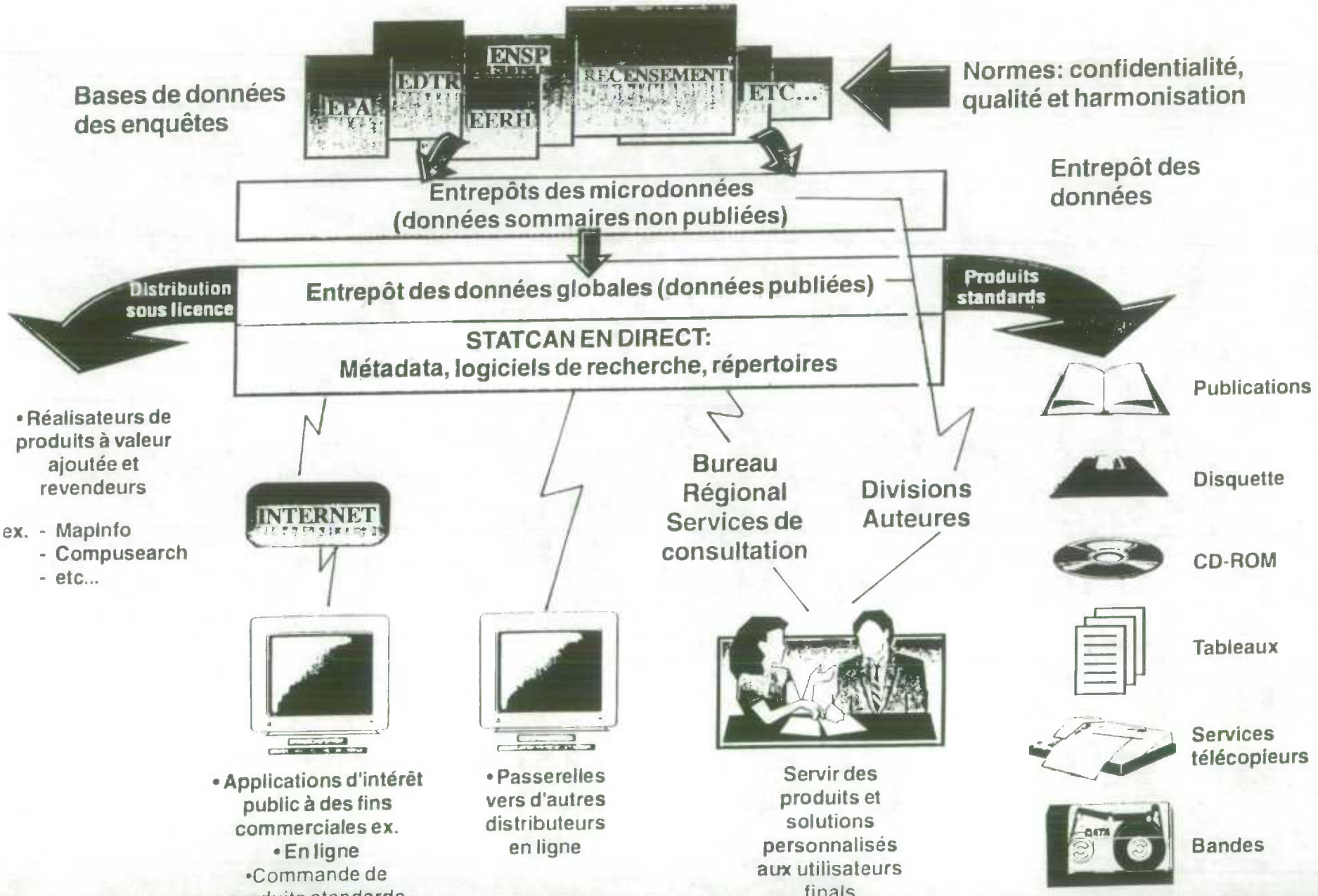
- la planification d'une restructuration pluriannuelle de CANSIM, en vue d'en faire le noyau de l'entrepôt de données;

- la transition du support papier au support électronique pour beaucoup de publications qui visent actuellement une clientèle restreinte;

- la poursuite, dans l'immédiat, du dialogue avec les distributeurs, afin de mieux comprendre les possibilités de partenariat, les conditions éventuellement acceptables et les répercussions pour le Bureau.

Je vous remercie de votre attention. Je serais heureux de répondre aux questions.

VISION STRATÉGIQUE
DIFFUSION À STATISTIQUE CANADA



Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

L'ALLIANCE VUE COMME UN MARIAGE ARRANGÉ

A. Foster[1]

Permettez-moi de prendre un peu de recul. En fait, toutes ces belles et bonnes paroles que nous avons entendues au sujet des partenariats stratégiques m'inquiètent un peu, parce que, justement, je voulais vous proposer de prendre un certain recul et de faire porter notre réflexion sur une analogie, l'alliance vue comme un mariage arrangé. On se sert souvent de la figure du mariage pour décrire le jeu des alliances stratégiques - mais il faut se rappeler que l'union dont on parle ici n'est pas le mariage moderne, sauf en ce qui touche la nécessité d'un moyen simple et pratique d'obtenir le divorce. Autrement, on n'évoque en rien le mariage romantique à la moderne, qui fait partie du décor depuis les années 1850. Il s'agit plutôt du mariage arrangé du XV[e] siècle, principalement, et les règles de «fonctionnement» de ce mariage sont des règles importantes à observer pour ceux qui désirent voir réussir les alliances stratégiques. À cause de nos discussions d'aujourd'hui, je devrai peut-être traiter en plus grand détail des causes d'échec et de réussite de ces alliances, thèmes qui n'ont semble-t-il pas été suffisamment approfondis au cours des discussions antérieures. Cela ne faisait pas partie de mon plan initial, mais j'essaierai d'intégrer ces éléments.

Trois facteurs sont décisifs pour l'efficacité de toute alliance de ce type. Si ces facteurs sont absents, rien n'est possible. Le premier de ces facteurs est la clarté d'intention, c'est-à-dire le but : pourquoi se marie-t-on?

Le deuxième facteur est l'intégrité des partenaires. On ne peut pas mentir au début de la relation et s'attendre à ce que celle-ci se poursuive. On ne peut pas commencer à mentir une fois la relation établie et s'attendre à ce que celle-ci survive. On ne peut pas non plus modifier les règles chemin faisant.

En troisième lieu, les parties doivent avoir une idée claire de leurs intérêts. Or, intérêts et intention ne sont pas une seule et même chose, d'où l'importance de l'analogie avec le mariage arrangé. Dans les deux cas, les intérêts des partenaires sont tout à fait distincts de la raison qui motive leur relation.

Quelle est l'essence même d'une intention? Premièrement, celle-ci doit être claire. Le mariage arrangé du XV[e] siècle était habituellement conclu entre deux familles nobles ou deux maisons royales. Le but de ce mariage, et sa fin publiquement proclamée par l'échange de personnes et de divers biens, de l'épouse au premier chef, était de sceller une promesse ou un pacte de paix. C'était là l'intention première, qui prenait tout son sens par l'échange de personnes.

Il faut aussi que l'intention soit partagée. Une fois que les parties ont déclaré pourquoi elles trouvent avantageux d'entrer en relation, il est donc fondamental d'établir que la relation et l'intention sont communes. Si, par exemple, un mariage arrangé avait lieu simplement parce que l'une des parties convoitait de s'enrichir et que l'autre aspirait à la paix, la relation ainsi nouée risquait fort de se dissoudre. Dans les familles de classe moyenne, on s'efforçait souvent de réunir autant d'argent que possible, afin que le fils ait, par le mariage, accès à un rang social supérieur. Cela entraînait des tensions énormes; en effet, si cette union permettait au jeune mari et à sa famille de s'élever socialement, l'épouse pour sa part se mariait au-dessous de sa condition, c'est-à-dire qu'il y avait «mésalliance». Ces relations comportaient donc une grande part de pressions et de tensions, et celles-ci s'exerçaient également sur les voies ultimes du maintien de la paix entre les pays.

Enfin, la fidélité à l'intention déclarée doit être rappelée à intervalles réguliers. Selon notre analogie, il y avait échange périodique de cadeaux, dans une mesure extrêmement bien calculée. Nous vous donnons notre fille, vous nous donnez de l'argent. Un peu plus tard, advenant par exemple que notre fille devienne veuve, vous lui remettrez l'argent nécessaire à sa subsistance; d'ici là, à l'occasion, je vous accorderai un titre de noblesse pour votre beau-frère ou votre cousin, et ainsi de suite. Il s'agissait d'un échange très calculé.

Cela dit, pourquoi faut-il comprendre l'intention et

---

[1] Anne Foster, Carswell & Thomson Professional Publishing, Professional Bldg. 1 Corporate Plaza, 2075 rue Kennedy, Scarborough (Ontario) M1T 3V4.

son rôle dans les alliances stratégiques? Premièrement, si l'on ne comprend pas l'intention de l'autre partie, on peut se méprendre sur ses besoins. Si vous ne comprenez pas pourquoi Statistique Canada conclut un partenariat avec Compusearch, ou si Compusearch ne le comprend pas, personne ne comprendra les décisions prises par la suite et les causes de malentendu seront multiples. Par exemple, j'ai vu ici la mise en commun de certains niveaux d'expertise. Ce que le personnel de Compusearch partage avec Statistique Canada, ce sont les moyens de rejoindre le marché et la capacité de créer un produit à valeur ajoutée; de son côté, Statistique Canada offre en partage des données brutes fiables et d'excellente qualité. Mais, quelle est l'intention qu'ont en commun les deux partenaires?

L'élément clé qui vient ensuite est l'intégrité - d'où la nécessité de la transparence. Au XIV[e] siècle, une certaine Lady Duborc avait rédigé un document, sous forme versifiée, à l'intention de son neveu. Voulant lui expliquer le sens du mariage, elle en soulignait une dimension cardinale. En substance, voici ce qu'elle écrivait. L'épouse de son neveu allait être bien seule dans son nouveau pays car, dans le cadre de cette alliance, elle arrivait d'une autre partie du monde, où elle avait laissé famille et amis. Il reviendrait donc à son mari de jouer un rôle de soutien à son égard et, notamment, de lui faire savoir clairement en quoi consistait sa dot. C'est là ce que veut dire le terme «transparence». Ainsi, lorsque s'établit une relation, particulièrement entre des partenaires qui nouent des rapports inusités, les partenaires doivent être francs quant aux conditions de la relation, afin de protéger leurs intérêts réciproques et aussi de réitérer en public les motifs de l'accord.

La tenue de comptes exacts représente un autre volet fondamental des relations, ce que n'avait pas non plus oublié Lady Duborc. Ayant déclaré franchement en quoi consistait la dot, le mari et administrateur devait ensuite faire périodiquement rapport à sa femme, jusqu'à la fin de sa gérance. Cette fonction est une partie importante de toute relation stratégique. Il faut être transparent quant à l'utilisation des fonds et des ressources. Je vous ai entendu vanter les mérites de Compusearch à cause des recettes; assurons-nous de savoir qui en profite, et de quelle façon.

Enfin, les partenaires doivent s'assurer un appui mutuel. Les intérêts de chacun doivent être exprimés. Mais - quelle est la différence entre intérêts et intentions? Il me semble que les intérêts, de part et d'autre, sont de toute évidence très différents, mais ils doivent coïncider sur un nombre suffisant de points si l'on veut que les intentions se réalisent.

Par exemple, l'intérêt du gouvernement du Canada, et non uniquement de Statistique Canada, consiste à assurer le bien public en prenant les moyens appropriés. L'intérêt de l'entreprise partenaire réside dans la création d'un produit et sa vente lucrative. Ces intérêts ne sont certes pas mutuellement exclusifs, mais ils sont différents; pour s'assurer du succès d'une alliance, il faut savoir cerner la fin commune qui est conciliable avec ces intérêts.

Pour illustrer des alliances stratégiques qui ont échoué et les raisons de leur échec, quelques exemples me viennent à l'esprit. Remontons au XV[e] ou, plutôt, au XVI[e] siècle, puisque je pense à Henri VIII, qui a eu six femmes. De ce nombre d'unions, je crois que deux seulement avaient été «arrangées», comme moyen d'atteindre une fin plus générale. Le premier mariage du roi, arrangé par ses parents, répondait sans doute assez bien à ses exigences puisque, si je ne me trompe, il a été marié pendant environ 18 ans avec Catherine d'Aragon, et il n'a jamais pu se décider à la faire décapiter. Les choses se sont bien passées pendant une période assez longue, mais les intérêts d'Henri ne coïncidaient pas avec le but de l'union. Ce but était de faire en sorte que le pape et l'Espagne s'allient à l'Angleterre, pour empêcher la France de songer à une invasion ou à une annexion de territoires; voilà ce qu'était en réalité la raison d'être de l'opération. Cependant, les intérêts d'Henri sont devenus un peu plus prosaïques, et sa deuxième femme, Anne Boleyn, ne lui apportait rien de dynastique; elle n'allait lui apporter rien d'autre non plus, à moins du mariage. On voit, en conséquence, que ce premier mariage dynastique a échoué, et ce, parce qu'il y avait conflit d'intérêts.

Le deuxième mariage dynastique du roi, avec Anne de Clèves, a eu lieu beaucoup plus tard. Cette union avait été arrangée par le chancelier, qui allait d'ailleurs le regretter amèrement, et elle devait favoriser une alliance des partisans de la cause protestante à l'époque. À ce stade, en fait, Henri s'intéressait à tout autre chose qu'aux alliances. Et Anne de Clèves a été son épouse pendant sept jours exactement; il est tout à son honneur que, ayant compris la nature des intérêts d'Henri aussi bien que l'intention des factions de la cour, elle ait survécu. Une fois hors de cette mauvaise passe, elle a pu mener une vie honorable de demi-divorcée au château de Hampton et se faire une réputation de cordon-bleu. Là encore, notons-le, l'intérêt avait tourné court.

Quels sont donc les mariages qui ont réussi?

J'allais citer Catherine de Médicis comme exemple d'une alliance stratégique couronnée de succès. Mais je ne peux me servir de cet exemple, parce que Catherine était en quelque sorte défavorisée comme partenaire. Elle n'était que la fille d'un banquier toscan, et les

membres de la famille royale le lui faisaient cruellement sentir. Cependant, ses intérêts coïncidaient, sur des points clés, avec ceux des diverses factions de la cour, situation qu'elle a su entretenir.

Revenons maintenant aux partenaires possibles du mariage dont nous traitons ici - qui sont le gouvernement et le secteur privé - et voyons un peu si nous pouvons découvrir leur fin ou intention commune. J'avancerais que le but commun de nos deux parties est d'instaurer un climat économique sain. On peut difficilement faire payer des impôts aux pauvres. On peut toujours essayer, mais on ne le devrait pas. Il faut arriver à dynamiser le climat économique. Voilà le but de toute relation entre un gouvernement et un partenaire du secteur privé.

En deuxième lieu, l'intégrité a deux exigences. Du côté gouvernemental, il ne doit pas y avoir d'abus de pouvoir. Et l'on serait étonné de savoir ce qui, aux yeux des gens, constitue un abus de pouvoir. C'est pourquoi j'insistais tout à l'heure sur la clarté. Voici, aux yeux du secteur privé, un exemple d'abus de pouvoir. Disons que, ayant mis au point un progiciel ou un plan de marketing, nous le partageons, dans le cadre d'une soumission, avec un organisme gouvernemental à titre de partenaire éventuel, mais que nous n'obtenons pas le contrat - ce qui est normal et ne constitue pas un abus de pouvoir. Or, n'ayant pas obtenu le contrat, nous constatons pourtant, un an plus tard, que le ministère en question ou même qu'un de nos concurrents exploite ce même progiciel ou plan (eh oui, cela s'est déjà vu) : alors, bien sûr, il y a réellement abus de pouvoir.

Le recours à la *Loi sur l'accès à l'information* ou à la protection du droit d'auteur comme moyen d'interdire l'accès à l'information constitue un abus de pouvoir et met donc l'intégrité du rapport en danger. Par ailleurs, le secteur privé doit connaître et respecter les compétences et la valeur des activités du secteur public. En ce qui me concerne, j'ai déjà travaillé ici même, au Conseil canadien de la documentation juridique. Je suis certaine que nous avons tous été témoins de situations où l'on voit le représentant du secteur privé faire la moue devant les efforts et les investissements consacrés à la création de produits, ou certaines personnes carrément sous-estimer les efforts nécessaires pour accomplir ce que nous exigeons du secteur privé. J'allais dire que si, dans le secteur privé, il y avait un plus grand nombre de femmes dans des postes de direction, ce genre de situation serait moins fréquent, mais Jan et moi-même faussons le présent échantillon.

Les intérêts, facteur que j'ai mentionné plus tôt, sont également différents. Les intérêts du gouvernement se rattachent à la diffusion en vue du bien public, et les intérêts du secteur privé se rattachent à la publication à but lucratif. Cela en soi ne présente pas de problème; il faut simplement le comprendre.

Leon Battista Alberti était marchand dans la Florence du XV[e] siècle. De nombreuses familles florentines tenaient alors des registres généalogiques; dans ces familles, les mariages occupaient de 50 à 60 % du temps de planification, ce qui veut dire que c'était un important sujet de réflexion. Notre personnage disait qu'«une foule de mariages ont été cause de l'anéantissement d'une famille, parce qu'ils avaient été conclus avec des partenaires exagérément fiers ou malveillants, portés au litige et à la chicane». Je prétends qu'il en va de même chez les partenaires stratégiques et que, du côté entreprise, si l'entente échoue, on perd toute maîtrise. On perd la maîtrise de son produit ou l'on perd la maîtrise de l'information diffusée par le gouvernement, on perd la maîtrise politique et la vie devient difficile.

En revanche, le mariage réussi débouche sur de nouvelles possibilités, de nouveaux contacts, de nouvelles sources d'information et de nouveaux alliés. Et, pour affronter le changement radical dont nous venons de discuter, nous avons tous besoin de nouveaux alliés. En effet, dans le cas contraire, il nous sera impossible de réaliser ce que j'ai défini comme notre fin commune, soit un climat économique vigoureux. Merci de votre attention.

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

INTRODUCTION DE PARTENARIATS DANS LE SECTEUR DE L'INFORMATIQUE OU LE CROQUET AU PAYS D'ALICE

P. Brandon[1]

Bonjour!

On m'a invité pour vous parler des partenariats dans le secteur de l'informatique. En tant que rédacteur en chef et éditeur d'une feuille de chou intitulée «Electronic Information Partnerships», on suppose que j'ai certaines connaissances dans le domaine des partenariats. Ne voulant décevoir personne, je n'ai pas contesté cette présomption. Me voici donc parmi vous aujourd'hui...

Avant que vous vous rendiez compte de l'étendue de ma méconnaissance du sujet, je veux m'assurer que nous avons une perception commune, ainsi qu'une expérience et des références similaires, au sujet du secteur de l'informatique et de sa situation actuelle. Laissez-moi donc vous donner un aperçu de la perception que j'ai du secteur de l'informatique et de ses activités en général. À cette fin, ma pensée s'est irrésistiblement tournée vers l'oeuvre de Lewis Caroll.

Vous vous rappelez sans doute du jeu de croquet dans *Alice au pays des merveilles*, un jeu fictif dont aucun élément ne reste stable très longtemps. Tout est animé et change autour d'Alice. Le maillet qu'elle utilise est un flamant qui lève sa tête juste au moment où Alice tente de frapper la boule. La boule est en fait un hérisson, qui fait comme bon lui semble. Plutôt que d'attendre d'être frappé par Alice, il s'étire, se redresse, se déplace dans le jeu et s'immobilise enfin. Quant aux arceaux, il s'agit de soldats de cartes à jouer qui, sous les ordres de la Reine de coeur, se déplacent sans arrêt sur le terrain.

Tentons maintenant d'appliquer ce modèle à notre contexte.

Vous pouvez d'ailleurs faire jouer les divers rôles par les personnes de votre choix. Voici, par exemple, ce à quoi un employé de Statistique Canada pourrait arriver. Remplaçons *Alice* par *un employé moyen de Statistique Canada qui travaille en informatique*, dans le rôle du joueur de croquet, le *maillet* par la *technologie de l'information (TI)*, le *hérisson* par *l'information*, les *arceaux* par les *attentes des employés*, et la *Reine de coeur* par le *statisticien en chef*. Je crois que cela correspond à peu près à l'image du jeu que se font la plupart des personnes présentes aujourd'hui!

Il va sans dire que vous pouvez faire les substitutions que vous voulez. En fait, j'aimerais que vous me fassiez part des plus farfelues, et je promets que je les publierai et que j'en accorderai le crédit à leurs auteurs.

Sur une note plus sérieuse, je crois que le *jeu de croquet au pays des merveilles*, ou peut-être *le jeu de croquet au pays de la cybernétique*, décrit bien l'environnement, les défis, le contexte et les règles tordues qui caractérisent le travail des spécialistes de l'informatique aujourd'hui. Je suis certain que chacun des autres panelistes sera d'accord avec moi pour assimiler son expérience quotidienne et son travail aux déboires, victoires et défis d'Alice jouant au *jeu à la mode* : le croquet au pays de la cybernétique.

Après cette digression métaphorique, laissez-moi situer à nouveau la question dans son contexte. Que signifient les partenariats en cette ère de la technologie de l'information? Quelles sont les règles du jeu qui s'appliquent aujourd'hui.

Sans vouloir parodier David Letterman, je vais vous soumettre dix propositions brèves qui pourraient servir de règles aux partenariats.

Je commencerai par la dixième et ainsi de suite jusqu'à la première.

---

[1] Peter Brandon, associé, Sysnovators Ltd., et éditeur et diffuseur, Electronic Information Partnerships, 17 Taunton Place, Gloucester (Ontario), K1J 7J7, email: pbrandon@fox.nstn.ns.ca.

PROPOSITION N° 10 : Nous devrons nous faire à l'idée qu'il n'y a pas de modèle unique. Nous devrons apprendre à structurer des partenariats spécialisés et personnalisés : partenariats de direction, partenariats d'exécution, partenariats de consultation.

Chacun d'eux diffère et comporte son propre «code génétique», ses priorités, son objectif et ses règles.

PROPOSITION N° 9 : Nous devons réinventer des partenariats de diffusion dans un contexte où les coûts de cette dernière sont sur le point de devenir nuls.

Il s'agit là d'une nouvelle réalité du monde de l'information aujourd'hui. Au fur et à mesure que l'information devient plus fluide et que les atomes (de papier ou de silicone) se transformant en éléments binaires, les coûts de la diffusion et les bénéfices à en tirer s'amenuisent. Je crois que nous devons rétablir la chaîne de valeurs entre le fournisseur d'information et l'utilisateur, et repenser entièrement la notion de «diffusion» ou de «chaîne de diffusion».

PROPOSITION N° 8 : Nous devrons apprendre à établir des partenariats dans un contexte de rentrées accrues.

Nous savons comment établir des partenariats dans un contexte de gagnant-perdant et de diminution des rentrées. Nous devons réapprendre à le faire dans un contexte de rentrées accrues, ce qu'en fait l'économie de l'information nous appelle à faire. Dans ce contexte, les idées reçues en matière économique (et plus particulièrement que la rareté crée la demande et fait diminuer les rentrées), sont modifiées totalement. Comme l'a si bien dit un de nos grands penseurs :

> «Dans le cas des biens de consommation, il existe une corrélation directe entre la rareté et la valeur. L'or vaut plus que le blé, même s'il est impossible d'en manger. Il existe des exceptions, mais la situation est souvent inverse en ce qui a trait à l'information. La plupart des produits doux prennent de la valeur au fur et à mesure que leur utilisation se répand. La familiarité est un actif important dans le monde de l'informatique. Et on s'aperçoit souvent que la meilleure façon d'accroître la demande pour un produit est de le distribuer gratuitement.» - **John Perry Barlow, parolier, fermier à la retraite et co-fondateur d'Electronic Frontier Foundation, tiré de *The Economy of Ideas*, magazine Wired, mars 1994.**

Dans un monde d'objets, la familiarité suscite l'indifférence du fait de la diminution des rentrées. Dans un monde d'information, la familiarité suscite la demande du fait de l'augmentation des rentrées. Les entreprises de logiciels voient la demande augmenter lorsqu'elles distribuent gratuitement leurs produits. Allez donc comprendre!

PROPOSITION N° 7 : Les partenariats doivent être axés sur les forces de base des partenaires.

Voici un exemple intéressant de partenariats qui permettent à chacun des partenaires de mettre à profit leurs propres compétences. Le Conseil de développement commercial de Hong Kong est devenu, de l'avis même de son président, un chef de file pour un nombre important d'efforts de promotion de Hong Kong à l'étranger.» Le mandat du conseil consiste à favoriser le commerce au niveau mondial, afin de diversifier les marchés d'exportation; à améliorer la conception des produits et leur image de marque; à faire de Hong Kong le centre de commerce et d'exposition de l'Asie; et à maintenir un contexte permettant une plus large diffusion des produits de Hong Kong partout dans le monde. Les efforts du conseil sont bien financés; son budget en 1992 était de 784 millions de dollars HK (100 millions de dollars US). *Au total, 58 % de ses revenus (419 millions de dollars HK) provenaient des droits de 0,05 % imposés par le gouvernement sur les importations et les exportations de Hong Kong*, le reste des revenus découlant des efforts déployés par le conseil.

Il s'agit là d'un bon exemple d'agencement des compétences de base du gouvernement dans le domaine de la perception de recettes fiscales, de la capacité particulière d'une organisation de fournir un service d'information stratégique. Ainsi, chacune des parties se charge de l'aspect pour lequel elle excelle, ce qui profite à l'ensemble du partenariat.

PROPOSITION N° 6 : Les partenariats devront être fondés de plus en plus sur la compatibilité culturelle, la confiance, ainsi que les codes et les principes d'éthique, plutôt qu'exclusivement sur des contrats et des ententes juridiques.

Les notions et concepts juridiques traditionnels ne pourront pas être maintenus à l'ère de l'informatique. La propriété intellectuelle prendra de nouvelles formes. Des codes d'éthique remplaceront les lois et règlements, un peu à la façon des règles du FarWest, qui remplaçaient des lois impossibles à faire appliquer. Toujours selon John Perry Barlow, dans son article *The Economy of Ideas*, publié en 1994 dans le magazine *Wired*:

«Avant la colonisation de l'Ouest, l'ordre était assuré selon un code informel, qui avait la souplesse de la Common Law plutôt que la rigidité des règles. L'éthique prenait le pas sur les règles établies. Les ententes avaient préséance sur les lois qui, de toute façon, étaient bien difficiles à faire respecter.»

PROPOSITION N^{O} 5 : La technologie de l'information favorise la coordination entre les marchés, plutôt que la coordination à l'intérieur des entreprises. Ainsi, les entreprises «achèteront» davantage qu'elles ne «fabriqueront». Cela mènera à un recours plus grand aux partenariats.

Étant donné que la technologie rend plus facile, rapide et économique la coordination de l'information, les entreprises découvrent qu'il peut être plus efficient et pratique d'acheter des biens ou des services plutôt que de les produire. Deux économistes du MIT, Thomas Malone et John Rockart, prétendent que l'intégration verticale au sein des entreprises devient de moins en moins efficiente du fait que l'on peut compter sur des entreprises plus petites davantage à l'écoute du marché.

Cela a pour effet d'embrouiller les limites qui existent actuellement entre la coordination interne et externe des entreprises. Les rapports entre les entreprises pourraient être axés davantage sur la collaboration que sur la concurrence du fait que les technologies de l'information rendent attrayantes et faisables la conclusion d'alliances stratégiques à long terme.

Peter Drucker compare la structure à venir des entreprises à celle d'un orchestre ou d'un hôpital : des gestionnaires supérieurs supervisent un ensemble de spécialistes largement autonomes qui «dirigent et organisent leur rendement à partir des réactions de leurs collègues, des clients et du bureau principal.» Il désigne cette structure sous le nom «d'organisation axée sur l'information». Grâce aux partenariats, nous pourrons élargir ces organisations axées sur l'information. Mais nous devrons d'abord apprendre comment y parvenir.

PROPOSITION N^{O} 4 : Les partenariats nécessiteront un partage plus important et plus rapide de l'information, des valeurs et des attentes.

Il est une notion qui devient de plus en plus populaire de nos jours, celle du «continuum d'information». Les organisations de plus en plus décentralisées, qui comptent sur un nombre croissant d'entrepreneurs, de fournisseurs et de partenaires, découvrent que la seule chose qui les réunit encore est l'information. C'est cette dernière, continue, partagée et largement diffusée, qui garantit la coordination des efforts, la cohérence des buts et la synchronisation de l'exécution.

Si l'on fait une analogie avec le hockey, le continuum d'information, c'est-à-dire l'information partagée par les partenaires, peut-être perçu comme la glace : lisse, sans fissure ni faille, qui permet aux joueurs d'évoluer sans crainte, à la rondelle de glisser sans anicroches, aux joueurs de se concentrer sur le jeu. Si on le compare au croquet, le continuum d'information pourrait être un carré d'herbe rase bien entretenu, qui facilite le jeu.

Si nous voulons que nos partenariats de croquet cybernétique fonctionnent, nous devons nous préoccuper sans arrêt de la qualité du terrain, c'est-à-dire du continuum d'information, une garantie de prévisibilité et d'uniformité pour tous les intervenants.

PROPOSITION N^{O} 3 : Le terrain de jeu, c'est-à-dire le continuum d'information, présente des besoins particuliers. Son entretien comporte des coûts élevés de transaction, des lacunes du point de vue de l'apport en fertilisant (métadonnées) et l'absence de similitudes quant aux contextes qui nuit à de nombreuses autres associations par ailleurs valables.

Parmi les aspects dont il faut être conscient et au sujet desquels il faut être particulièrement vigilant au fur et à mesure de l'établissement du continuum d'information, on note les suivants.

- Le fait que l'information comporte des coûts inhérents très élevés de transaction. Les coûts nécessaires pour trouver les bonnes données au moment opportun, et pour les diffuser au bon endroit et sous la forme appropriée, sont encore déraisonnablement élevés.
- Il existe des lacunes chroniques du point de vue des métadonnées; c'est-à-dire l'information sur l'information. Cela a évidemment pour effet d'augmenter les coûts de la recherche. Le problème vient du fait qu'il y a peu de prestige et d'argent à tirer des métadonnées. Le corollaire direct du manque chronique de métadonnées est le suivant : le peu de normes d'étiquetage dans le monde de l'information et, en fait, le peu d'étiquettes fiables pour les objets d'information qui traversent notre espace électronique.
- L'information ne prend son sens que dans un contexte particulier. Le fait de la sortir de son contexte lui fait perdre toute signification. Le fait de la placer dans un contexte différent lui donne une signification totalement différente. Nous présumons souvent, sans motif valable, qu'il existe un contexte

commun. Par ailleurs, nous ne sommes pas particulièrement portés à faire la promotion d'un contexte et à le transmettre avec nos messages. Par conséquent, nos messages sont souvent hors contexte, ou n'ont pas de contexte du tout, et ont des résultats décevants.

Que devons-nous faire? Tout d'abord, nous devrons demander aux intervenants qui sont là pour nous aider d'établir des espaces d'information communs facilement accessibles, ces terrains de jeu bien entretenus qui sont le gage du succès. Les courtiers électroniques, par exemple les intermédiaires du commerce électronique ou les fournisseurs de services en réseau à valeur ajoutée, font partie des «préposés à l'entretien» qui peuvent nous venir en aide. Toutefois, ces intervenants auront besoin d'une certaine légitimité ainsi que d'un cadre juridique et réglementaire pour fonctionner, selon le niveau de sécurité et de fiabilité du continuum d'information qui doit être établi et maintenu. Le défi des «préposés à l'entretien de l'information» est de taille.

PROPOSITION N° 2 : Nous devrons apprendre la signification de la *subsidiarité* dans la structure des partenariats.

La notion de subsidiarité est à la base de la perception qu'a Jacques Delors de la Communauté européenne (CE). Dans ce contexte, cela signifie uniquement que le pouvoir est entre les mains de chaque pays de la Communauté. Ce n'est qu'avec l'accord des pays que Bruxelles peut exercer son autorité. Dans ce contexte, la subsidiarité est le contraire de l'habilitation. Les pouvoirs ne sont pas transférés ou délégués à partir du centre. On part plutôt du principe que le pouvoir est entre les mains des intervenants au niveau le plus bas de la hiérarchie. L'Église catholique, par exemple, applique les mêmes principes en disant que chaque prêtre joue le rôle de Pape dans sa propre paroisse. (Au point où j'en suis, pourquoi ne pas faire intervenir l'Église catholique pour vous convaincre encore davantage!)

Que signifie la notion de subsidiarité dans le cadre d'un partenariat? Elle signifie que les partenaires ne sont pas simplement des rouages, qu'ils ont réellement des pouvoirs, qu'ils peuvent s'exprimer et qu'ils ont la responsabilité et la capacité de faire des choses par eux-mêmes et non pas uniquement dans le cadre de leur partenariat avec d'autres. Ils ont ces pouvoirs de façon inhérente, du fait du rôle qu'ils jouent.

Si l'on transpose cette notion sur le terrain de croquet, la subsidiarité signifie que la légitimité de l'équipe repose sur les talents de chacun, les efforts concertés des joueurs et de leur volonté de faire équipe. L'équipe correspond au partenariat fonctionnel. (Veuillez me pardonner si j'oublie parfois que le croquet n'est pas réellement un sport d'équipe. Je dois vous avouer que ces petits oublis me fournissent parfois l'occasion de pousser mes comparaisons au-delà des limites acceptables.)

ENFIN, LA PROPOSITION N° 1: Nous allons assister à l'avènement de principes politiques sur le terrain de croquet : celui-ci deviendra un espace fédéré, un espace où les rapports nous permettront de constater réellement la rareté des ressources.

On nous lave le cerveau en nous disant que la rareté des ressources a à voir avec le spectre, la largeur de bande ou la vitesse des ordinateurs. La progression de la technologie au cours des 30 dernières années démontre tout à fait le contraire. Ces ressources sont devenues pour nous sans limite.

La rareté des ressources continue toutefois de se faire sentir dans les rapports humains :

1) l'attention - Nous n'avons pas encore trouvé de façon de donner aux personnes une plus grande capacité d'attention, afin qu'elles puissent écouter de façon plus intense pendant de plus longues périodes de temps, et retenir davantage de matière.

2) «largeur de bande» des personnes -- Saviez-vous que nous communiquons (c.-à-d. que nous échangeons de l'information) les uns avec les autres au rythme ridicule de 55 bits par seconde? (Par comparaison, la transmission électronique d'information se fait à 10 millions de bits par seconde, et on atteindra bientôt 155 millions de bits par seconde!). Le fait de bien connaître la personne à qui vous envoyez un message par télécopieur ou par courrier électronique peut grandement accroître la qualité de la communication. Plus vous connaissez une personne et moins la «largeur de bande» doit être importante.

3) la capacité des personnes de tirer parti des enseignements des données; comme le disait T.S. Eliot dans son poème, *The Rock* :

Où est la vie que nous avons perdue en la vivant?
Où est la sagesse que nous avons perdue dans la connaissance?
Où est la connaissance que nous avons perdue dans l'information?

Et, on pourrait ajouter, où est l'information que nous avons perdue dans les données?

4) l'éthique des personnes en matière d'information (ne pas déformer le message; ne pas dissimuler de l'information; communiquer avec ses pairs de façon facile à comprendre et significative; faire part de ses connaissances et de sa sagesse).

Je suis d'avis que c'est là que réside la rareté aujourd'hui, et non dans pas la taille canaux électroniques ou dans la rapidité du matériel. Voici une citation intéressante d'une dénommée Linda Ray Pratt (je n'ai aucune idée de qui il s'agit), dont j'ai pris connaissance l'autre jour :

> «L'aptitude à philosopher ne vient pas plus facilement avec la fibre optique. La clarté de la pensée ne va pas avec la précision de la résolution à l'écran. La connaissance de soi-même et la perception des autres n'a rien à voir avec les mises à niveau de logiciels.»

En fait, si vous vous y arrêtez, vous vous rendrez compte que la plupart de nos problèmes, y compris la protection des renseignements personnels, l'accès à l'information, les erreurs de communication, la sécurité dans les communications ou les échecs dans les partenariats, surviennent par suite de lacunes graves dans les ressources vraiment rares.

Laissez-moi résumer

Je crois que notre capacité à établir des réseaux efficaces et des partenariats réussis dépendra de notre capacité et de notre volonté à faire de notre espace de jeu de croquet un environnement fédéré. Dans cet espace, nous devrons :

- nous faire à l'idée qu'il n'y a pas de modèle unique;
- réinventer des partenariats de diffusion dans un contexte où les coûts de diffusion diminuent;
- apprendre à établir des partenariats dans un contexte de rentrées accrues;
- tirer parti des compétences de base de chaque partenaire;
- fonder les partenariats sur la compatibilité culturelle, la confiance et les codes d'éthique, plutôt qu'exclusivement sur des contrats et des ententes juridiques (ayez confiance, mais vérifiez tout de même!);
- faire de nos partenariats un mécanisme de coordination plus efficace axé sur le marché;
- partager davantage l'information, les valeurs et les attentes, avec nos partenaires;
- assurer l'entretien de notre terrain de jeu (le continuum d'information) en tout temps;
- apprendre la signification de la subsidiarité dans nos ententes de partenariat.

Avant de terminer, j'aimerais vous faire part de deux pensées, qui ne sont pas de moi, pour faire changement.

- Lorsque vous planifierez votre prochain projet, il serait sans doute approprié de vous rappeler la pensée de William Gibson, l'inventeur de *Cyberspace*: «L'avenir est déjà là, mais il n'est pas réparti également».
- Enfin, en ce qui a trait aux questions abordées, je crois que ces paroles, attribuées à F. Scott Fitzgerald, vous seront très utiles : «l'intelligence se manifeste par la capacité d'avoir deux idées opposées en même temps et de pouvoir continuer à fonctionner. Quelqu'un peut par exemple voir que la situation est sans espoir, mais être quand même déterminé à trouver une solution.»

Merci et bonne partie de croquet.

ALLOCUTION DE CLÔTURE

Recueil du Symposium 95 de Statistique Canada
Des données à l'information - méthodes et systèmes
Novembre 1995

ALLOCUTION DE CLÔTURE

G.J. Brackstone[1]

Depuis deux jours et demi, nous faisons le survol d'un sujet aussi vaste que multiforme. Nous avons abordé certaines questions de contenu des données statistiques, de technologie, de méthodologie, ainsi que la dimension pratique de la diffusion de l'information. Nous avons ainsi parcouru un terrain appréciable, à commencer par la vision évoquée par Peter Hick du type de base de données qu'il croit nécessaire à la formulation des politiques publiques, en matière sociale tout particulièrement, jusqu'aux discussions de ce matin, davantage centrées sur les technologies de gestion et les partenariats pouvant faciliter aux usagers l'accès à l'information.

À mon humble avis, il est impossible de résumer tout ce dont nous avons traité; de toute manière, je ne pourrais guère y prétendre, n'ayant pu être présent à toutes les séances. Je voudrais cependant relever deux des thèmes qui m'ont paru dominer nos propos. En premier lieu et de toute évidence, il y a la technologie, tant par l'impact de la puissance de calcul sur nos méthodes de traitement, de gestion et d'analyse des données que par l'impact des médias électroniques sur nos modes de diffusion et de distribution des données. En second lieu, il y a ces pressions d'ordre financier, qui ont ponctué une grande partie de nos discussions. C'est le problème d'accomplir et de produire toujours davantage avec des ressources restreintes et, comme on l'a montré ce matin, sur le plan des recettes, les ramifications de la nécessité, pour les organismes de statistique, du recouvrement de leurs coûts par la diffusion de leurs données.

Avant de clore ce volet du symposium, je voudrais exprimer ma gratitude, avant tout aux membres du Comité organisateur. Le Comité travaille depuis des mois à la préparation du colloque. Je tiens à remercier chaudement le président du Comité, Jean-Louis Tambay, et ses trois autres membres, Georgia Roberts, John Berigan et Jean Dumais. De nombreux collaborateurs leur ont prêté main-forte, notamment quatre personnes-ressources : Josée Morel, Sophie Arsenault, Christine Larabie et Nick Budko. Une foule de bénévoles, qui restent dans l'ombre, nous ont aidés de mille et une façons, et je voudrais les saluer ici. Une pensée toute spéciale va à nos interprètes, qui nous ont assidûment accompagnés pendant ces deux journées et demie. Je crois que nous avons tous apprécié leurs services. Je voudrais également souligner l'apport déterminant de nos conférenciers et panélistes au déroulement du colloque. Nous savons gré de leur collaboration aux démonstrateurs, que nous verrons à l'oeuvre cet après-midi.

Nous vous invitons à nous faire parvenir vos commentaires, positifs ou négatifs, sur la tenue du colloque et sur les éléments qui, d'après vous, pourraient être améliorés. Au fil des pauses, on m'a fait des remarques tout à fait intéressantes et utiles. Plus ces suggestions seront nombreuses, plus nous pourrons améliorer notre formule. Nous avons traité d'un vaste éventail de sujets. Tous n'ont certes pas présenté le même intérêt pour chacun d'entre nous, mais j'espère vivement que tous et chacun ont trouvé ici des propos stimulants et des pistes neuves.

Soyez donc tous remerciés de votre participation, au nom de Statistique Canada.

---

[1] G.J. Brackstone, Statisticien en chef adjoint, Secteur de l'informatique et de la méthodologie, 26-J, Édifice R.H. Coats, Parc Tunney, Statistique Canada, Ottawa, (Ontario), Canada K1A 0T6.

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010224049
Ca 005