11-617F
no.88-02
ex. 2

CONTRIBUTION DE STATISTIQUE CANADA À L'ÉCHELLE INTERNATIONALE DANS LES DOMAINES DE L'INFORMATIQUE ET DE LA MÉTHODOLOGIE

Document préparé par
I.G. Sande
en vue d'un exposé
Statistique Canada
Janvier 1988

Z-194F
c.2

STATISTICS CANADA STATISTIQUE CANADA
AUG 9 1988
LIBRARY
BIBLIOTHÈQUE

CAHIER DE TRAVAIL NO. BSMD-88-002F
DIRECTION DE LA MÉTHODOLOGIE
Division des méthodes d'enquêtes-entreprises

CONTRIBUTION DE STATISTIQUE CANADA À L'ÉCHELLE INTERNATIONALE DANS LES DOMAINES DE L'INFORMATIQUE ET DE LA MÉTHODOLOGIE

Document préparé par
I.G. Sande
en vue d'un exposé
Statistique Canada
Janvier 1988

1. INTRODUCTION

Les organismes statistiques des autres pays éprouvent des problèmes similaires à ceux auxquels se heurte Statistique Canada en matière de collecte, de traitement et de diffusion des données. Une grande partie de nos travaux traitant de ces problèmes ont attiré l'attention d'autres pays et ont influé sur leurs propres efforts à cet égard.

Il sera question ici d'une partie des travaux réalisés à Statistique Canada dans les domaines de la théorie et des méthodes statistiques ainsi que des applications informatiques qui ont été reconnues à l'échelle internationale. Il ne s'agit pas d'un exposé exhaustif des réalisations de Statistique Canada. En raison du manque de temps, il ne sera pas fait mention de certains thèmes, qui sont néanmoins intéressants et importants. Il ne faut pas oublier que Statistique Canada continue de se consacrer à des travaux inédits et qu'il y a un échange continuel d'idées entre Statistique Canada et les autres organismes statistiques. Il arrive souvent que des idées qui semblent peu importantes de prime abord influent sur la réflexion faite en d'autres lieux.

Les domaines auxquels Statistique Canada a apporté sa contribution et qui seront abordés ici sont les suivants.

a) Confidentialité : la non-divulgation des renseignements relatifs aux répondants dans les tableaux produits par le Bureau.

b) Contrôle et imputation : la détection et la correction, à l'aide de moyens automatisés, des erreurs entachant les données.

c) Analyse des données provenant d'enquêtes dont le plan d'échantillonnage est complexe : la possibilité de tirer des conclusions valables à partir des données lorsque les hypothèses habituelles ne s'appliquent pas.

d) Désaisonnalisation : la possibilité de faire disparaître les tendances saisonnières présentes dans les données recueillies régulièrement au cours d'une longue période afin de faire ressortir les tendances à long terme.

e) Gestion des bases de données statistiques : l'organisation des données à des fins statistiques.

f) Géocartographie : l'automatisation des aspects liés à la géographie en ce qui concerne la collecte des données au Canada, et la présentation des données de façon à illustrer les variations ayant trait aux caractéristiques pour l'ensemble du pays.

Le présent exposé ne vise pas à rendre hommage aux personnes à qui nous sommes redevables des innovations dont il est question ici. Il s'agit plutôt de célébrer le travail de l'ensemble du Bureau, étant donné qu'un grand nombre de personnes ont apporté leur concours d'une façon ou d'une autre. C'est pourquoi personne n'est nommé expressément.

2. PROTECTION DE LA CONFIDENTIALITÉ

La protection de la confidentialité des données fournies par les répondants constitue une priorité pour Statistique Canada et est exigée explicitement par la Loi sur la statistique. Les données administratives, qui sont recueillies par d'autres organismes à des fins diverses et que le Bureau acquiert afin d'éviter les contacts inutiles avec les répondants, font aussi l'objet de mesures de protection. Il ne fait aucun doute que la capacité du Bureau de recueillir des données et de produire des statistiques fiables serait sérieusement compromise si les répondants n'avaient pas l'assurance que les données se rapportant à chaque personne ne seront pas divulguées, ni de propos délibéré à d'autres organismes, ni par inadvertance dans les tableaux statistiques publiés. La protection de la confidentialité des données relatives à chaque personne est prise très au sérieux.

Le Bureau produit surtout des tableaux de toutes sortes, de même que des fichiers et des microdonnées à grande diffusion (à l'occasion); dans chaque cas, les méthodes de protection sont différentes. Statistique Canada a contribué à

l'élaboration de mesures de protection de la confidentialité des données figurant dans les tableaux.

Les tableaux peuvent être de deux types : tableaux des fréquences et tableaux des grandeurs. Dans le premier cas, il s'agit simplement de compter le nombre de membres d'une population qui se classent dans des catégories données. Voici un exemple de tableau des fréquences dans lequel tous les membres d'un certain groupe sont classés selon le sexe et le revenu.

		SEXE		
		M	F	TOTAL
Revenu (milliers de dollars)	Moins de $ 25	0	2	2
	$ 25 à $ 35	6	0	6
	$ 35 à $ 50	15	14	29
	$ 50 à $ 70	13	3	16
	Plus de $ 70	2	1	3
	TOTAL	36	20	56

Exemple de tableau des fréquences

Il est probable que la femme ayant le revenu le plus élevé est facilement identifiable et qu'elle n'aimerait pas que son salaire soit révélé. Il en est de même pour les deux femmes ayant un faible revenu.

Afin de régler de tels problèmes, Statistique Canada a mis au point la technique de l'arrondissement aléatoire. Nous connaissons tous la méthode de l'arrondissement systématique qui consiste, comme dans le calcul des pourcentages, à

arrondir au nombre entier le plus près. Lorsqu'il s'agit de l'âge, cependant, il faut arrondir au nombre entier inférieur et donner l'âge au dernier anniversaire. Selon la méthode de l'arrondissement aléatoire, les nombres sont habituellement arrondis à un multiple de 5 près, et on lance un dé imaginaire à 5 côtés afin de déterminer s'il faut arrondir vers le haut ou vers le bas. Ainsi, les nombres se terminant par 0 ou 5 ne changent pas et,

	1		4		0	
	2		3		0	
	3		2		0	Sinon, il sera
si le dernier	4		1	chances sur	0	arrondi au
chiffre du	6	il a	4	5 d'être	5	multiple de
nombre est un	7		3	arrondi à	5	5 supérieur.
	8		2		5	
	9		1		5	

La validité statistique de la méthode susmentionnée repose sur le fait que, si nous arrondissions un grand nombre de fois les nombres figurant dans un tableau donné, il suffirait de faire la moyenne des nombres arrondis pour obtenir le tableau exact. S'il nous arrivait de publier le même tableau plusieurs fois et de refaire chaque fois l'arrondissement aléatoire, une personne persévérante pourrait découvrir le tableau original en faisant la moyenne de tous les tableaux publiés.

Voici une version du tableau donné en exemple où figurent des nombres arrondis

		SEXE		TOTAL
		M	F	
Revenu (milliers de dollars)	Moins de $ 25	0	5	5
	$ 25 à $ 35	5	0	5
	$ 35 à $ 50	15	15	30
	$ 50 à $ 70	15	0	15
	Plus de $ 70	5	5	10
	TOTAL	40	25	65

Tableau des fréquences selon la méthode de l'arrondissement aléatoire

Maintenant, la femme ayant un revenu élevé et les deux femmes à faible revenu ne sont plus identifiables, mais le nombre total de membres de ce groupe est très différent du total original (56). Il serait possible, et peut-être même souhaitable, d'arrondir les totaux séparément, mais alors la somme des lignes et des colonnes ne correspondrait pas nécessairement aux totaux indiqués.

Pour faire face à ce problème, on a élaboré la technique de l'arrondissement aléatoire contrôlé. Cette méthode permet d'arrondir les nombres au multiple de 5 supérieur ou inférieur et de maintenir les totaux à plus ou moins 5 unités des nombres originaux. Voici une version du tableau original auquel la technique d'arrondissement aléatoire contrôlé a été appliquée.

		SEXE		
		M	F	TOTAL
Revenu (milliers de dollars)	Moins de $ 25	0	5	5
	$ 25 à $ 35	5	0	5
	$ 35 à $ 50	15	15	30
	$ 50 à $ 70	15	0	15
	Plus de $ 70	0	0	0
	TOTAL	35	20	55

Tableau des fréquences selon la méthode de l'arrondissement aléatoire contrôlé

Il s'agit en fait d'un tableau qui ressemble grandement à l'original, qui n'est pas très différent du premier tableau arrondi et qui est beaucoup plus acceptable en ce qui a trait au totaux.

La technique de l'arrondissement contrôlé, qui consiste à additionner les nombres inscrits dans les colonnes, à arrondir tous les nombres, puis à soustraire les diverses sommes les unes des autres afin de déterminer chacun des nombres inscrits arrondis, est aussi applicable aux tableaux plus complexes et de plus grande taille, qui renferment trois variables ou plus; toutefois, dans les tableaux à plusieurs variables, les totaux peuvent différer de beaucoup plus que de 5 unités des totaux originaux.

Statistique Canada est en outre considéré comme un chef de file dans le domaine de l'élaboration de techniques automatisées applicables aux tableaux des grandeurs. Voici un exemple de tableau des grandeurs.

	EST DU CANADA	CENTRE DU CANADA	OUEST DU CANADA	CANADA
Pêche	66	17	50	133
Chasse	1	25	45	71
TOTAL	67	42	95	204

Recettes totales (millions de dollars)

Exemple de tableau des grandeurs

Ici, il ne s'agit pas de compter quoi que ce soit, mais plutôt d'indiquer les recettes totales réalisées par l'ensemble des entreprises appartenant à un secteur d'activité donné. Bien entendu, quelqu'un peut très bien savoir le nombre d'entreprises faisant partie de ce secteur, et c'est là qu'un problème se pose. Supposons qu'il y a seulement trois entreprises oeuvrant dans le domaine de la pêche dans l'Est du Canada (ce n'est pas le cas) et que l'information "commerciale" nous permet de savoir que deux d'entre elles sont de très petite taille. Nous pouvons ainsi déterminer approximativement (disons à 10 % ou à 20 % près), les recettes réalisées par la troisième entreprise, ce qui risque de ne pas plaire à l'entreprise en question. La case où figure cette donnée est alors déclarée "confidentielle". Nous refusons alors de la publier et proposons le tableau suivant.

	EST DU CANADA	CENTRE DU CANADA	OUEST DU CANADA	CANADA
Pêche	X	17	50	133
Chasse	1	25	45	71
TOTAL	67	42	95	204

Recettes totales (millions de dollars)

Tableau dont une case confidentielle a été supprimée

De toute évidence, il s'agit d'une mesure inutile puisque quelques opérations arithmétiques de base permettent de déterminer la valeur manquante. Il faut aussi supprimer les valeurs figurant dans certaines autres cases afin de protéger la case confidentielle. On obtient donc le tableau suivant.

	EST DU CANADA	CENTRE DU CANADA	OUEST DU CANADA	CANADA
Pêche	X	X	50	133
Chasse	X	X	45	71
TOTAL	67	42	95	204

Recettes totales (millions de dollars)

Tableau comportant des suppressions complémentaires

Plus les tableaux sont complexes, plus le problème de la protection des données est difficile à régler. La solution au problème se compose de trois éléments : les règles servant à déterminer si une case est confidentielle; la vérification du schéma de publication, afin de déterminer s'il y a divulgation; la détermination, à l'aide de moyens automatisés, des autres cases qu'il faut supprimer afin d'empêcher la divulgation, c'est-à-dire l'établissement du schéma de suppression complémentaire.

Pour établir la confidentialité d'une case, on applique habituellement la règle de la concentration. Par exemple, une case peut être jugée confidentielle si l'apport des trois répondants les plus importants représente plus de 80 % du total. Habituellement, ce critère est lui-même considéré comme étant confidentiel par les organismes statistiques. Il est fondé sur le principe que le regroupement avec d'autres répondants dont l'apport n'est pas trop différent aide à protéger la confidentialité des données relatives à chaque répondant. Ce genre de critère de confidentialité se prête à l'analyse mathématique, ce qui permet l'élaboration de techniques automatisées susceptibles de régler ce problème.

Par schéma de suppression, on entend la façon dont les X sont répartis dans la publication. Il peut être déterminé manuellement ou il peut être identique au schéma utilisé le mois précédent ou l'année précédente pour la même enquête. Dans ce dernier cas, il peut être souhaitable de vérifier le schéma en fonction des données courantes afin de déceler les possibilités de divulgation par approximation. Un programme élaboré par Statistique Canada effectue les opérations arithmétiques nécessaires pour déterminer les valeurs minimales et maximales que peut prendre chaque case supprimée, compte tenu des valeurs restantes. Une vérification du dernier tableau dont quatre cases ont été supprimées permettrait de produire le tableau suivant.

	EST DU CANADA	CENTRE DU CANADA	OUEST DU CANADA	CANADA
Pêche	41 - 67	16 - 42	50	133
Chasse	0 - 26	0 - 26	45	71
TOTAL	67	42	95	204

Recettes totales (millions de dollars)

Tableau vérifié indiquant les fourchettes de valeurs possibles

Ici, la vérification peut se faire à l'aide de quelques opérations arithmétiques simples, mais elle nécessite habituellement des calculs complexes. Si les limites coïncident, il y a divulgation. Si elles sont très rapprochées, disons à un faible pourcentage l'une de l'autre, il peut y avoir divulgation par approximation. Nous devons toujours supposer qu'une personne astucieuse et intéressée essaie d'obtenir des données confidentielles à partir de nos tableaux et qu'elle a grand besoin de ces données. Il faut que les utilisateurs apprécient à leur juste valeur les compétences et le savoir-faire dont il faut faire preuve afin de leur cacher certains renseignements.

La détermination d'un schéma de suppression complémentaire visant à protéger un ensemble de valeurs confidentielles pose un problème plus difficile. Lorsqu'il s'agit de très petits tableaux, on peut simplement examiner toutes les possibilités. Cependant, les publications d'envergure peuvent contenir plusieurs milliers ou dizaines de milliers d'inscriptions et dans ce cas, même les méthodes automatisées de recherche exhaustive sont peu pratiques. On a adopté des techniques de recherche opérationnelle et des programmes ont été élaborés afin de régler le problème de la détermination du schéma de publication.

3. CONTRÔLE ET IMPUTATION

La correction des erreurs et des omissions identifiables dans les données d'enquête peut se révéler très coûteuse. Communiquer à nouveau avec les répondants lorsque la collecte des données est terminée est une tâche fastidieuse qui est beaucoup plus coûteuse que la collecte initiale, quand elle n'est pas tout simplement impossible. Il est cependant inacceptable de laisser des données erronées figurer dans les tableaux. Elles entravent sérieusement l'analyse des données, jettent le doute sur les conclusions et causent beaucoup d'inconvénients aux utilisateurs.

Lorsqu'il se présente un problème d'erreurs dans les réponses, on adopte habituellement une stratégie de "contrôle et imputation". Le contrôle permet de déceler les données erronées, incohérentes ou manquantes. L'imputation vise à modifier les réponses de sorte que l'enregistrement soit exempt d'erreur, c'est-à-dire qu'il soit accepté au contrôle.

Le contrôle consiste seulement à vérifier les liens (aussi appelés contrôles, ce qui prête à confusion) entre les éléments d'une réponse. Cette dernière est jugée valide si chacun de ces éléments satisfait à certaines exigences. Voici un certain nombre de contrôles qui pourraient être applicables dans le cadre d'une enquête sur les employés de Statistique Canada.

Âge inférieur à 80 ans
Âge supérieur à 15 ans
Si le niveau d'instruction = universitaire, alors l'âge est supérieur à 20 ans.

Si certains de ces liens ne s'appliquent pas, la réponse est rejetée au contrôle. Il faut alors déterminer quels éléments doivent être considérés comme étant erronés, ce qui n'est pas toujours évident. Les éléments erronés ou manquants sont ensuite imputés d'une façon acceptable d'un point de vue statistique, ce qui signifie que, même si l'imputation ne permet pas de trouver la

réponse exacte dans chaque cas, elle se traduit par une amélioration de la qualité globale des données totalisées.

Pourquoi un organisme d'enquête tel que Statistique Canada corrige-t-il les données plutôt que de les laisser telles qu'elles ont été fournies par les répondants? La correction des données, surtout lorsqu'elle consiste à modifier les enregistrements, peut ressembler à une falsification des données. Il y a pourtant plusieurs bonnes raisons justifiant le recours à cette méthode.

Dans le cadre de l'analyse des données, il faut inévitablement effectuer des corrections en fonction des données manquantes. Supposons que 2 % des personnes recensées n'ont pas indiqué leur âge. Les démographes doivent répartir ces personnes selon les groupes d'âge pour être en mesure d'établir des estimations ou des projections. S'ils utilisent seulement les données valides sans tenir compte des données manquantes, ils effectuent une correction, mais cette dernière est de piètre valeur puisqu'elle se traduit par une sous-estimation de la population totale. Il aurait été préférable et très simple de se servir du compte total comme une estimation de la taille de la population et d'utiliser les données exemptes d'erreur comme source de renseignements sur la répartition par âge. Cette méthode de correction, qui est une forme simple de pondération, est préférable au rejet pur et simple des données incomplètes mais pourrait encore être améliorée.

Chaque utilisateur pourrait corriger lui-même les données qu'il obtient. Cette méthode se révélerait toutefois inefficace. Chaque utilisateur devrait déployer des efforts considérables pour effectuer ses propres corrections, qui seraient différentes de celles apportées par d'autres utilisateurs. En fait, la correction des données avant leur diffusion est un service offert aux utilisateurs, qui évite à ces derniers d'avoir à régler des problèmes relatifs aux données et qui assure l'uniformité entre les utilisateurs.

L'organisme d'enquête peut en général faire mieux que les utilisateurs lorsqu'il s'agit de corriger les données. Le producteur de données a accès à tous les

renseignements confidentiels qui ne peuvent être divulgués à l'utilisateur, et il peut acquérir un savoir-faire spécial dans le domaine des techniques de correction des données.

Pour toutes les raisons susmentionnées, Statistique Canada corrige depuis longtemps les questionnaires remplis. À l'origine, la correction était effectuée manuellement, et encore aujourd'hui, de vastes opérations de correction manuelle sont entreprises dans le cadre de certaines enquêtes. Depuis la venue des ordinateurs rapides, on a assisté à l'automatisation de la plupart des opérations de ce genre effectuées dans le cadre des principales enquêtes. Comme c'est souvent le cas, les opérations automatisées étaient effectuées à l'origine selon les procédures manuelles. Il a fallu élaborer des règles complexes pour remplacer ces procédures, souvent incomplètes ou non documentées, qu'utilisait le personnel des opérations de bureau. Dans certains cas, les règles complexes ont cependant donné lieu à des programmes informatiques complexes peu maniables, étant donné qu'ils peuvent détecter toutes les erreurs et toutes les exceptions possibles.

3.1 Contrôle et imputation des données qualitatives

À l'occasion du recensement de 1971, l'ampleur et la complexité du processus de contrôle et d'imputation ont retardé de beaucoup la diffusion des données définitives.

Pour remédier à cette situation, Statistique Canada a élaboré une méthode standard de traitement des variables qualitatives (comme l'âge, le niveau d'instruction et l'état matrimonial) qui composent la majeure partie des données de recensement. Il s'agit du système Can-Edit, qui a été utilisé pour la première fois à l'occasion du traitement des données du recensement de 1976. Ce système a permis de traiter les données de ce recensement beaucoup plus rapidement que celles du recensement de 1971, et de mieux comprendre tout ce qu'impliquent le contrôle et l'imputation des données d'enquête.

Le système Can-Edit s'est révélé très efficace mais ne s'appliquait pas dans tous les cas. Il ne permettait pas de traiter efficacement les données quantitatives ni les données qualitatives figurant dans une longue liste de codes. Il imposait en outre des critères beaucoup trop restrictifs quant aux données pouvant être imputées. Dans certains cas, il est facile de distinguer la bonne réponse, qui est évidente, de la mauvaise.

En 1981, le système Can-Edit s'est encore une fois révélé très efficace. Il a permis de réaliser des économies importantes au titre des ressources consacrées aux travaux d'élaboration et de réduire de beaucoup le temps de traitement. À cette occasion, on a aussi élaboré un deuxième système, appelé Spider, visant à traiter les variables qui ne peuvent l'être à l'aide du système Can-Edit. Le système Spider reprend certains des principes de la méthode établie à l'origine. Ce système est plus flexible mais moins rigoureux, et il fait plus appel aux jugements de valeur quant à la façon dont les imputations doivent être faites.

En 1986, les deux systèmes ont été réutilisés, ce qui a permis de réaliser d'autres gains substantiels en matière de productivité.

Les principes établis pour l'imputation des données étaient les suivants.

1) Il faut faire en sorte que les données figurant sur chaque enregistrement soient acceptées à tous les contrôles en modifiant l'information le moins possible.

2) Dans la mesure du possible, il faut conserver les caractéristiques relatives à la répartition que présentent les données.

3) Les règles d'imputation doivent découler des règles de contrôle.

Les deux premiers principes visaient à accroître au maximum la qualité des données tout en réduisant au minimum les modifications devant être

apportées par l'enquêteur aux données recueillies. Le troisième principe avait pour objectif de préserver l'uniformité des enregistrements et d'éviter ainsi les cas où l'apport de modifications à des enregistrements entraîne d'autres erreurs, c'est-à-dire les cas qui nécessitent des corrections plus importantes que celles qui étaient prévues ou les cas qui, au pire, peuvent se traduire par un cercle vicieux.

Can-Edit est un système général, c'est-à-dire qu'il n'est pas adapté à une enquête précise. Il peut donc être réutilisé à l'occasion de chaque recensement sans qu'il soit nécessaire de le modifier, ce qui permet de réaliser des économies considérables au titre des ressources consacrées aux travaux d'élaboration. De tels logiciels sont chose courante aujourd'hui en raison de l'augmentation rapide des applications micro-informatiques. Il y a quinze ans, lorsque l'on a entrepris l'élaboration de ce système, il s'agissait de travaux très avant-gardistes.

Le système Can-Edit se compose des modules suivants.

a) L'analyseur des règles de contrôle

Ce module permet de vérifier si les règles de contrôle proposées sont logiquement compatibles et ne renferment aucune redondance. Il fournit en outre au sélectionneur et pour les étapes suivantes du système toutes les règles de contrôle que suppose l'ensemble initial, ce qui garantit que le contrôle se fera si des données sont manquantes et permet au sélectionneur d'effectuer une contre-vérification des règles. Lorsque la vérification est terminée, les règles sont intégrées dans le système.

b) Le contrôle

L'étape du contrôle permet de vérifier si chaque enregistrement subit avec succès tous les contrôles et, si ce n'est pas le cas, d'indiquer à quels contrôles il a été rejeté.

c) La détermination de la zone à imputer

Ce module sert à déterminer les zones (éléments ou variables) qu'il faut modifier afin d'obtenir un enregistrement exempt d'erreur. De par sa conception, l'algorithme permet de déterminer, grâce à l'ensemble complet de contrôles provenant de l'analyseur des règles de contrôle, le nombre minimal de modifications qu'il faut apporter à chaque enregistrement.

d) L'imputation

L'imputation est une méthode "hot deck", ce qui signifie que la valeur des variables qu'il faut modifier sur un enregistrement erroné est remplacée par une valeur provenant d'un bon enregistrement, ou enregistrement "donneur". Le système permet de trouver un enregistrement donneur qui correspond le plus possible à l'enregistrement erroné. Toutes les variables à changer sont modifiées en même temps. Dans le cadre du processus d'appariement servant à déterminer le donneur, toutes les variables faisant partie des enregistrements rejetés au contrôle sont utilisées mais ne sont pas modifiées, ce qui garantit que l'enregistrement modifié sera accepté au contrôle lorsque l'imputation sera terminée.

Si aucun donneur ne peut être trouvé, le système impute chaque variable séparément tout en veillant à ce que l'enregistrement définitif soit accepté au contrôle.

Le système Can-Edit a ouvert la voie à tous les travaux subséquents d'élaboration de systèmes généraux de contrôle et d'imputation à Statistique Canada, et a eu sans aucun doute une incidence marquée sur les travaux entrepris à cet égard par les organismes statistiques d'autres pays. Comme il a été réutilisé à l'occasion des recensements de 1981 et de 1986 sans avoir subi aucune modification, sa robustesse n'est plus à démontrer.

3.2 Contrôle et imputation des données numériques

Lorsque la réponse à une question est une quantité, la méthode utilisée dans le cadre du système Can-Edit pour les données qualitatives ne s'applique pas. La raison en est qu'il est impossible de répartir des quantités, pour une raison quelconque, selon un nombre commode de catégories. Si l'on souhaite connaître les recettes totales réalisées dans un certain secteur d'activité, il faut additionner les recettes de toutes les entreprises de ce secteur, mais il est impossible d'additionner des catégories.

Lorsqu'il s'agit de quantités, les contrôles sont surtout exprimés sous forme de fourchettes. Pour donner un exemple très simple, supposons que l'enquête porte sur des entreprises et que nous désirons connaître les ventes totales et les bénéfices exprimés en dollars.

Les contrôles pourraient ressembler à ceux-ci.

VENTES	>	0
VENTES	≤	$ 100,000
BÉNÉFICES	≥	0 (nous sommes optimistes)
BÉNÉFICES	≤	40 % des VENTES
BÉNÉFICES	≤	$ 20,000

Voici la représentation graphique de ces contrôles.

BÉBÉFICES
(milliers
de dollars)

DONNÉES VALIDES
DANS CETTE ZONE
SEULEMENT

VENTES (milliers de dollars)

Exemple de contrôles de données numériques

Toute combinaison de bénéfices et de ventes qui ne se situe pas dans la zone hachurée sera rejetée au contrôle.

Supposons les données suivantes,

VENTES = $30,000

BÉNÉFICES = $15,000

Ces données seraient rejetées au moment du contrôle BÉNÉFICES $\leq$ 40 % des VENTES. Nous pourrions vérifier sur le questionnaire si les données ont été saisies correctement ou téléphoner au répondant pour les confirmer. Ces deux possibilités sont assez coûteuses et prennent beaucoup de temps; de plus, nous ne voulons pas déranger le répondant plus qu'il n'est nécessaire. Nous décidons plutôt de modifier les VENTES ou les BÉNÉFICES.

Il convient de prendre note que, s'il arrivait qu'un grand nombre de réponses soient rejetées au contrôle, nous soupçonnerions que quelque chose ne va pas avec le contrôle ou la procédure et examinerions la situation en conséquence. Cependant, si le nombre de rejets est faible, notre expérience peut nous amener à apporter la correction appropriée même si la donnée ainsi obtenue n'est pas vraiment exacte. La différence par rapport à la valeur totale des VENTES ou des BÉNÉFICES établie à partir de toutes les réponses sera probablement négligeable.

On peut aussi alléguer que, si la différence est négligeable, il est inutile d'apporter une correction. Cependant, comme les données peuvent être totalisées de différentes façons, des incohérences peuvent se faire jour dans les tableaux.

La situation devient beaucoup plus complexe lorsqu'il s'agit de données plus réalistes, en raison du nombre accru d'éléments de données et de contrôles. Il est alors impossible de donner une représentation graphique de la situation, mais les problèmes demeurent les mêmes.

a) Il faut déceler les réponses rejetées au contrôle.

b) Il faut déterminer les éléments de réponse qu'il faudrait modifier afin d'obtenir des réponse acceptées au contrôle.

c) Il faut déterminer les valeurs qui remplaceront celles devant être modifiées.

À la fin des années soixante-dix, on a conçu un prototype de système destiné à régler le problème du contrôle et de l'imputation des données numériques, qui a pour nom NEIS (système de contrôle et d'imputation des données numériques).

S'il s'agit de contrôles linéaires, c'est-à-dire de contrôles exprimés sous la forme

$$aX + bY + cZ + \ldots\ldots \leq d$$

où X, Y et Z sont des éléments de données quantitatives et a, b et c sont des constantes (c'est-à-dire des nombres fixes), et si nous convenons de mofidier aussi peu d'éléments que possible lorsqu'une réponse est rejetée, le système effectuera les tâches suivantes :

i) analyser les contrôles afin de s'assurer qu'ils ne sont pas contradictoires ni redondants;
ii) appliquer les contrôles aux données;
iii) déterminer les éléments à modifier à l'aide d'une méthode qui nécessite seulement un ensemble minimal de contrôles;
iv) rechercher, parmi les bonnes réponses, une réponse qui ressemble à la réponse rejetée et à partir de laquelle on pourra copier certains éléments servant à remplacer les éléments "erronés" figurant dans l'enregistrement rejeté.

La dernière étape est l'imputation. Au contraire de Can-Edit, le système NEIS ne choisit pas au hasard parmi les possibilités d'appariement; il détermine d'abord les éléments à apparier, puis repère les quinze (ce nombre peut changer) réponses complètes qui se rapprochent le plus de ces éléments. Les réponses sont ensuite soumises une par une à un test visant à déceler les réponses susceptibles de fournir les données nécessaires. Le test est très simple; si les réponses modifiées sont acceptées au contrôle, c'est que le processus d'imputation a été couronné de succès.

Le système NEIS est assez perfectionné et présente en fait plus de caractéristiques que les autres systèmes susmentionnés. Toutefois, il ne s'agit que d'un prototype qui, bien entendu, ne peut être considéré à lui seul comme "le" système de contrôle et d'imputation de données numériques applicable dans tous les cas. On a toutefois emprunté un grand nombre de caractéristiques de ce système afin de les intégrer dans d'autres systèmes.

Nous avons récemment entrepris un projet d'élaboration d'un système général de contrôle et d'imputation (GEIS). Ce système comprendra presque toutes les caractéristiques du système NEIS et offrira plus de possibilités en ce qui concerne les genres de contrôles et les méthodes d'imputation. Le problème est que la base mathématique sur laquelle repose le système NEIS ne peut être élargie pour inclure un bon nombre des possibilités souhaitées. Par exemple, si les contrôles ne sont pas linéaires, ils ne peuvent pas être analysés rapidement et ne permettent pas le recours à la méthode de localisation des erreurs du NEIS. Il faudrait établir des procédures spéciales à cet égard. Ces problèmes constituent des obstacles importants qu'il nous faudra surmonter au cours des prochaines années.

4. ANALYSE DES DONNÉES PROVENANT D'ENQUÊTES DONT LE PLAN D'ÉCHANTILLONNAGE EST COMPLEXE

L'analyse des données est le processus qui consiste à tirer des conclusions à partir des données. C'est l'une des principales fonctions des statisticiens,

difficile, en raison du manque de données, d'établir des comparaisons entre les différents taux de chômage selon le groupe d'âge et le sexe. Cet échantillon entraînerait des dépenses élevées puisqu'il faudrait joindre des personnes réparties dans l'ensemble du pays.

Selon la méthode actuelle de réalisation de l'enquête sur la population active, on fait appel non pas à une liste de personnes, mais à une liste de secteurs. Des secteurs sont échantillonnés dans des proportions différentes selon les provinces; ces secteurs sont ensuite partagés en segments, qui font à leur tour l'objet d'un échantillonnage et au sein desquels des ménages sont finalement échantillonnés. L'échantillon ainsi obtenu est appelé échantillon à plusieurs degrés. Toutes les personnes de plus de 15 ans faisant partie d'un ménage échantillonné participent à l'enquête. Il est ainsi possible d'éviter un grand nombre de problèmes opérationnels et de réduire les coûts. Nous pouvons garantir que l'échantillon est réparti de façon qu'il soit possible d'établir des estimations raisonnables de statistiques pertinentes pour toutes les provinces. Il est cependant beaucoup plus difficile d'établir des comparaisons et d'analyser les données en raison de la complexité du plan d'échantillonnage.

Il existe deux types d'analyses utilisées couramment : l'analyse de tableau de contingence et l'analyse de la régression. L'analyse de tableau de contingence consiste à classer la population selon des facteurs présentant un intérêt particulier et à déterminer, d'après la proportion de la population se trouvant dans chaque case, quels pourraient être les liens entre ces facteurs. Voici un exemple de tableau de contingence.

qu'ils soient théoriciens ou praticiens. Grâce à l'analyse des données, on peut déterminer le lien entre la consommation de tabac et les maladies cardiaques, les conditions dans lesquelles un procédé industriel est le plus efficace ou les liens existant entre l'emploi et des variables telles que l'âge, le niveau d'instruction et l'état matrimonial.

Dans le cadre des méthodes d'analyse des données qui ont été mises au point au cours des quelque cent dernières années, on posait généralement comme hypothèse que les observations sont faites complètement au hasard. Les unités observées, comme les malades participant à la mise à l'essai d'un traitement, les étudiants prenant part à une expérience psychologique ou les parcelles de terrain utilisées pour une expérience agricole, constituent les représentants d'une "importance égale" d'une population donnée et n'influent pas les unes sur les autres. Ainsi, dans le cadre des techniques statistiques auxquelles elles sont soumises, les unités sont toutes traitées de la même façon.

Lorsqu'il s'agit d'enquêtes-échantillon, on ne peut pas se payer le luxe de traiter de la même façon toutes les unités observées. Il faut obtenir le maximum de précision au coût le plus bas. Par exemple, supposons que l'enquête sur la population active est réalisée selon une méthode qui se prête davantage aux techniques statistiques classiques. Il faudrait d'abord mettre à jour chaque mois la liste d'adresses de toutes les personnes âgées de plus de 15 ans au Canada. Ensuite, il faudrait constituer un échantillon aléatoire (ce qui signifie que tous les échantillons de même taille sont semblables) comprenant (disons) 100,000 de ces personnes, par exemple en tirant leur nom un à un d'un chapeau. Il nous resterait ensuite à calculer le taux de chômage au Canada, dans chaque province, pour chaque groupe d'âge, etc., et il serait relativement facile d'établir des comparaisons et de déterminer si le taux de chômage du Québec est différent de celui de l'Ontario.

L'échantillon présenterait malheureusement certains défauts évidents. Les provinces de petite taille (comme l'Île-du-Prince-Édouard) ne seraient pas nécessairement bien représentées dans l'échantillon, et il serait probablement

		PERSONNES OCCUPÉES	CHÔMEURS	TOTAL	
HOMMES	MOINS DE 25 ANS	100	20	120	460
	PLUS DE 25 ANS	300	40	340	
FEMMES	MOINS DE 25 ANS	150	25	175	570
	PLUS DE 25 ANS	350	45	395	
TOTAL		900	130		1030

Exemple de tableau de contingence

Il s'agit d'un "Tableau à trois variables". Si les données proviennent d'un échantillon aléatoire, l'analyse du tableau ne pose aucun problème. Par contre, si les inscriptions sont des estimations établies à partir d'un plan d'échantillonnage complexe, nous ne pouvons pas utiliser les mêmes techniques.

L'analyse de la régression, quant à elle, consiste à représenter à l'aide d'une équation les liens qui existent entre une variable donnée et d'autres variables.

Exemple :

Q.I. = a + b (QI de la mère) + c (QI du père) + d (revenu de la famille)

Afin d'estimer les valeurs a, b, c et d et de vérifier si le modèle est approprié, il nous faut obtenir des données auprès d'un grand nombre de personnes (et de leurs parents), puis avoir recours à des astuces statistiques pour établir

nos procédures. Encore une fois, s'il s'agit d'une expérience psychologique, il ne se pose aucun problème particulier; cependant, si les données proviennent d'une enquête complexe, nous faisons face à certains problèmes.

Statistique Canada effectue depuis la fin des années soixante-dix des travaux de recherche visant à appliquer les méthodes classiques aux données d'enquête provenant de plans d'échantillonnages complexes. Le Bureau a élaboré un certain nombre de méthodes qui sont maintenant utilisées partout dans le monde pour toute une gamme d'applications, ce qui a confirmé la validité globale des conclusions auxquelles étaient arrivés les chercheurs et les analystes qui utilisaient auparavant les procédures classiques, lesquelles n'étaient pas vraiment applicables.

A titre d'exemple, examinons les résultats d'une étude analytique des données provenant de l'enquête sur l'éducation des adultes, enquête supplémentaire de l'enquête sur la population active réalisée en février 1986. Cette analyse visait à déterminer les facteurs contribuant au fait qu'une femme donnée suive ou non des cours de formation à temps plein. Les facteurs en question étaient l'âge, le niveau d'instruction, la catégorie de travailleurs et l'état matrimonial.

Dans le cas qui nous occupe, on s'est servi de la méthode analytique appelée régression logistique, qui a été appliquée de plusieurs façons. Nous n'allons cependant considérer que deux cas d'application : l'analyse effectuée sans tenir compte du plan d'échantillonnage et l'analyse réalisée en tenant compte de ce plan.

Le modèle applicable dans cet exemple est le suivant : si P est la probabilité qu'une femme donnée suive un cours de formation, alors P peut être exprimé en fonction de son âge, de son niveau d'instruction, de sa catégorie de travailleurs et de son état matrimonial sous la forme suivante :

$$\log \frac{P}{1 - P} = a + b \text{ (groupe d'âge)} + c \text{ (niveau d'instruction)} + d \text{ (catégorie de travailleurs)} + e \text{ (état matrimonial)}.$$

Il est quelque peu difficile de l'exprimer explicitement, mais l'analyse permet d'établir, pour les différents groupes d'âge, niveaux d'instruction, etc., une estimation de la contribution d'un facteur donné à la probabilité qu'une femme suive des cours de formation. Pour chaque estimation de la contribution, on a aussi établi une estimation de l'erreur d'échantillonnage. Si certaines estimations de la contribution semblent manquer, c'est parce que les estimations pour les autres catégories du même facteur ont été établies en fonction de celles-ci, c'est-à-dire que les estimations de la contribution sont évaluées d'après une norme, qui prend la valeur zéro. Par exemple, les estimations de la contribution de l'âge sont établies en fonction du groupe d'âge 18-20 ans, qui sert de norme en l'occurrence.

Facteur	Catégorie	Estimations			
		En tenant compte du plan		Sans tenir compte du plan	
		Estimation	Écart type	Estimation	Écart type
Âge	18-20 21-24 25-29 30-34 35-39 40-45				
Niveau d'instruction	Primaire Secondaire (en partie) Secondaire Postsecondaire (en partie) Collégial Universitaire				
Catégorie de travailleurs	Occupée Chômeuse Travailleuse à temps partiel Inactive				
État matrimonial	Mariée Célibataire Séparée/Divorcée/Veuve				
Constante					

Comparaison entre les estimations établies en tenant compte du plan et celles établies sans tenir compte du plan
Femmes qui suivent des cours de formation à temps plein

Exemple d'analyse de données provenant d'enquêtes complexes

Dans l'exemple ci-devant, on peut voir qu'il existe d'assez grandes différences entre les estimations, comme entre celles pour les catégories d'âge 21 à 24 ans, 25 à 29 ans et 30 à 34 ans. On ne peut pas affirmer que ces différences sont "significatives d'un point de vue statistique" parce que les écarts types sont aussi plutôt élevés. En fait, cet exemple démontre que le plan est, dans une large mesure, indépendant des variables à l'étude. Ce renseignement est bon à connaître, mais il ne peut pas être tenu pour acquis et n'est certainement pas toujours vrai.

5. DÉSAISONNALISATION DES SÉRIES CHRONOLOGIQUES

Lorsque des statistiques portant sur des caractéristiques économiques ou sociales, comme l'indice des prix à la consommation et le taux de chômage, sont recueillies régulièrement pendant une longue période, on voit habituellement apparaître une forte tendance saisonnière dans les données. Cette tendance saisonnière est liée aux changements plus ou moins prévisibles qui surviennent au cours d'une année; elle empêche l'étude des tendances ou des cycles à long terme et l'établissement de comparaisons à court terme entre les périodes.

Les personnes qui s'occupent d'analyser les séries chronologiques considèrent généralement que ces dernières comprennent trois composantes : la composante tendancielle ou cyclique, la composante saisonnière et la composante irrégulière; cette dernière est imprévisible et représente les événements isolés ou les fluctuations aléatoires qui marquent une période donnée.

Afin d'être en mesure d'établir des comparaisons éloquantes entre les périodes, les analystes désirent faire disparaître la composante saisonnière. Le processus utilisé est la désaisonnalisation. L'élaboration et l'amélioration des méthodes de désaisonnalisation continuent de faire l'objet de travaux de recherche considérables.

Le graphique 3 fait ressortir les différences entre les séries définitives et provisoires pour X-11-ARIMA et pour X-11. Dans le cas de la méthode X-11-ARIMA, les discordances sont non seulement moins marquées, mais aussi moins systématiques. Dans le cas de la méthode X-11, les discordances pourraient laisser croire que la saisonnalité n'a pas été complètement éliminée des séries désaisonnalisées.

Statistique Canada a élaboré un logiciel fondé sur la méthode X-11-ARIMA. Ce logiciel remplit quatre fonctions de base : l'établissement de prévisions; l'estimation des composantes des séries chronologiques, y compris les fluctuations attribuables aux jours ouvrables et l'incidence de la fête de Pâques; l'agrégation des données originales et des données désaisonnalisées; la production de statistiques contribuant à la qualité des estimations.

Plus de 5,000 clients du monde entier se sont procurés le logiciel X-11-ARIMA. Des services de consultation ont été offerts à de nombreux organismes.

Le logiciel X-11-ARIMA a d'abord été conçu pour les unités centrales, mais on le retrouve aujourd'hui sous forme de procédure SAS et de micrologiciel.

6. ÉTABLISSEMENT DE BASES DE DONNÉES STATISTIQUES

6.1 Stockage efficace des données statistiques

Il existe deux différences importantes entre les exigences relatives au stockage efficace de données statistiques et celles ayant trait au stockage de données devant être utilisées à des fins opérationnelles. Dans le premier cas, les données, relativement exemptes d'erreur, sont souvent accessibles peu de temps après leur production, et font ensuite l'objet d'un grand nombre de demandes, à des fins d'analyse, portant sur quelques variables seulement, sans qu'il soit nécessaire de les modifier. Par contre, dans le cas d'un système bancaire par exemple, il faut accéder

Au cours des années soixante, le Bureau of the Census des États-Unis a élaboré la méthode de désaisonnalisation X-11, méthode par ailleurs fort utilisée. Cette méthode n'était pas sans problème; elle permettait notamment une meilleure désaisonnalisation au milieu de la série qu'aux extrémités. On s'est rendu compte de cette particularité pour la raison suivante : lorsqu'il faut ajouter des années à la série et répéter le processus de désaisonnalisation, les modifications apportées aux valeurs les plus anciennes et les plus récentes de la série désaisonnalisée sont plus importantes.

Afin de régler ce problème, Statistique Canada a apporté une modification considérable à la méthode X-11, qui est devenue X-11-ARIMA. Cette méthode consiste d'abord à extrapoler la série aux extrémités à l'aide d'un modèle statistique appelé ARIMA (moyenne mobile intégrée et autorégressive), puis à appliquer la méthode X-11.

À titre d'exemple, voici au graphique 1 le nombre de chômeurs de sexe masculin âgés de 25 ans ou plus, pour chaque mois entre janvier 1975 et décembre 1986 (la ligne pointillée). Cette série est extrapolée pour une année (les étoiles) à l'aide d'un modèle ARIMA et est ensuite désaisonnalisée (la ligne continue) au moyen de la méthode X-11-ARIMA, ce qui permet d'obtenir la tendance combinée aux composantes irrégulières.

Le graphique 2 illustre les valeurs désaisonnalisées "définitives" (obtenues à la fin de 1986) pour la période allant de juin 1982 à juin 1983 comparativement aux valeurs provisoires établies à l'aide des méthodes X-11-ARIMA et X-11 en 1983 (avant que l'on ne dispose des données pour les trois années suivantes). On peut facilement constater que les différences entre les estimations provisoires et les estimations définitives sont beaucoup plus marquées s'il s'agit de la méthode X-11 que s'il s'agit de la méthode X-11-ARIMA, ce qui signifie que cette dernière méthode de désaisonnalisation permet de faire des prévisions beaucoup plus justes.

souvent à tous les renseignements au sujet d'un compte, à l'occasion de chaque transaction effectuée dans ce compte.

Il est très avantageux de structurer les données statistiques de façon à rendre la plus efficace possible l'utilisation qui est typique à ces données.

6.2 Modèle relationnel

Au cours des années soixante-dix, une grande partie des travaux théoriques ayant trait à la conception de bases de données portaient plus précisément sur les bases de données "relationnelles". Le modèle relationnel repose sur le principe que les données doivent figurer dans des tableaux à deux entrées; chacun de ces tableaux renferme un certain nombre de variables (appelées colonnes ou domaines) et de lignes, qui représentent différentes observations ou différents enregistrements logiques. Il a été démontré qu'il est possible de réduire les structures complexes de données sous cette forme au moyen d'un ensemble approprié de transformations et d'y gagner ainsi en simplicité. Le modèle relationnel est en outre plus souple parce que les liaisons ne sont pas intégrées dans les données, mais qu'elles peuvent être refaites à volonté en fonction d'un ensemble particulier de demandes ou d'exigences relatives à la mise à jour.

Par la même occasion, il est devenu possible de déterminer un ensemble de procédures standard applicables à ces tableaux, qui permettent de manipuler les données des tableaux de diverses façons. Par exemple, le choix d'un sous-ensemble de lignes et de colonnes faisant partie d'une "relation" donnée peut se faire à l'aide d'une composante de logiciel standard, tout comme la liaison logique de deux relations données en fonction d'une commande de liaison arbitraire.

Le langage SQL (langage d'interrogation structuré) est issu de la mise en oeuvre de ces concepts; il est de plus en plus reconnu comme étant "le"

langage de la manipulation des données, de la même façon que différents langages de programmation ont été adoptés pour l'établissement d'algorithmes.

6.3 RAPID

Au début des années soixante-dix, Statistique Canada a décidé qu'il était nécessaire de remettre en oeuvre les systèmes de soutien au recensement de la population. Il a alors décidé d'adopter une technologie relative aux bases de données et a passé en revue les logiciels offerts sur le marché. À cette époque, aucun progiciel ne permettait de traiter des fichiers d'aussi grande taille que ceux du recensement. Le Bureau a donc mis sur pied un programme interne d'élaboration à cet égard. C'est ainsi que RAPID a été créé; il s'agit d'un système général de bases de données, dont voici les principales caractéristiques.

a) Il peut recevoir des fichiers établis selon le modèle relationnel, dont le nombre limite de lignes et de colonnes permises dans une relation donnée est très élevé.

b) Il est possible d'ajouter des lignes et des colonnes dans une relation donnée, ou d'en supprimer, sans qu'il soit nécessaire de la restructurer.

c) Les données sont stockées en colonnes. (Cette façon de stocker les données est appelée structure "transposée" dans la documentation à cet égard.) Cette structure offre l'avantage suivant : on peut avoir accès à toutes les valeurs d'un sous-ensemble de variables sans qu'il soit nécessaire de transférer les autres valeurs dans une mémoire rapide. Cette caractéristique convient particulièrement aux données statistiques, comme on l'a expliqué ci-dessus.

d) On a pu arriver à un équilibre entre le stockage physique très compact des données et la haute performance du traitement sur unité centrale.

Le logiciel RAPID a été utilisé avec succès à l'occasion des recensements de 1976, 1981 et 1986, et sera employé de nouveau en 1991. On s'en est aussi servi dans le cadre d'un certain nombre d'autres projets à Statistique Canada.

6.4 Coopération internationale

RAPID a suscité l'intérêt d'autres organismes; trente d'entre eux se le sont procurés. Dans bien de cas, il s'agit de bureaux nationaux de la statistique. Certains d'entre eux, dont celui de la Suède, de la Hongrie et de la République démocratique allemande, en ont fait un très grand usage.

Le logiciel a reçu le soutien officiel du Projet d'informatique statistique (parrainé par le Programme des Nations-Unies pour le développement et la Commission économique pour l'Europe), qui vise à réunir un ensemble de logiciels statistiques devant être utilisés dans les pays en voie de développement. Les pays participant à ce projet ont enrichi RAPID en fournissant un ensemble d'"opérateurs de base", qui ressemblent beaucoup aux commandes du SQL, permettant d'extraire des données d'une base de données se composant d'un ensemble de relations RAPID.

7. GÉOCARTOGRAPHIE

7.1 Codage suivant une grille géographique

Avant le recensement de 1971, il n'existait aucun système permettant l'extraction de données régionales du recensement pour des zones précisées par les utilisateurs. Ces derniers devaient déterminer approximativement

à quels groupes de secteurs de dénombrement ou à quelles unités geostatistiques normalisées de grande taille, comme les secteurs de recensement ou les municipalités, correspondait la zone qui les intéressait. De plus, ces zones "standard" variaient souvent d'un recensement à l'autre. Afin de régler ces problèmes, Statistique Canada a élaboré le système de stockage et d'extraction des données codées suivant une grille géographique (GRDSR), qui permet le stockage des données du recensement à un niveau géograhique très bas, le côté d'îlot, dans les régions urbaines. Grâce à ce système, les utilisateurs peuvent obtenir des données personnalisées portant sur les zones spécifiées et sur plusieurs recensements.

Le système GRDSR et le fichier du réseau des rues qui peut le recevoir, appelé fichier principal de la région (FPR), ont attiré l'attention sur la scène internationale. En outre, les secteurs public et privé du Canada y ont de plus en plus recours pour la planification, l'établissement de cartes et la répartition de véhicules automobiles.

7.2 Établissement de cartes pour le recensement

L'un des problèmes auxquels se heurtent les organismes chargés d'effectuer un recensement national de la population et du logement est l'obtention d'un ensemble complet de cartes à jour et faciles à lire que pourront utiliser les recenseurs pour la livraison des questionnaires le jour du recensement. En vue du recensement de 1986, on avait élaboré un système de production de cartes assistée par ordinateur.

Les résultats ont été assez satisfaisants : la production de 1,200 cartes de secteurs de recensement à l'aide du FPR de la Division de la géographie a été réalisée à temps et dans les limites permises par le budget. Le Bureau a retiré certains avantages de ce système, comme une plus grande compatibilité entre les différentes bases de données géographiques et une réduction des coûts de production pour les nouvelles zones. Pour le

recensement de 1991, les cartes des secteurs de recensement seront produites par ordinateur toutes les fois que le FPR permettra d'accéder aux données nécessaires.

Le Bureau of the Census des États-Unis a suivi le projet de très près. En fait, des visites entre les deux pays ont eu lieu régulièrement; ces échanges avaient trait aux techniques et aux algorithmes en vue de l'établissement de cartes du recensement à l'aide de moyens automatisés.

7.3 Établissement de cartes statistiques

Statistique Canada dispose de nombreuses données qui présentent souvent des variations régionales importantes. Des cartes statistiques peuvent se révéler utiles afin de faire connaître les tendances géographiques et la répartition de certaines caractéristiques dans l'ensemble du pays. Au fil des ans, une gamme de techniques cartographiques ont été élaborées ou acquises, adaptées, mises à l'essai et mises en application.

8. CONCLUSION

Les exemples susmentionnés illustrent bien la contribution de Statistique Canada, à l'échelle internationale, au titre des méthodes statistiques et des techniques informatiques servant à la production de statistiques officielles. Il est essentiel pour un organisme tel que Statistique Canada de maintenir un programme actif de recherche et d'élaboration afin de continuer à surmonter les nouveaux obstacles et de trouver des solutions de pointe aux problèmes auxquels il se bute et qu'il partage avec les autres intervenants du monde entier. Il ne fait aucun doute que Statistique Canada continuera d'apporter une contribution importante et intéressante dans les domaines de la statistique et de l'informatique.

Le présent exposé n'aurait pas été possible sans le concours de plusieurs personnes. Je tiens à remercier spécialement C. Hill, D. Binder, M. Jeays, E. Dagum, J. Yan et G. Sande d'avoir fourni la documentation sur laquelle est fondé cet exposé.

STATISTICS CANADA LIBRARY
BIBLIOTHÈQUE STATISTIQUE CANADA
1010221863

c. 2