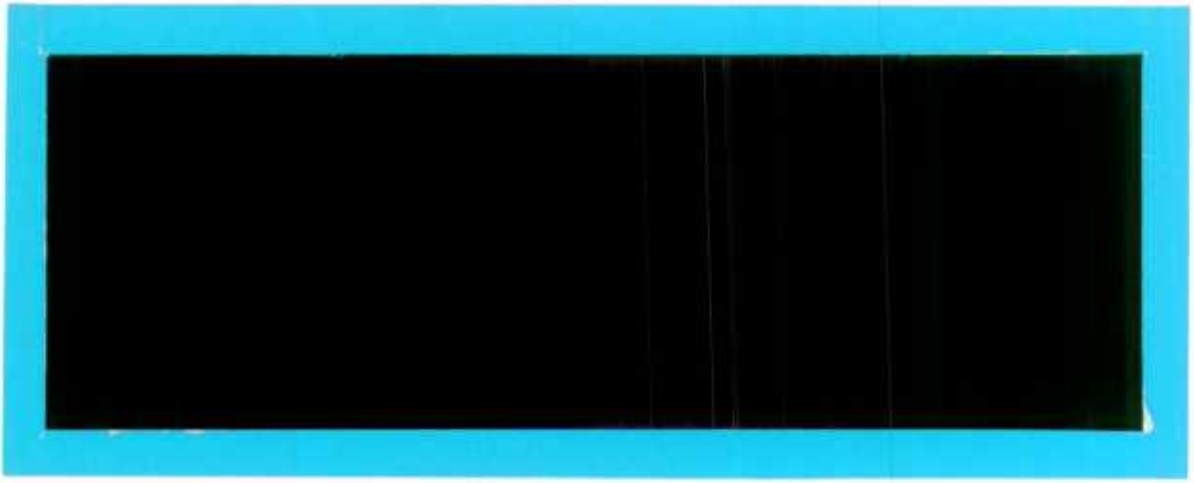

Methodology Branch

Social Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes sociales

Canada

CAHIER DE TRAVAIL NO. SSMD-88-009 E/F

MÉTHODOLOGIE

MÉTHODES DE PROTECTION DU SECRET STATISTIQUE DANS LES RECENSEMENTS DE LA POPULATION ET DE L'AGRICULTURE DU CANADA

SSMD-88-009 E/F

M.J. March et D.A. Norris

Présenté au
Joint Statistical Meetings
17 août 1987
San Francisco

1 INTRODUCTION

Statistique Canada exécute les recensements de la population et de l'agriculture suivant les dispositions de la Loi sur la statistique. Cette loi autorise l'organisme fédéral à recueillir et à diffuser des données de recensement. Toutefois, la diffusion de ces données doit être faite de manière à ne pas divulguer de renseignements sur un répondant en particulier.

À cause de cette obligation de préserver le caractère confidentiel des renseignements, le programme de diffusion des données du recensement prévoit l'élaboration et l'application de méthodes qui garantissent le secret statistique sans que cela n'ait de répercussions sensibles sur la quantité et la qualité des données.

La protection du secret fait partie intégrante de tous les programmes statistiques. Toutefois, les données de recensement soulèvent des problèmes particuliers à cet égard puisqu'elles concernent l'ensemble de la population ou du moins une forte proportion de celle-ci, et qu'elles renferment des données détaillées sur de très petites régions géographiques.

La question du secret statistique a fait l'objet de nombreuses études. Cox et coll. (1985) ont traité le sujet en profondeur dans la perspective du Bureau of the Census des États-Unis; des études non moins complètes ont paru dans Cox (1983) et dans le Report on Statistical Disclosure and Disclosure Avoidance Techniques. Fellegi (1972) et Fellegi et Phillips (1974) ont étudié la question sous une perspective canadienne.

Le présent rapport a pour but de décrire les méthodes de protection du secret utilisées dans les recensements de la population et de l'agriculture du Canada. Les difficultés ne sont pas les mêmes dans les deux cas et les méthodes adoptées pour résoudre ces difficultés diffèrent aussi beaucoup. Dans le cas du recensement de la population, il s'agit surtout de préserver le caractère confidentiel des données portant sur de très petits effectifs tandis que dans le cas du recensement de l'agriculture, qui produit des totalisations de valeurs agrégées (par exemple: dépenses), le problème ne vient pas seulement des petits effectifs mais aussi du danger d'associer des montants à des personnes en particulier. Le présent rapport concerne particulièrement les méthodes utilisées pour les recensements de 1986; toutefois, pour introduire une certaine perspective dans notre analyse, nous comparons les méthodes de 1986 à celles qui ont été utilisées dans des recensements antérieurs et décrivons le point de vue des utilisateurs sur les nouvelles méthodes.

En ce qui concerne le recensement de la population, l'arrondissement aléatoire à un multiple de 5 est la principale méthode utilisée depuis 1971. À cela se sont ajoutées diverses méthodes de suppression de données, qui ont été légèrement modifiées depuis. En ce qui a trait au recensement de l'agriculture, les méthodes de protection de la confidentialité actuellement en usage n'ont réellement été élaborées que vers la fin des années 1970. En 1981, l'application du prototype d'un système généralisé de regroupement et de suppression a donné des résultats acceptables. Comme ces résultats n'étaient pas entièrement satisfaisants, un nouveau système spécialement élaboré pour le recensement de 1986.

II MÉTHODES DE PROTECTION DU SECRET STATISTIQUE UTILISÉES POUR LE RECENSEMENT DE LA POPULATION

Autrefois, les données du recensement de la population étaient publiées sous forme de tableaux spécifiques dans une série de bulletins de recensement. Cette formule de diffusion élémentaire permettait habituellement de résoudre le problème de la confidentialité des données par la suppression manuelle de cases. Comme la quantité de tableaux augmentait et qu'il n'y avait plus assez de temps pour les vérifier manuellement, d'autres méthodes ont dû être adoptées.

En outre, avec l'évolution des techniques informatiques, on a exigé des données plus détaillées et des données régionales enregistrées sur bande magnétique ou microfiche. Pour répondre à ces exigences, il a fallu créer de nouveaux produits du recensement. Parmi ces nouveaux produits, soulignons les séries de données complètes sur les petites régions géographiques, qui peuvent servir d'unités de base pour regrouper et manipuler les données selon les besoins des utilisateurs. Ces séries de données rendent beaucoup plus difficile la protection de la confidentialité. Comme la suppression de cases a une incidence majeure sur les données de ce genre, il faut choisir avec soin les méthodes de protection du secret statistique. Au Canada, le secteur de dénombrement (taille moyenne de la population = 550) sert de région géographique de base pour le recensement. Toutefois, en ce qui concerne les grandes régions urbaines, où est concentrée environ 60% de la population, on utilise aussi comme unité de codage le côté d'îlot, qui sert d'unité d'agrégation pour l'extraction de données. Le côté d'îlot coïncide le plus souvent avec le secteur désigné par le code postal à six caractères; du reste, ce secteur servira aussi d'unité de diffusion pour le recensement de 1986. Le fait de pouvoir totaliser des données par groupe de côtés d'îlot selon les demandes des utilisateurs peut donner lieu à une divulgation de renseignements par différence.

Par surcroît, la diffusion des données du recensement de la population se fait de plus en plus au moyen de totalisations spéciales ainsi désignées du fait que les utilisateurs définissent eux-mêmes le contenu et les unités géographiques de ces totalisations. Il est plus facile aujourd'hui d'obtenir des renseignements de ce genre parce que les données du recensement sont stockées dans une base structurée de manière à permettre l'extraction de totalisations spéciales sur demande. Pour le recensement de 1981, on a produit environ 2,000 totalisations régulières sur support de papier ou sous forme ordinoligne et plus de 10,000 sur demande spéciale.

Les totalisations spéciales constituent une précieuse source de renseignements pour les utilisateurs des données du recensement. Toutefois, le traitement d'une quantité de données aussi considérable exige que la méthode de protection de la confidentialité soit généralisée et facile à appliquer.

Il y a aussi les demandes de plus en plus nombreuses de données sur des groupes cibles bien circonscrits (par exemple sur les groupes ethniques et les professions). La production de totalisations concernant ces groupes, surtout dans le cas de petites régions définies par les utilisateurs, peut donner lieu à une divulgation de renseignements par différence. Cela pourrait se produire, par exemple, durant la construction d'une série de données spéciale sur la

population autochtone. Une forte proportion de la population autochtone du Canada vit dans des réserves qui sont définies comme des subdivisions de recensement dans la classification géographique type. Comme les subdivisions de recensement sont des unités de totalisation de base, il existe déjà beaucoup de données sur ces unités géographiques. Même si les réserves indiennes sont constituées en très grande partie d'autochtones, elles comptent souvent un petit nombre de non-autochtones. Si, dans ces circonstances, on publiait des données portant uniquement sur les autochtones, les non-autochtones pourraient être victimes d'une divulgation de renseignements par différence.

Il existe deux grandes méthodes de protection du secret statistique pour le recensement de la population:

1. l'arrondissement aléatoire;
2. la suppression.

1. ARRONDISSEMENT ALÉATOIRE

L'arrondissement aléatoire est la principale méthode de protection du secret utilisée dans le recensement de la population du Canada depuis 1971. À chaque recensement, toutes les totalisations de données, sauf le total des chiffres de la population et des logements, sont arrondies aléatoirement à un mutiple de 5. Chaque fréquence de case est arrondie au multiple de 5 supérieur ou inférieur à l'aide d'une méthode non biaisée. Plus précisément, on arrondit une fréquence de case au multiple de 5 supérieur ou inférieur selon les modalités suivantes:

Soit r le reste du quotient de f par 5, alors:

1. f est arrondie au multiple de 5 supérieur avec une probabilité r/5, r = 1, 2, 3, 4
2. f est arrondie au multiple de 5 inférieur avec une probabilité 1-r/5, r = 1, 2, 3, 4
3. f n'est pas arrondie si r = 0.

Tous les nombres d'une totalisation, y compris les sous-totaux et les totaux, sont arrondis individuellement. Par conséquent, il y aura rarement une correspondance parfaite entre la somme des parties et le total indiqué.

Nous avons dit plus haut que l'arrondissement aléatoire était en usage depuis 1971. Cette méthode a graduellement été acceptée par les utilisateurs, qui la considèrent comme un moyen acceptable de préserver le caractère confidentiel des informations. Il arrive que de nouveaux utilisateurs se plaignent de ce que les totaux ne correspondent pas à la somme des cases, mais cela n'a pas vraiment été un problème dans les derniers recensements. Le recensement de 1986 sera le quatrième où l'on utilisera cette méthode et les utilisateurs se sont faits à ce léger inconvénient. Certains utilisateurs s'inquiètent de l'effet que peut avoir l'arrondissement aléatoire sur les petites populations, surtout lorsqu'il faut agréger des données arrondies aléatoirement. Pour la plupart des utilisateurs toutefois, cela ne pose pas de problème majeur. De fait, les

données arrondies aléatoirement pour de petites régions géographiques ont maintes fois fait l'objet d'agrégation. Il est en outre possible d'obtenir au besoin des totalisations spéciales lorsque l'arrondissement aléatoire a été effectué après l'agrégation géographique.

Enfin, malgré l'absence de réticence de la part des utilisateurs, il est nécessaire d'approfondir nos recherches pour évaluer l'effet que peut avoir l'agrégation de données arrondies aléatoirement sur la qualité des données.

2. SUPPRESSION

Même si l'arrondissement aléatoire est la principale méthode de protection du secret statistique utilisée dans le recensement de la population, il existe également des méthodes fondées sur la suppression de données.

Il y a deux formes principales de suppression: i) la suppression de régions, où l'on élimine les données se rattachant aux très petites régions, et ii) la suppression de cases, où l'on élimine les petites cases d'un tableau.

Dans le recensement de la population, la suppression de données s'ajoute à l'arrondissement aléatoire quand on craint une divulgation même avec des fréquences arrondies, surtout s'il s'agit de caractéristiques comportant de nombreuses catégories détaillées, comme le revenu, la profession ou l'industrie. En outre, lorsqu'on recueille des données sur certaines caractéristiques à l'aide d'un échantillon de 20%, où le poids attribué à un répondant est habituellement de 5, les cases arrondies de 5 ou 10 reflètent les caractéristiques d'une ou de deux personnes seulement, ce qui peut entraîner une violation du secret statistique.

La suppression de données est fortement souhaitable lorsqu'on produit et diffuse des séries régulières concernant de très petites régions. Le tableau 1 donne la répartition des secteurs de dénombrement (SD) et des subdivisions de recensement (SDR) (c.-à-d. des municipalités) selon leur taille dans le recensement de 1981; les SD et les SDR sont les deux types de petites régions géographiques pour lesquelles il existe des données. Le tableau indique qu'il y a des régions, peu nombreuses toutefois, qui comptent une seule personne ou un petit nombre de personnes, souvent membres d'un seul ménage. Dans ce cas, l'arrondissement aléatoire ne suffit évidemment pas pour préserver le caractère confidentiel des informations.

TABLEAU 1

NOMBRE DE RÉGIONS GÉOGRAPHIQUES ET POPULATION PAR TRANCHE DE POPULATION

Secteurs de dénombrement	Nombre	Population totale - Tous âges
Dont population = 0	2,444	0
Dont population = 1-24	1,096	10,423
Dont population = 25-39	372	11,682
Dont population = 40-49	191	8,379
Dont population = 50-99	936	70,636
Dont population = 100-249	4,154	741,756
Dont population = 250 ou plus	32,004	23,240,615
Subdivisions de recensement		
Dont population = 0	201	0
Dont population = 1-24	124	1,743
Dont population = 25-39	90	2,894
Dont population = 40-49	43	1,869
Dont population = 50-99	197	14,163
Dont population = 100-249	487	82,230
Dont population = 250-499	885	331,746
Dont population = 500-999	1,232	890,494
Dont population = 1000-4999	1,801	3,895,742
Dont population = 5000-9999	321	2,253,482
Dont population = 10000-24999	189	2,821,939
Dont population = 25000-49999	66	2,265,548
Dont population = 50000-99999	44	2,991,868
Dont population = 100000-249999	16	2,170,966
Dont population = 250000-499999	8	2,570,129
Dont population = 500000 et plus	6	3,788,769

Aux fins du recensement de 1981, on a établi des règles pour la suppression des données concernant les très petites régions. Par exemple, la suppression devait être appliquée pour les secteurs d'autodénombrement qui comptaient moins de 50 personnes et les secteurs de recensement par interview qui en comptaient moins de 25. Les données relatives à ces secteurs étaient toutefois comprises dans celles publiées pour des niveaux d'agrégation supérieurs. Les règles variaient légèrement selon que les données étaient destinées à une publication, à une totalisation sommaire, à une bande magnétique ou encore à une totalisation spéciale. Ce manque d'uniformité a créé une certaine confusion parmi les utilisateurs puisque dans quelques cas, les données qui n'étaient pas disponibles sous une forme l'étaient sous une autre. Nous utiliserons de nouveau la suppression de régions pour le recensement de 1986, mais nous appliquerons cette fois une règle uniforme qui prévoira la suppression des données se rattachant à toutes les régions géographiques types qui comptent moins de 40 personnes. En ce qui concerne les régions définies par les utilisateurs, le critère minimum sera de 100 personnes (le nombre est plus élevé pour prévenir

la divulgation de renseignements par différence).

Pour les recensements de 1981 et 1986, le total de la population hors institution a servi à établir le critère minimal pour une totalisation de données recueillies auprès d'un échantillon de 20% puisque seule la population hors institution a fait l'objet de l'échantillonnage. En ce qui concerne la répartition du revenu, aucune donnée n'est publiée pour les régions où la population hors institution est inférieure à 250.

La suppression de régions n'est pas la seule méthode utilisée; en effet, on a appliqué la suppression de cases pour certaines variables, notamment le revenu, la profession et l'industrie. Dans les publications, on supprime les cases relatives à ces trois variables de même que les chiffres obtenus à l'aide de ces cases si la somme de ligne ou de colonne est inférieure à 250. Toutefois, les données pertinentes sont comprises dans les sous-totaux ou les totaux qui figurent dans ces publications.

Une dernière forme de suppression appliquée est la suppression de petites cases individuelles dans le cas des totalisations spéciales. Dans le recensement de 1981, toutes les cases dont l'effectif était inférieur à 25 et qui se rattachaient au revenu, à la profession ou à l'industrie ont été supprimées.

Le seuil a été fixé à 25 pour des raisons ayant trait à la fois à la qualité des données et à l'exploitation puisqu'il n'était pas facile d'appliquer la suppression de régions pour les très petites régions. L'application de ce critère pour les totalisations spéciales a été particulièrement mal reçue chez les utilisateurs, et c'est pourquoi la règle du "25" sera remplacée par la règle des "5" pour le recensement de 1986; cette règle est définie dans le paragraphe suivant.

La demande croissante de totalisations spéciales a remis en question l'efficacité de la méthode de l'arrondissement aléatoire, surtout en ce qui a trait aux petites régions géographiques. Des études ont confirmé que, dans le cas de totalisations fondées sur un échantillon de 20%, une très forte proportion des "5" (après arrondissement) figurant dans les tableaux se rapportaient en réalité à une seule personne. Compte tenu du caractère très détaillé de certaines totalisations spéciales, on craignait que l'arrondissement aléatoire ne suffise pas à protéger le secret statistique. Il a donc été décidé, pour le recensement de 1986, d'éliminer tous les "5" (après arrondissement) des totalisations spéciales fondées sur un échantillon de 20%. Même si cette mesure créait des problèmes d'agrégation, ceux-ci étaient atténués par la possibilité de définir les totalisations spéciales en fonction de nombreux sous-totaux. Il convient de souligner que cette mesure ne visait que les totalisations spéciales; elle ne s'appliquait pas aux publications régulières ni aux totalisations de données régionales sous forme ordinoligne puisque ces dernières renferment très peu de renseignements détaillés et que les publications présentent des totalisations détaillées uniquement au niveau du pays, des provinces et des régions métropolitaines de recensement. La décision d'appliquer cette mesure était, pour ainsi dire, le résultat d'un compromis, et toute la question sera réexaminée pour le recensement de 1991.

III. PROTECTION DU SECRET STATISTIQUE DANS LE RECENSEMENT DE L'AGRICULTURE

Beaucoup de renseignements très précis sont recueillis dans le recensement de l'agriculture. Chaque enregistrement de la base de données renferme plus de 300 zones ou variables.

La plupart des renseignements recueillis (par exemple la valeur de l'actif de l'exploitation agricole, l'état détaillé des dépenses et les ventes totales) sont des données économiques qui ont un caractère relativement confidentiel.

Pour répondre aux besoins des utilisateurs, les données recueillies au moyen du recensement de l'agriculture sont toujours agrégées en fonction de cinq niveaux géographiques: Canada, provinces, régions agricoles (régions relativement étendues qui correspondent habituellement aux districts agricoles), divisions de recensement (régions habituellement moins étendues - -on en compte 266 au Canada) et subdivisions de recensement unifiées (qui sont encore moins étendues mais dont la superficie est d'au moins 25 kilomètres carrés). Ces cinq niveaux d'agrégation représentent environ 20,000 pages de tableaux réguliers.

Statistique Canada considère qu'un agrégat ne respecte pas le principe de la confidentialité des données dans l'un ou l'autre des cas suivants:

1) L'agrégat ne concerne qu'une seule exploitation agricole.

2) L'agrégat concerne plus d'une exploitation mais ce nombre est si faible que les voisins peuvent déduire sans trop de peine des renseignements sur les personnes visées.

3) L'agrégat concerne un nombre suffisant d'exploitations agricoles mais une ou deux de ces exploitations ont tellement d'importance par rapport aux autres que la publication de cet agrégat équivaudrait à fournir des estimations assez précises sur ces exploitations.

4) Même si l'agrégat ne contribue pas directement à la divulgation de renseignements confidentiels, un utilisateur peut déduire des données confidentielles en combinant deux ou plus de deux agrégats qui ne constituent pas en soi de l'information confidentielle. C'est ce qu'on appelle la divulgation <u>par différence</u> (ou par déduction).

Cette définition a été convertie en une série de règles formelles qui stipulent des pourcentages et des nombres d'exploitations précis. Ces chiffres sont confidentiels.

L'expérience et l'étude de distributions de variables ont montré qu'une diffusion générale de données agrégées du recensement de l'agriculture aux niveaux géographiques mentionnés plus haut donnerait lieu à de nombreux cas de violation du secret statistique. À cet égard, on fait les observations suivantes.

1) Beaucoup de subdivisions de recensement unifiées et même certaines divisions de recensement et régions agricoles comptent un très petit nombre d'exploitations agricoles soit parce que le sol est impropre

à la culture, parce que la région est en majeure partie urbaine ou encore parce que les exploitations agricoles sont regroupées graduellement pour former des entités plus importantes. Une proportion appréciable des agrégats produits pour ce genre de régions sont susceptibles de révéler des renseignements confidentiels.

2) Certains produits comme les fruits ou le tabac ne peuvent être cultivés que dans certaines régions du pays; on les trouve donc rarement à l'extérieur de ces régions. D'autres produits sont rares à cause d'une demande restreinte ou parce que leur production exige des conditions ou des installations spéciales. Pour ces produits, il y a risque de divulgation même au niveau provincial.

3) Même si les subdivisions de recensement unifiées comptent un nombre suffisamment élevé d'exploitations agricoles, les agrégats relatifs à une bonne partie des 300 variables ne respectent pas le principe de la confidentialité puisque peu de variables sont réparties uniformément dans le pays.

4) La distribution de la production de certains produits est parfois très asymétrique. Par exemple, il peut y avoir un petit nombre de très gros producteurs de dindes à côté d'un grand nombre de petits producteurs, tout comme un petit nombre d'exploitations agricoles produisent des légumes pour les industries des conserves alimentaires et des aliments congelés tandis qu'un grand nombre de maraîchers produisent les mêmes légumes pour la consommation directe. Pour ces produits, les valeurs déclarées par les gros producteurs sont suffisamment élevées pour constituer la majeure partie de la plupart des agrégats dans ce domaine.

La production des nombreux tableaux de données agrégées du recensement de l'agriculture a été automatisée il y a un certain nombre d'années. Avant 1981, diverses règles de protection du secret statistique ont été utilisées. L'une d'elles (utilisée jusqu'en 1981 inclusivement et appelée "règle des dix exploitations agricoles") consistait à ne publier de données que pour les régions géographiques qui comptaient au moins dix fermes.. Une autre règle consistait à définir certaines variables pour lesquelles il n'y aurait pas de diffusion de données au-dessous d'un niveau géographique donné. Une troisième, appliquée aux distributions de fréquence, consistait à rattacher aux catégories adjacentes les catégories comptant moins de trois exploitations. Avant 1981, on supprimait les cases occupées en majeure partie par un ou deux producteurs très importants, mais cette opération n'était pas systématique. L'arrondissement aléatoire (c.-à-d. l'arrondissement aléatoire du nombre d'exploitations déclarantes à un multiple de cinq et la modification de l'agrégat déduit des déclarations pour conserver la moyenne initiale) a été utilisé pour les tableaux demandés par les utilisateurs et ceux produits à l'aide de la base de données servant au couplage agriculture-population.

Peu de temps avant la diffusion des données du recensement de 1981, un nouveau système généralisé de confidentialité a vu le jour. Ce système pouvait éliminer toute forme de divulgation susceptible de survenir dans le recensement de l'agriculture, à l'aide de règles de confidentialité et d'une méthode de regroupement et de suppression définie par Sande (1977) et Cox et Sande (1979). Cette méthode prévoit la suppression des valeurs contenues

dans les cases complémentaires et les cases de nature délicate afin d'éliminer tout risque de divulgation (par différence). Il s'agit en fait d'un regroupement de cases. Là où il a été appliqué, ce système s'est avéré très efficace. Malheureusement, nous n'avons pas pu appliquer le système à tous les tableaux produits puisque nous nous sommes rendu compte qu'il fallait énormément de ressources humaines et informatiques pour déterminer les cases à supprimer dans des tableaux très détaillés comme ceux du recensement de l'agriculture.

Nous croyons que le système utilisé en 1981 ne convient pas pour le recensement de 1986 pour les raisons suivantes.

1. Il ne fait pas l'objet d'une maintenance régulière et par conséquent, on peut très difficilement l'incorporer à un vaste système de production et, à plus forte raison, l'exploiter à l'intérieur de ce système.

2. Il s'agit d'un "prototype" qui n'est pas conçu pour être utilisé directement avec une large base de données de consultation. Il faudrait élaborer des programmes d'interface qui serviraient à créer des fichiers d'entrée.

3. Il n'est pas conçu pour assurer une liaison directe avec les tableaux prêts pour l'impression et faire les changements nécessaires. Il faudrait donc des programmes d'interface additionnels pour pouvoir utiliser les résultats du système de confidentialité.

4. Son exploitation est coûteuse.

Nous avons aussi rejeté l'arrondissement aléatoire parce que les utilisateurs préfèrent la suppression à la distorsion.

Nous en sommes venus à la conclusion que nous n'aurions pas assez de temps ni assez de ressources pour mettre au point une nouvelle méthode et un nouveau système et, par ailleurs, nous ne pouvions trouver aucun système déjà existant qui pouvait répondre à nos besoins. Nous avons donc décidé de reporter l'élaboration d'un système permanent (réutilisable) en 1991 et d'élaborer pour les besoins du recensement de 1986 un système spécifique simplifié qui résoudrait les principaux problèmes de confidentialité avec l'intervention des spécialistes. Ce système a effectivement été mis au point et servi à produire les tableaux publiés en juin de cette année.

Nous allons décrire brièvement les diverses étapes d'application du système.

La première étape consistait à introduire dans le système un certain nombre de fichiers d'entrée qui avaient été définis par les spécialistes. Un des principaux fichiers contenait les codes géographiques des régions à regrouper afin qu'il n'y ait pas de publication de données pour des régions où, de toute façon, le trop petit nombre d'exploitations agricoles nécessitait la suppression de la plupart des données. Les régions qu'il fallait regrouper on été déterminées à la suite d'une analyse des tableaux de 1981 et au moyen des résultats d'un essai préliminaire du système de 1986.

Venait ensuite l'étape de l'analyse. Nous avons exécuté une seule fois le programme d'analyse et obtenu des agrégats pour les 309 variables du recensement de l'agriculture et pour chaque niveau géographique pour lequel

des données sont publiées. En faisant la somme des valeurs de chaque variable, nous trouvions les valeurs qui formaient la majeure partie de chaque agrégat et une fois la somme connue, nous procédions à des calculs et à des comparaisons pour savoir s'il y avait violation des règles de confidentialité. Les renseignements concernant les cas de violation, notamment le nom de la variable, la taille du complément requis (dans le cas à une seule variable) et le genre de règle qui a été violée, étaient enregistrés dans un fichier d'ordinateur.

La seconde étape, c'est-à-dire le programme de suppression, était appliquée aux tableaux finals produits à l'aide de la base de données d'extraction. Chaque tableau était décrit dans un des fichiers d'entrée introduits antérieurement par terminal à l'aide d'un programme d'interaction. La description contenait les codes correspondant aux variables dans chaque colonne, le mode de répartition géographique utilisé et les codes d'identification des colonnes dont la somme était considérée comme un sous-total. Pour chaque tableau, le système consultait le fichier d'analyse pour déterminer si les variables du tableau présentaient des cas de violation des règles de confidentialité aux niveaux géographiques concernés. Si le système ne relevait aucun cas de violation, il n'était pas nécessaire de supprimer des données du tableau. Si le fichier d'analyse contenait des enregistrements en tout point conformes au tableau, on supprimait automatiquement les cases en question et on choisissait, au besoin, d'autres cases qui devaient être supprimées pour éliminer tout risque de divulgation complémentaire. On choisissait les cases complémentaires à l'aide d'un fichier de compléments géographiques prescrits, qui reliait les régions adjacentes puisque les utilisateurs croyaient préférable de regrouper des régions adjacentes si un regroupement était nécessaire. Le complément géographique adjacent ne servait pas si une autre case dans la région avait été supprimée à cause de la violation d'une règle de confidentialité et l'on pouvait alors utiliser les deux cases qui avaient été supprimées automatiquement comme compléments l'une de l'autre. Si le complément géographique prescrit ne renfermait aucune valeur, un algorithme du système permettait de choisir une autre région géographique à la place, de préférence une région où les autres variables connexes avaient aussi été supprimées. Les compléments additionnels étaient choisis à l'aide d'un ensemble de règles précises si, dans le tableau, une variable était additionnée à d'autres variables connexes (par exemple, vaches laitières, vaches de boucherie et génisses).

Les tableaux ainsi obtenus étaient ensuite soumis à un contrôle visant à vérifier s'il n'y avait plus aucun risque de divulgation par différence et si les compléments choisis étaient suffisamment grands pour que le secret statistique soit respecté. Les rejets à la vérification étaient consignés dans un rapport destiné aux spécialistes; ceux-ci analysaient ce rapport ainsi qu'une version provisoire des tableaux et pouvaient, par un accès en direct, supprimer des cases additionnelles et rétablir celles qui ne devaient plus servir de complément.

Une fois que le spécialiste était satisfait d'un tableau, on créait une version prête pour l'impression.

Le programme d'analyse traitait également les tableaux croisés réguliers en calculant les totaux pour toutes les cases incluses dans ces tableaux et en appliquant en même temps les règles de confidentialité. Toutefois, le

programme de suppression ne permettait pas de traiter ce genre de tableaux. Les spécialistes utilisaient plutôt un rapport produit par le programme d'analyse de même qu'une version imprimée du tableau pour choisir manuellement les cases à caractère confidentiel et les compléments qu'il fallait supprimer. Ils se servaient de l'accès en direct pour modifier les tableaux.

Le principal avantage de ce système était de pouvoir être élaboré et testé à temps pour la production des résultats. En outre, le coût de l'élaboration de ce système et d'exploitation s'est avéré raisonnable et les résultats ont été produits dans les délais requis, malgré qu'il ait fallu l'intervention du personnel des services spécialisés.

Un second avantage du système est d'être bien compris du personnel des services spécialisés et des utilisateurs de données puisque ces deux groupes ont contribué largement à son élaboration.

La principale faiblesse du système est son manque de flexibilité. En effet, il ne peut traiter que les tableaux inclus dans les produits réguliers du recensement. Toutes les variables, y compris les variables calculées et les variables de tableaux croisés, de même que les modes de répartition géographique doivent être définis au préalable. Le programme d'analyse est conçu de manière à être exécuté une fois ou deux tout au plus au début de la production. En outre, il faudrait réviser en profondeur le système si les modes de présentation des tableaux et les variables devaient être modifiés. Le programme de suppression présente aussi un inconvénient en ce sens qu'il a été conçu uniquement en fonction des tableaux produits par le système Statpak de Statistique Canada.

Le manque de flexibilité du système fait qu'il ne peut traiter les demandes de tableaux spéciaux. Nous avons reçu environ 1,000 demandes de ce genre pour les données de 1981 et nous nous attendons à en recevoir autant sinon plus pour les données du recensement de 1986. Les analystes doivent continuer de traiter ces demandes avec les outils qu'ils ont à leur disposition (c.-à-d. l'arrondissement aléatoire, la suppression des cases ne contenant qu'un faible nombre d'exploitations déclarantes et l'analyse visuelle). On songe à utiliser aussi les rapports produits par le système de confidentialité. (Un de ces rapports contenant des données sur chacune des cases à caractère confidentiel qui ont été repérées par le programme d'analyse et l'autre contenant un fichier d'évaluation avec de l'information concernant chaque cas de suppression.)

Nous pouvons peut-être voir une troisième faiblesse dans ce système; elle a trait à la sélection des compléments. L'algorithme de sélection des compléments est conçu de manière à choisir un complément selon une règle arbitraire qui ne tient pas compte de la taille de ce complément. Des contrôles ultérieurs permettent de repérer les compléments trop petits. Toutefois, il n'y a pas d'autre contrôle qui permette de vérifier si l'analyste responsable du tableau a fait un choix juste. Les changements faits par l'analyste doivent être validés manuellement. Cela entraîne automatiquement une possibilité d'erreur.

Nous prévoyons faire une évaluation complète du système de confidentialité du recensement de l'agriculture de 1986. Nous étudierons alors les points

suivants:

a) coûts;
b) qualité des résultats dans l'optique des utilisateurs;
c) capacité du système à préserver le secret statistique;
d) opérations manuelles nécessaires et cohérence des décisions du personnel des services spécialisés.

Au moment de la rédaction de ce rapport, les totalisations produites à l'aide du nouveau système de confidentialité commençaient à peine à être diffusées. Bien qu'il soit trop tôt pour évaluer la réaction des utilisateurs, nous ne croyons pas qu'elle soit négative puisque les méthodes de regroupement appliquées dans ce système s'inspirent de consultations qui ont été menées auprès des utilisateurs. Les analystes affectés à la préparation des tableaux ont exprimé leur appréhension devant le fait que certains tableaux de subdivisions de recensement unifiées ont fait l'objet de nombreuses suppressions et de nombreux regroupements de niveaux géographiques et que cela peut mécontenter les utilisateurs. Nous surveillerons attentivement la réaction des utilisateurs pour voir si les craintes de ces analystes sont justifiées.

IV CONCLUSIONS

Nous venons de décrire les méthodes de protection du secret statistique utilisées dans les recensements de la population et de l'agriculture du Canada.

La principale méthode de protection utilisée dans le recensement de la population est l'arrondissement aléatoire à un multiple de 5. En usage depuis le recensement de 1971, cette méthode est maintenant acceptée par la majorité des utilisateurs. Elle a toutefois l'inconvénient de produire des totaux qui ne correspondent pas à la somme de ligne ou de colonne du fait qu'ils sont arrondis individuellement. L'application de l'arrondissement aléatoire a favorisé la création d'un programme flexible de diffusion de données qui prévoit la production de nombreuses totalisations dont le contenu et les unités géographiques sont définis par les utilisateurs. Les unités géographiques peuvent être formées à l'aide de données recueillies au niveau du côté d'îlot.

Tandis que l'arrondissement aléatoire est la principale méthode utilisée, il arrive aussi qu'on utilise diverses règles qui visent à supprimer les données relatives aux très petites régions ou les petites cases des totalisations concernant certaines variables.

En outre, pour le recensement de 1986, on éliminera toutes les cases de 5 (après arrondissement) des totalisations spéciales fondées sur des données d'échantillon parce qu'il a été démontré que ces cases concernent le plus souvent une seule personne. La méthode dite "de l'arrondissement aléatoire double" pourrait peut-être résoudre le problème des très petites cases (surtout les cases de 5). Cette méthode consisterait à éliminer les "5" des tableaux par un second arrondissement aléatoire qui transformerait le "5" en "15", "10" ou "0". Cette méthode serait préférable à la méthode de la suppression actuellement en usage parce qu'elle serait non biaisée. Toutefois, il n'a pas été possible d'évaluer pleinement les conséquences de l'application de cette méthode pour le recensement de 1986 ni d'adapter cette méthode au

logiciel d'extraction actuellement en usage.

Même si beaucoup de recherches peuvent encore être faites en matière de protection de la confidentialité dans le recensement de la population, il est peu probable qu'on trouve une méthode qui réponde à toutes les exigences à la fois. Le choix des règles à appliquer procédera toujours du souci de trouver un juste milieu entre les risques de divulgation et un niveau de qualité qui ne rende pas les données inutilisables pour la recherche et l'élaboration de programmes.

La protection de la confidentialité est une des questions qui feront l'objet d'une étude approfondie dans les séances preparatoires des recensements de 1991. Bien que les utilisateurs n'aient pas semblé trop mal accueillir le fait que la méthode de l'arrondissement aléatoire produit des totaux qui ne correspondent pas à la somme de ligne ou de colonne, on étudiera avec soin la possibilité d'appliquer une méthode d'arrondissement aléatoire contrôlé. On suivra de près l'application de l'arrondissement contrôlé dans le recensement de 1990 aux États-Unis.

On envisagera également d'autres moyens que les méthodes de suppression pour traiter les petites cases, surtout en ce qui concerne les très petites régions géographiques.

Avant de décider de l'élaboration d'un nouveau système de confidentialité du recensement de l'agriculture pour 1991, il faudra évaluer entièrement le système actuel au point de vue du coût, de l'utilité, de la satisfaction des utilisateurs et de la qualité des données (compte tenu de distributions variables). Pour des raisons de qualité, on pourrait devoir supprimer les données publiées au niveau géographique le plus bas (subdivision de recensement unifiée). Il faut évaluer la possibilité de créer un système généralisé de confidentialité qui pourrait répondre aux besoins du recensement de l'agriculture et être facilement adapté aux systèmes d'extraction qui seront utilisés dans l'avenir. Les demandes des utilisateurs et la souplesse qu'elles exigent sont des questions qui méritent aussi une attention particulière.

BIBLIOGRAPHIE

1. COX, LAWRENCE H. (1983): "Some Mathematical Problems Arising from Confidentiality Concerns", Statistical Review, 21,5, Statistics Sweden, Stockholm, 179-189.

2. COX, L.H., JOHNSON, B., MCDONALD, S.-K., NELSON, D. et VASQUEZ, V. (1985): "Confidentiality Issues at the Census Bureau", Proceedings of First Annual Census Bureau Research Conference, Reston (VA.)

3. COX, L. et ERNST, L. (1982): "Controlled Rounding". INFOR, 20, 4, pp. 423-432.

4. COX, L. et SANDE, G. (1979): "Techniques for Preserving Statistical Confidentiality.", Institut international de statistique, vol. 3, Actes du 42e congrès, Manille.

5. FELLEGI, Ivan P. (1972): "On the Question of Statistical Confidentiality", J.A.S.A., mars 1979, vol. 67, no. 337, pp. 7-18.

6. FELLEGI, Ivan P. (1975): "Controlled Random Rounding", Survey Methodology, 1, Statistique Canada, p. 123-135.

7. FELLEGI, I et PHILLIPS, J. (1976): "Statistical Confidentiality: Some Theory and Applications to Data Dissemination". Annals of Economic and Social Measurement. pp. 399-409.

8. NARGUNDAR, M.S. et SAVELAND, W. (1972): "Random Rounding: A Means of Preventing Disclosure about Individual Respondents in Aggregate Data". American Statistical Association, Proceedings of the Social Statistics Section, Washington, D.C. pp. 382-385.

9. SANDE, G. (1977): "Towards Automated Disclosure Analysis for Enterprise-Based Statistics", Statistique Canada, document non publié.

10. U.S. DEPARTMENT OF COMMERCE (1978): "Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure Avoidance Techniques". U.S. Government Printing Office, Washington, D.C.