



**Conseil Économique
et Social**

Distr.
GÉNÉRALE

CES/2001/18
28 mars 2001

FRANCAIS
Original : FRANÇAIS et ANGLAIS

COMMISSION DE STATISTIQUE et COMMISSION ÉCONOMIQUE POUR L'EUROPE

CONFÉRENCE DES STATISTICIENS EUROPÉENS

Quarante-neuvième réunion plénière
(Genève, 11-13 juin 2001)

Stratégies et approches relatives aux statistiques régionales

Communication sollicitée auprès de Statistique Canada

I. Introduction

1. La plupart des bureaux statistiques nationaux (BSN) ont essentiellement pour mandat de suivre les conditions sociales, économiques et environnementales au niveau national, ainsi qu'à l'échelon des principales unités administratives (provinces, États, régions métropolitaines importantes) à l'intérieur du pays. Toutefois, la demande de données à des niveaux géographiques plus restreints est toujours présente, particulièrement chez les administrations locales et les entreprises qui doivent prendre des décisions en matière d'investissement, de marketing et d'implantation, décisions qui nécessitent une connaissance des régions locales. Nous utiliserons le terme « statistiques régionales » pour parler des statistiques s'appliquant à des régions plus restreintes qu'un État, une province ou une région métropolitaine importante, soit toute une gamme de régions qui va des villes importantes aux villages ruraux, en passant par les quartiers urbains.

2. La portée des responsabilités des BSN en ce qui a trait aux statistiques régionales dépend de la répartition des responsabilités gouvernementales au sein d'un pays. Par exemple, dans certains pays, les administrations locales sont créées par les provinces et la responsabilité relative à leurs besoins statistiques sont du ressort des administrations provinciales. Toutefois, dans de

nombreux pays, quelle que soit la répartition officielle des pouvoirs, on s'attend dans les faits à ce que le BSN réponde aux besoins de statistiques régionales, à partir de ses propres ressources, ou en collaboration avec d'autres niveaux de gouvernement. Le BSN doit, à tout le moins, établir les normes et le cadre s'appliquant aux données régionales, afin que celles-ci ne deviennent pas un enchevêtrement de statistiques inégales et non comparables au niveau national.

3. Du fait des budgets limités, le BSN est aux prises avec un compromis difficile entre des investissements à l'égard de statistiques nationales et la production de données régionales détaillées. Doit-il choisir de couvrir davantage de domaines ou de couvrir les domaines existants de façon plus détaillée, aux niveaux national et provincial, ou encore de fournir davantage de données régionales détaillées pour des sujets qui sont déjà couverts au niveau national. Il n'existe pas de formule établie pour résoudre ce problème. L'équilibre obtenu dépend pour une large part des besoins nationaux, des pouvoirs relatifs et de la tradition, et peut-être de certaines considérations statistiques en parallèle. Néanmoins, il existe une série de mesures et d'approches qu'un BSN doit envisager pour répondre dans la plus large mesure possible aux demandes de statistiques régionales, à partir d'un budget limité.

4. Il existe quatre sources possibles de données statistiques régionales qui, individuellement ou collectivement, sont à la base de la majeure partie de la production des données régionales par les organismes statistiques. Les recensement ou les dénombrements complets de la population constituent une source traditionnelle. Les dossiers administratifs, y compris les registres nationaux qui couvrent la totalité ou la presque totalité d'une population définie, sont à de nombreux égards équivalents à un recensement. Les enquêtes nationales sur échantillon sont rarement suffisamment exhaustives pour produire des données régionales directement, mais elles constituent une source valable d'information à jour qui peut être utilisée, sur la base de certaines hypothèses et en combinaison avec d'autres sources, pour produire des données régionales. Enfin, des études locales axées sur des régions particulières permettront de produire des données régionales, mais elles ne peuvent s'appliquer à un ensemble complet de régions. L'imagerie par satellite et la photographie aérienne peuvent être perçues comme des recensements ou des études au niveau local selon leur couverture.

5. Dans le présent document, nous passons d'abord en revue le rôle important que joue le recensement de la population, avec ou sans registre de la population, quant à la production de données socio-économiques régionales (Section 2), puis nous mettons l'accent sur le rôle fondamental d'une infrastructure géographique à jour pour appuyer la production des statistiques régionales, y compris particulièrement les recensements de la population (Section 3). Nous examinons ensuite des approches pour la production de données régionales concernant les personnes et les familles, entre les recensements (Section 4), les activités des entreprises (Section 5) et les questions environnementales (Section 6). Nous concluons par quelques observations générales au sujet de la diffusion des statistiques régionales et de la gestion de ces statistiques à l'intérieur d'un BSN.

II. Recensement de la population

6. Dans la plupart des pays, le recensement de la population joue un rôle de premier plan quant à la production de données régionales sur les personnes, les familles et les ménages. Les estimations qu'il permet de produire, qui sont fondées sur un dénombrement complet de la population (à tout le moins pour les caractéristiques de base), sont exemptes d'erreurs d'échantillonnage, ces dernières limitant la capacité de production d'estimations régionales à partir des enquêtes sur échantillon. Si les ménages sont codés selon un niveau géographique détaillé (p. ex., un îlot ou un côté d'îlot), les totalisations directes des ménages permettent de produire des agrégations statistiques pour toute région géographique qui peut être définie ou faire l'objet d'une approximation, selon le niveau de codage géographique le plus détaillé.

7. Toutefois, les recensements comportent des lacunes. Ils sont coûteux et, de ce fait, ne sont pas tenus fréquemment. Les données du dernier recensement peuvent fournir une mauvaise représentation des régions qui connaissent des progrès rapides. Dans nombre de pays, on a recours à l'échantillonnage pour nombre des questions du recensement. Même si cela amène des erreurs d'échantillonnage dans les estimations, ces échantillons sont encore considérables comparativement à ceux d'une enquête type sur échantillon. Par ailleurs, les échantillons sont généralement répartis dans tous les secteurs de dénombrement du pays, ce qui permet de produire des estimations régionales, même si les régions en question seront parfois plus importantes que dans le cadre d'un véritable recensement.

8. En ce qui a trait à la précision, les erreurs non dues à l'échantillonnage, par exemple, les erreurs de couverture et le biais dans les réponses, sont peut-être plus graves. La plupart des recensements laissent de côté certaines personnes, ou dénombrent deux fois la même personne, et il a été démontré à maintes reprises que les personnes non dénombrées ne sont généralement pas représentatives de la population dans son ensemble. Il se peut par conséquent que les estimations de recensement soient biaisées par rapport à certains sous-groupes de population. Si ces sous-groupes (p. ex., certains groupes d'immigrants) ont tendance à être concentrés géographiquement, cela peut avoir des répercussions graves sur les estimations pour certaines régions. Un biais dans les réponses se produit lorsqu'une question du recensement est systématiquement mal comprise par de nombreux répondants. Ces erreurs touchent à la fois les estimations pour les petites et les grandes régions.

9. Les pays qui maintiennent un registre de la population ont la possibilité de produire des données s'apparentant aux données de recensement pour des petites régions, plus fréquemment que dans le cadre des cycles traditionnels de 5 à 10 ans des recensements. La tenue à jour des données sur le lieu de résidence constitue clairement une exigence pour la production de données régionales précises à partir de ces registres. Il se peut que la portée des données disponibles à partir d'un système de registre soit moins grande que celle des données découlant d'un recensement conventionnel, étant donné que le premier se limite aux caractéristiques figurant dans des registres administratifs qu'il est possible de coupler. Dans certains pays, le registre de la population peut servir de base à un recensement qui sert à recueillir les caractéristiques

additionnelles nécessaires qui ne sont pas présentes dans les registres existants. Redfern (1989) donne une description utile des pratiques utilisées en Europe à cet égard.

10. Étant donné que le recensement permet la production d'estimations pour de très petites régions, des règles visant à empêcher la divulgation directe ou par recoupement des données individuelles doivent être mises en place. Il peut s'agir notamment d'imposer un minimum quant à la population des régions au sujet desquelles des données seront diffusées, de procéder à la perturbation aléatoire des données, de supprimer des données ou d'avoir recours à d'autres techniques (Jabine (1993), Zayatz et coll. (2000)).

11. Même si un recensement, avec ou sans registre de la population, constitue une source précise de données régionales en date du recensement, la valeur de ces données diminue au fur et à mesure que le temps passe. Toutefois, le rôle des données du recensement à l'égard de la production des statistiques régionales dépasse de beaucoup l'utilisation directe des résultats de chaque dénombrement périodique. Au cours des périodes intercensitaires, les données de recensement peuvent être utilisées comme repères, comme base d'échantillonnage ou comme données auxiliaires, à l'égard d'autres sources de données qui sont disponibles. Ces utilisations sont décrites de façon plus détaillée à la Section 4. Une solution de rechange innovatrice au recensement traditionnel est décrite à la Section 4.4.

III. Infrastructure géographique

12. Pour qu'un recensement national permette la production de données régionales précises, il faut disposer au préalable d'une infrastructure géographique des limites et d'une fonction de cartographie pour l'ensemble du pays. Une telle infrastructure nécessite que chaque logement soit lié à un emplacement géographique précis sur le terrain, le degré de précision étant déterminant du niveau de définition des régions. Même si la technologie moderne de positionnement global fait en sorte qu'il est possible de repérer chaque logement selon une paire précise de coordonnées, il suffit généralement, à des fins statistiques, de lier chaque logement d'une région urbaine à un côté d'îlot (c.-à-d. un côté d'une rue entre deux intersections), ou à un immeuble, dans le cas des immeubles en hauteur. Dans les régions rurales, le niveau de précision choisi dépend des limites administratives et naturelles locales, l'utilisation de coordonnées précises pour chaque logement permettant d'assurer le maximum de souplesse.

13. L'infrastructure géographique, qui est nécessaire dans le cadre du recensement, est aussi utile pour la production de statistiques régionales à partir d'autres sources. Essentiellement, chaque point de données, quelle qu'en soit la source, doit être lié à un emplacement géographique, selon un niveau de détail suffisant pour permettre l'agrégation en régions présentant un intérêt statistique. Par exemple, si la source des données est un registre administratif, ou un registre des entreprises, l'adresse de chaque enregistrement doit pouvoir être transposée en une paire de coordonnées géographiques, ou à tout le moins correspondre à la région dans laquelle se trouve cette adresse. Étant donné que les registres administratifs utilisent souvent les adresses postales,

un fichier qui convertit les codes postaux en emplacements géographiques constitue un outil valable pour l'élaboration de données régionales.

14. Une infrastructure géographique précise et à jour, qu'elle soit établie par le BSN ou obtenue au niveau externe, est essentielle pour que l'on puisse choisir les régions pour lesquelles des statistiques seront produites, dans le cadre d'un programme de statistiques régionales.

IV. Statistiques régionales sur les personnes et les ménages - périodes intercensitaires

15. Nous aborderons maintenant la question de la production de données régionales pour les personnes ou les ménages, au cours des périodes qui séparent les recensements. Il ressort clairement que l'existence d'un registre à jour de la population fait toute la différence en ce qui a trait aux possibilités et à la façon de faire. Étant donné que d'autres pays ont beaucoup plus d'expérience que le Canada dans ce genre de situation, nous nous limiterons aux cas où il n'existe pas de registre de la population régulièrement mis à jour.

16. Dans de tels cas, trois catégories principales d'approches s'offrent. La première consiste à utiliser des fichiers s'apparentant à ceux du recensement qui proviennent de systèmes administratifs et qui visent à couvrir l'ensemble d'une population bien définie. La deuxième consiste à exploiter les données des enquêtes sur échantillon et, grâce à des hypothèses modèles additionnelles, à produire des estimations pour des régions plus petites (quoique pas encore très petites) que celles visées par l'estimation directe découlant d'enquêtes. La troisième catégorie représente une combinaison de ces deux approches et utilise les données du recensement le plus récent. Dans les paragraphes qui suivent, nous passerons en revue certaines des caractéristiques de ces approches.

a) Fichiers administratifs

17. Parmi les fichiers administratifs qui comportent des possibilités statistiques au niveau régional figure le fichier annuel des déclarations de revenu des particuliers¹. Si chaque enregistrement comprend une adresse résidentielle qui peut être liée à un point géographique ou à une région, des données peuvent être totalisées directement pour les régions, sous réserve des exigences relatives à la confidentialité (comme c'est le cas pour les données du recensement). Les caractéristiques disponibles ne dépassent généralement pas les variables démographiques et les variables de revenu, et la couverture se limite aux déclarants. Néanmoins, un tel fichier représente une source importante de données annuelles pour des régions relativement petites. La couverture de la population peut être améliorée grâce à l'imputation des personnes à charge pour lesquelles des déductions sont demandées dans la déclaration de revenu. Au Canada, la

¹ Parmi les autres exemples, pour des populations plus restreintes, figurent les fichiers des titulaires de permis de conduire, des prestataires de l'assurance-emploi ou des bénéficiaires de l'assurance-maladie.

couverture de ces fichiers imputés s'apparente à celle du recensement, étant donné que la couverture des personnes à faible revenu qui doivent produire des déclarations de revenu pour obtenir des prestations d'aide sociale augmente.

18. En ce qui a trait aux données administratives en général, le statisticien doit tirer parti de ce qui est disponible (même s'il peut influencer sur le contenu à plus long terme), tenir compte des écarts entre les concepts, la définition ou la couverture des fichiers administratifs et les objectifs statistiques, et évaluer les problèmes de déclaration ou de codage des enregistrements. Sous réserve de ces précautions, les données administratives peuvent constituer une source géographique valable de données régionales (Brackstone, 1987).

b) Données des enquêtes sur échantillon

19. Le problème que présentent les données des enquêtes sur échantillon comme sources de statistiques régionales a trait à la taille de l'échantillon. Il arrive fréquemment que le nombre de cas compris dans l'échantillon ne soit pas suffisant dans une région pour permettre des estimations directes fiables. Parfois même, il est impossible de produire des estimations. Dans le cadre des enquêtes sur échantillon nationales importantes, on peut concevoir des stratégies d'échantillonnage permettant d'assurer un niveau acceptable de précision pour des régions définies, par exemple, les régions infra-provinciales, sans diminuer de façon significative la fiabilité des estimations à des niveaux plus élevés (Singh et coll., 1994). Toutefois, il n'est pas possible de produire des estimations fiables pour des régions plus petites, ou pour des régions de taille similaire qui n'ont pas été prises en compte au moment du plan d'échantillonnage. Les échantillons plus grands sont utiles et peuvent permettre une estimation directe pour certaines des grandes régions, mais les budgets limitent généralement cette approche en tant que solution générale. Si aucune autre source de données n'est disponible, les statisticiens doivent se fier uniquement aux méthodes fondées sur des modèles, lesquelles nécessitent que l'on fasse des hypothèses au sujet des liens qui existent entre les données régionales et les autres données. Ces méthodes reposent souvent sur le transfert d'information, c'est-à-dire qu'elles consistent à emprunter de l'information d'autres secteurs de l'enquête sur échantillon pour augmenter le nombre d'unités qui contribuent aux estimations pour une région donnée. Le transfert peut se faire à partir d'autres périodes, d'unités d'échantillonnage extérieures à la région donnée, ou d'autres variables mesurées à partir de la même unité d'échantillonnage. Voici des exemples de ces cas. La plupart d'entre eux permettront d'élargir la gamme des estimations régionales qui peuvent être produites à partir des enquêtes comportant des échantillons relativement grands. Ils ne permettent toutefois pas de convertir de façon magique les enquêtes reposant sur de petits échantillons en sources importantes de données régionales.

- i) Dans le cadre d'une enquête mensuelle, il est possible de combiner des données pour une région, pendant plusieurs mois consécutifs, en vue de produire des estimations directes d'une moyenne mobile sur plusieurs mois pour la région. Cela peut par exemple

permettre la production d'estimations trimestrielles, alors que des estimations mensuelles ne sont pas possibles.

ii) On doit être prêt à faire l'hypothèse que les moyennes ou proportions estimées pour une région plus grande s'appliquent également à une région plus petite qu'elle englobe. Si la taille de la région est connue, on peut obtenir une estimation en multipliant par la moyenne ou la proportion supposée. Cette hypothèse est plus réaliste pour des sous-groupes de population (p. ex., des groupes d'âge), plutôt que pour la population dans son ensemble. Dans ce cas, si la taille de chaque sous-groupe est connue pour la région, un estimateur synthétique peut être établi grâce à la multiplication des tailles et des moyennes supposées et à l'agrégation des résultats.

iii) Si des variables connexes additionnelles sont disponibles à partir de l'enquête, des modèles plus élaborés peuvent être établis et comporter un lien entre la variable faisant l'objet de l'estimation et ces variables auxiliaires. Les paramètres du modèle peuvent être estimés à un niveau géographique plus élevé, lorsque l'échantillon est suffisant pour procéder à cette estimation de façon fiable. Le modèle est par la suite appliqué avec les paramètres estimés aux données pour la région déterminée.

20. Pour toutes ces approches, on manque de données de base fiables pour chaque région. Si de telles données étaient disponibles, par exemple, à partir d'un recensement récent ou de dossiers administratifs, elles pourraient être utilisées en combinaison avec l'une ou l'autre des sources pour produire des estimations plus fiables que chacune des sources isolément.

c) Sources combinées

21. Les méthodes qui combinent des données de recensement ou des données administratives récentes et des données d'enquête sur échantillon courantes profitent du transfert de données extérieures à l'enquête. Elles nécessitent aussi des hypothèses modèles, mais ces dernières sont souvent plus faibles (étant donné qu'elles reposent sur des hypothèses quant au changement par rapport au repère, plutôt qu'au sujet de niveaux absolus pour chaque région) et sont donc davantage acceptables, ou plus plausibles, que les données d'enquêtes sur échantillon prises isolément.

22. Une gamme variée de méthodes d'estimation (que nous ne décrivons pas ici) a été élaborée pour traiter cette situation. Certaines de ces méthodes peuvent être perçues comme une estimation des changements par rapport au repère le plus récent, d'autres comme une répartition d'estimations d'enquête sur échantillon à jour et fiable entre les composantes, à partir des données repères, et d'autres encore, comme une recalibration des anciens chiffres repères en fonction des nouvelles estimations courantes. Essentiellement, toutes ces méthodes nécessitent qu'un équilibre soit établi entre ces trois types d'estimations : a) variance élevée, mais

estimations d'enquête directe courante non biaisées pour la région en question; b) faible variance pour les estimations d'enquête courante à l'égard d'une région voisine plus grande comprenant la région en question; c) estimations de type recensement pour la même région à partir des données administratives récentes, ou d'un recensement passé, qui peuvent contenir un biais indéterminé, en raison de la source et du délai écoulé. Toutes les données auxiliaires disponibles peuvent être intégrées pour améliorer la précision de chaque estimation de composante. La façon de combiner ces trois types d'estimations est déterminée par le choix et les paramètres du modèle.

23. En résumé, les méthodes énoncées dans la présente section et dans la section précédente ont essentiellement pour effet de réduire la variance, du fait qu'elles utilisent davantage de données, mais elles peuvent entraîner un biais en raison de l'utilisation d'hypothèses modèles qui ne seront jamais exactement correctes. Il est très important d'analyser le rendement de ces méthodes avant de les utiliser, par exemple, en procédant au processus d'estimation au cours d'une année de recensement, lorsque des estimations directes sont disponibles aux fins de la comparaison, et de façon périodique par la suite. Pour des descriptions plus détaillées des méthodes disponibles dans cette catégorie, voir, par exemple, Schaible (1979), Purcell et Kish (1979), Fay et Herriott (1979), Ghosh et Rao (1994), Singh, Gambino et Mantel (1994), Rao (1999), Gambino et Dick (2000).

d) Cumul des recensements

24. Une solution de remplacement innovatrice au recensement est envisagée dans au moins deux pays. La méthode de production de données régionales à partir d'un échantillon cumulatif important est préconisée depuis longtemps par Leslie Kish, comme solution de remplacement au recensement traditionnel (Kish, 1990, 1998). On procède au cumul des résultats de l'enquête sur échantillon, c'est-à-dire qu'au cours d'une période prolongée (par exemple, une décennie), chacune des régions plus petites pour laquelle des estimations sont nécessaires est incluse une fois dans l'échantillon, ce qui permet la production d'une estimation directe pour cette région au moins une fois pour chaque période. Des régions de plus en plus grande (agrégation des régions plus petites) sont représentées plus souvent dans l'échantillon, ce qui permet des estimations plus fiables ou plus fréquentes pour ces régions. Dans le cas des régions encore plus grandes, par exemple, des provinces ou l'ensemble du pays, l'échantillon cumulatif est suffisant pour fournir des estimations annuelles fiables, ou plus fréquentes, selon certains niveaux de détail. L'approche peut être envisagée, parallèlement à des recensements périodiques, ou en l'absence de tels recensements, pour recueillir des données de base et calibrer les estimations des enquêtes intercensitaires.

25. Le cumul des recensements permet d'éviter l'élaboration de modèles, mais il repose sur le principe que des estimations non biaisées de moyennes sur plusieurs années, ou des estimations asynchrones pour diverses régions du pays, constituent des options de rechange satisfaisantes aux estimations simultanées à un point donné dans le temps des recensements traditionnels. Le coût relatif constitue aussi un facteur clé, particulièrement dans les cas où un recensement de base est

aussi effectué. Par ailleurs, cette approche tient compte du fait que les estimations de recensement peuvent dater de 12 ans lorsque les suivantes sont produites, à partir d'estimations annuelles fiables pour de nombreuses régions plus grandes, le contenu de ces estimations s'apparentant à celui des recensements du point de vue du niveau de détail. Elle fait aussi suite aux préoccupations croissantes concernant les difficultés et les coûts de plus en plus grands liés à la tenue d'un recensement traditionnel.

26. Cette approche fait l'objet d'un essai aux États-Unis, sous le nom d'American Community Survey (Alexander, 1999), et en France, sous le nom de Recensement continu de la population (Isnard, 1999).

V. Statistiques sur les entreprises

27. Les problèmes de production de données régionales pour les entreprises diffèrent de nombreux égards de ceux relatifs aux données sur les personnes ou les ménages.

28. Même si l'association d'une personne avec un « lieu habituel de résidence » représente, pour la grande majorité de la population, un concept² relativement clair et non ambigu, la question de la répartition géographique des diverses caractéristiques des entreprises est moins évidente dans nombre de cas. Pour ce qui est des entreprises qui ne comptent qu'un établissement, et dont toutes les activités se déroulent à un seul emplacement, il n'y a pas de problème conceptuel, même s'il peut y avoir un problème pratique lorsque la source des données est un fichier administratif qui comporte, par exemple, l'adresse d'un comptable, plutôt que l'adresse de l'entreprise. Dans le cas de certaines variables, comme l'emploi, aucun problème conceptuel majeur ne se pose, même pour les grandes entreprises (sauf peut-être pour celles du secteur du transport, ou pour certaines industries de service). Toutefois, dans le cas des variables comme le revenu et les bénéfices, des questions réelles peuvent se poser quant à leur répartition géographique pour les entreprises comptant plusieurs établissements. Plus la région géographique est grande, moins le problème est grave, l'emplacement dans une province ne faisant pas de différence lorsque l'on s'intéresse aux totaux provinciaux seulement. En général, toutefois, les règles de répartition géographique doivent être déterminées avant d'envisager la production d'estimations régionales pour les activités des entreprises, et dans le cas de certains aspects de ces activités, les estimations régionales n'ont pas de sens au niveau conceptuel.

29. Même si pour les enquêtes auprès des ménages, le principal obstacle quant à la production d'estimations régionales est la taille de l'échantillon, pour les enquêtes auprès des entreprises, ce sont les considérations de confidentialité qui constituent généralement l'obstacle majeur. Plus la

²Concept qui est peut-être moins bien défini avec l'augmentation du nombre de résidences secondaires, de l'incidence des absences prolongées pour des destinations plus au sud, et des conditions de logement plus souples.

région est petite, plus il y a de chances qu'une ou quelques entreprises majeures dominent un secteur d'activité particulier, ce qui empêche la production d'estimations pour cette région en raison des risques de divulgation. Les méthodes de vérification des produits statistiques relatifs aux entreprises qui permettent de déterminer les risques possibles du point de vue de la divulgation sont assez bien développées (FCSM, 1994), mais elles nécessitent une attention constante de la part des BSN. Le problème de la confidentialité est moins important pour les branches d'activité caractérisées par des petites unités, et il peut s'agir des mêmes branches d'activité pour lesquelles les problèmes conceptuels abordés au paragraphe précédent ne sont pas aussi graves. Pour ces branches d'activité, les considérations relatives à la taille de l'échantillon peuvent en fait constituer le facteur restrictif, auquel cas on peut avoir recours aux familles de méthodes décrites dans la section précédente.

30. Il existe un troisième domaine de divergence par rapport aux données sur les personnes, à tout le moins pour les pays qui ne tiennent pas de registre de la population, à savoir l'existence de nomenclatures relativement à jour des entreprises. Cela fournit non seulement une base d'échantillonnage et une source de données auxiliaires pour l'estimation, mais aussi une source possible d'estimations directes de la démographie des entreprises, à tout le moins sur une base annuelle. Dans nombre de pays, le registre des entreprises est mis à jour grâce aux opérations transmises par le système d'imposition des entreprises, lequel constitue en soi une source annuelle de données administratives s'apparentant à un recensement sur les activités des entreprises. Toutefois, l'utilisation des données fiscales nécessite encore un examen soigneux des questions conceptuelles et géographiques ainsi que des questions de confidentialité soulevées ci-dessus.

VI. Statistiques environnementales

31. Les statistiques sur l'environnement présentent des défis encore différents pour la production de statistiques régionales. Même si certaines questions environnementales sont de portée nationale ou même mondiale, nombre sont de nature locale. De nombreuses sources de pollution sont généralement localisées, et leurs répercussions se font davantage sentir dans le voisinage d'une usine ou d'un accident écologique. Les répercussions socio-économiques des problèmes environnementaux plus larges (p. ex., la perte de stocks de poissons) se font fréquemment sentir dans de petites collectivités, souvent isolées, dont l'industrie est axée sur les ressources.

32. Certaines données environnementales sont recueillies auprès des ménages ou des particuliers (p. ex., sur les pratiques de recyclage, la consommation de carburant) et leurs possibilités en tant que source de données régionales sont sujettes à des considérations déjà soulevées à la Section 4. D'autres données environnementales (p. ex., sur la production de déchets, les dépenses pour la protection de l'environnement, l'utilisation des ressources naturelles) proviennent des entreprises et sont assujetties aux considérations énoncées à la Section 5. Toutefois, une part importante des données environnementales sont obtenues à partir d'enquêtes sur le terrain (p. ex., géologiques,

physiographiques et hydrographiques), de mesures au moyen d'instruments (p. ex., température, qualité de l'air, qualité de l'eau, épaisseur de la couche d'ozone), et d'observations directes (p. ex., utilisation du territoire). Des considérations différentes régissent les rapports qui existent entre ces sources de données et les données régionales.

33. Du fait que les données environnementales ne respectent pas les limites administratives, la nécessité d'une infrastructure géographique souple, qui est soulignée à la Section 3, revêt une importance particulière dans ce cas. La détermination géographique des limites régionales est nécessaire pour regrouper les données selon des unités géographiques qui se prêtent davantage à l'analyse environnementale. Par exemple, la production de déchets attribuable à un certain type d'activité agricole peut être agrégée pour l'ensemble des producteurs avoisinants un bassin fluvial. Les unités géographiques environnementales sont, soit définies au préalable (écozones, bassins de drainage), soit dictées par des événements spéciaux (des régions couvertes de différentes épaisseurs de glace, terres inondées par des pluies abondantes ou par la fonte des neiges au printemps). Dans certains cas, la région étudiée peut se limiter à un emplacement très petit, par exemple, un parc.

34. Des données sur la quantité ou la qualité matérielle peuvent être difficiles à agréger ou à résumer. Dans certains cas, des données de source ponctuelle, comme les mesures de la qualité de l'air, ne peuvent être considérées comme représentatives d'unités géographiques plus importantes. La qualité de l'eau peut faire l'objet de résumés ou de comparaisons, grâce à un indicateur, par exemple, le nombre de jours d'ouverture des plages pour la baignade, mais ne peut pas constituer simplement une agrégation ou une moyenne des données sur la qualité. Pour nombre de mesures, l'accent est mis sur les changements qui se produisent au fil du temps, plutôt que sur les comparaisons régionales. Dans d'autres cas, des techniques d'échantillonnage et d'estimation peuvent devoir reposer sur des techniques d'analyse spatiale, comme l'établissement de cartes en courbes de niveau ou l'interpolation.

35. Les préoccupations en matière de protection des renseignements personnels et de confidentialité liées aux données environnementales dépendent de la source de ces données. Les données recueillies auprès des ménages ou des entreprises, même si elles comportent des mesures matérielles, sont protégées par les mêmes règles de confidentialité que les autres données provenant de ces sources. Des mesures directes du stock de ressources naturelles ou de la qualité de l'environnement ne soulèvent pas de telles préoccupations. La représentation cartographique des modèles spatiaux peut constituer une façon de surmonter certaines des frustrations analytiques découlant de la suppression de données régionales. Les cartes choroplèthes (qui illustrent la répartition des variables ou des caractéristiques grâce à des zones colorées ou ombrées correspondant aux fourchettes de répartition) peuvent représenter de façon explicite les fourchettes implicites dans les lignes et les colonnes des tableaux publiés.

36. Les flux transfrontaliers de polluants et leurs effets globaux donnent un caractère international aux données environnementales. La coopération des pays voisins est nécessaire pour faire en sorte que les frontières nationales n'empêchent pas l'analyse des répercussions des procédés matériels qui dépassent ces frontières.

37. En résumé, la dimension régionale est particulièrement importante dans le cas des données environnementales, non seulement parce que le caractère local représente fréquemment le centre d'intérêt, mais aussi parce que les données doivent souvent être réagréguées selon des régions géographiques qui se prêtent mieux à l'analyse environnementale, comme les écozones et les bassins hydrographiques.

VII. Questions d'organisation et de diffusion

38. La plupart des BSN sont organisées en secteurs spécialisés. La production d'estimations régionales recoupe ces secteurs, et elle nécessite le soutien du personnel de la Géographie pour l'infrastructure géographique, de la Méthodologie pour les méthodes d'estimation et d'évaluation, et parfois de la Diffusion, pour l'intégration des données d'un secteur à l'autre. La question de l'organisation des estimations régionales à l'intérieur des BSN se pose donc.

39. Le fait de confier aux secteurs spécialisés la responsabilité de la gestion des estimations régionales les concernant, avec l'appui du personnel de la Méthodologie et de la Géographie, au besoin, représente un choix naturel, étant donné que ces secteurs sont probablement davantage au courant des exigences et des limites s'appliquant aux données dans leur domaine. Ce qui présente davantage un problème, c'est la façon de regrouper les données régionales en vue de leur diffusion aux utilisateurs. Qui devrait être responsable d'obtenir les données des divers secteurs spécialisés pour une région particulière? Devrait-il s'agir d'un programme régulier, ou devrait-on parfois le faire « sur demande »? Divers modèles s'offrent, et Statistique Canada les a essayés, pour la plupart, au fil des ans.

40. À un moment donné, une division axée sur les statistiques régionales ou urbaines a été mise en place pour assurer le caractère régional des données statistiques. À un autre moment, le programme de recensement, qui constitue évidemment la source la plus importante de données régionales, a été à l'avant-plan de la production de profil de données régionales. Parfois aussi, on a eu recours à un projet faisant intervenir plusieurs divisions pour gérer un programme de profils relatifs aux districts électoraux ou à d'autres régions géographiques. Parallèlement, les employés des bureaux régionaux ont joué un rôle clé pour regrouper des données régionales en réponse à des demandes des clients. Aucune de ces façons de faire n'est idéale. La production de profils a toujours nécessité beaucoup de travail, ainsi qu'une vaste compréhension des domaines spécialisés et de nombreuses recherches et manipulations de données. En dépit de l'existence de régions géographiques normalisées, la combinaison de données fondées sur plusieurs bases

géographiques différentes pose généralement un problème. Le contrôle de la qualité présente un déficit majeur.

41. Les profils sur papier planifiés au préalable n'ont jamais connu un très grand succès, ce qui fait que l'on a privilégié une stratégie répondant au maximum aux demandes des clients, au fur et à mesure qu'elles se présentent. Du fait des progrès récents de la technologie et de la couverture plus large des données régionales dans la base de données du Bureau, une approche davantage automatisée est possible. Une composante du site Internet de Statistique Canada (www.statcan.ca), appelée Profil statistique des communautés canadiennes et fondée pour une large part sur les données du recensement, constitue notre tentative la plus récente pour rendre les données régionales plus accessibles, et semble présager des orientations à venir dans ce domaine.

VIII. Conclusions

42. La production de statistiques régionales par un BSN soulève des problèmes qui sont différents au point de vue qualitatif de ceux découlant de la production régulière de données nationales et provinciales ou de données concernant d'autres grandes régions. La théorie statistique selon laquelle les données fondées sur des mesures individuelles suffisantes sont fiables de façon inhérente dans le cas des grandes régions (sans tenir compte du biais pour le moment), ne semble pas pouvoir être soutenue pour les régions plus petites. À moins de disposer d'un recensement ou d'une source administrative à jour, comportant une couverture complète, le BSN doit avoir recours à une forme de mesure axée sur un modèle pour produire des estimations. Étant donné que les divers modèles permettent de produire des estimations différentes, cela amène un degré d'arbitraire dans les estimations, et peut être perçu par certains comme nuisant à l'objectivité d'un BSN et à ses méthodes. Le principe fondamental d'ouverture et de transparence au sujet des méthodes, y compris le choix des modèles utilisés et les répercussions des différentes hypothèses, est encore plus important dans le domaine des estimations régionales.

43. Par-dessus tout, un BSN doit s'attendre à ce que les estimations régionales fassent l'objet d'une attention beaucoup plus étroite que nombre d'estimations relatives à de grandes régions. Même si les estimations pour les grandes régions reçoivent une attention plus marquée, peu de personnes ont la capacité de confirmer ou de réfuter une estimation au niveau national. Toutefois, au niveau local, nombreux sont ceux qui pensent savoir de quoi il retourne. Et de façon générale, les estimations régionales ne fonctionnent pas uniformément bien pour toutes les régions. Une méthode qui fonctionne bien en moyenne n'est pas à l'abri des critiques dans les régions où elle n'a pas bien fonctionné, à moins que cela n'ait été à l'avantage de ces régions! Le BSN doit être prêt à faire face au double obstacle que présentent des estimations plus faibles faisant l'objet d'une attention plus étroite.

44. Et comme si cela n'était pas suffisant, les considérations en matière de confidentialité sont plus grandes au niveau régional. Du fait même que des estimations sont produites pour des

régions, il est possible que des personnes puissent être identifiées, même si le BSN a pris suffisamment de précautions pour prévenir une telle divulgation. Certains utilisateurs des données régionales dans le secteur du marketing contribuent à accentuer le problème en soulignant dans leurs annonces qu'ils peuvent cibler le courrier qu'ils envoient aux ménages, selon les caractéristiques des personnes ou des ménages, lorsqu'ils utilisent les données régionales pour faire une distinction entre les divers quartiers. Certaines méthodes d'estimations régionales nécessitent le couplage d'enregistrements, ce qui soulève aussi des questions du point de vue de la protection des renseignements personnels. Encore une fois, une politique d'ouverture et un examen soigneux de toutes les applications, à un niveau supérieur et avant même que ces applications soient mises en oeuvre, est nécessaire pour s'assurer que les avantages pour le public surpassent les intrusions dans la vie privée des gens.

45. En dépit de ces difficultés, la demande de données régionales demeure élevée, la technologie offre de nouvelles approches pour la gestion et la diffusion des données régionales, et les travaux méthodologiques relatifs aux estimations régionales constituent un secteur actif de recherche parmi les statisticiens. Même si la production des données régionales ne constitue généralement pas la première priorité d'un BSN, la pertinence des programmes statistiques sera grandement améliorée si les BSN peuvent répondre aux besoins les plus importants en matière de données régionales.

Bibliographie

- Alexander, C.H. (1999). A rolling sample survey for yearly and decennial uses. *Proceedings of the 52nd Session of the International Statistical Institute*, Helsinki.
- Brackstone, G. (1987). Utilisation des dossiers administratifs à des fins statistiques, *Techniques d'enquête*, Vol. 13, No. 1, pp. 35-51.
- Fay, R.E., and Herriott, R.A. (1979). Estimates of Income for Small Places: an Application of James-Stein Procedure to Census Data, *Journal of American Statistical Association*, Vol. 74, pp. 269-277.
- Federal Committee on Statistical Methodology (May 1994). Report on Statistical Disclosure Limitation Methodology (Statistical Policy Working Paper #22). Washington, D.C., Office of Management and Budget, Office of Information and Regulatory Affairs, Statistical Policy Office.
- Gambino, J., and Dick, P. (2000). Small Area Estimation Practice at Statistics Canada, *Statistics in Transition*, Vol. 4, pp. 597-610.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: an appraisal, *Statistical Science*, Vol. 9, pp. 55-93.
- Isnard, M. (1999). *Alternatives to Traditional Census Taking: The French Experience*. Paris: INSEE.
- Jabine, Thomas B. (1993). Statistical Disclosure Limitation Practices of United States Statistical Agencies, *Journal of Official Statistics*, Vol. 9, No. 2, pp. 427-454.
- Kish, L. (1990). Recensement par étapes et échantillons avec renouvellement complet, *Techniques d'enquête*, Vol. 16, No. 1, pp 67-86.
- Kish, L. (1998). Space/time variations and rolling samples. *Journal of Official Statistics*, Vol. 14, pp 31-46.
- Purcell, N.J., and Kish, L. (1979). Estimation for Small Domains, *Biometrics*, Vol. 35, pp. 365-384.
- Rao, J.N.K. (1999). Quelques progrès récents concernant l'estimation régionale fondée sur un modèle, *Techniques d'enquête*, Vol. 25, No. 2, pp 199-212.

Redfern, P. (1989). L'expérience européenne relative à l'utilisation des données administratives pour recenser la population : questions d'ordre politique, *Techniques d'enquête*, Vol. 15, No.1, pp 85-103.

Singh, M.P., Gambino, J., and Mantel, H.J. (1994). Les petites régions : problèmes et solutions, *Techniques d'enquête*, Vol. 20, No. 1, pp 3-23.

Zayatz, L., Steel, P, and Rowland, S. (2000). Disclosure Limitation for Census 2000. Paper presented at the annual meeting of the ASA, August 2000. To appear in Proceedings.