Novembre 1991

SYMPOSIUM 91

Questions spatiales liées aux statistiques

RECUEIL

Statistique Canada Statistics Canada

Canada

SYMPOSIUM 91

Questions spatiales liées aux statistiques

12 au 14 novembre 1991

Ottawa (Ontario) Canada

RECUEIL

Juillet 1992

Comité organisateur du Symposium 91

Mary March
Caroline Weiss
Liane Chatterton
Joel Yan

Publication autorisée par le ministre responsable de Statistique Canada

PRÉFACE

Le Symposium 91 était le huitième de la série des symposiums internationaux qui ont été tenus annuellement à Statistique Canada depuis 1984. Chaque année, le symposium porte sur un thème particulier. Le thème en 1991 était la qualité des données.

En 1991, plus de 400 personnes de plusieurs pays ont assité au symposium. Les participants se sont regroupés pendant 3 jours dans la salle de conférences Simon Goldberg à Ottawa pour écouter les experts en qualité des données provenent de nombreuses agences gouvernementales, des universités, et de l'industrie privée. Au cours du symposium, les participants ont écouté 28 communications invitées réparties dans 10 sessions. Ces sessions ont touché à un large évantail de questions sur la qualité incluant: Perpectives géographiques sur la modélisation des données; Considérations spatiales lors de la conception des enquêtes ou des bases de sondage; Analyse spatiale des données sur la santé et l'environnement; Développements spatiaux au niveau du traitement des données; Innovations géographiques au niveau de la collecte des données; Qualité des produits de données spatiales; Géographie médicale; Analyse spatiale des données d'enquêtes; Cadres géographiques pour les données statistiques; Analyse des données sous des perspectives géographiques.

À part la traduction et la mise en page, le recueil du Symposium 91 est une copie conforme des articles tels que soumis par les auteurs. L'ordre de présentation des articles est le même que celui du Symposium.

Le comité du Symposium 91 désire souligner les contributions de plusieurs personnes qui ont aidé à préparer ce Recueil.

Naturellement, nous remercions les orateurs du Symposium 91 qui ont pris le temps de rédiger leur communication et de la soumettre pour fin d'inclusion dans ce livre. Les efforts de plusieurs autres personnes ont été d'importance vitale pour la publication de ce Recueil. Christine Larabie, Carmen Lacroix et Judy Clarke se sont occuppées du traitement des manuscrits d'une façon experte. La révision de la traduction a été faite par de nombreux méthodologistes et experts du contenu: S. Auger, Y. Beaucage, Y. Bélanger, J.-R. Boudreau, R. Boyer, M. Bureau, E. Castonguay, P. Daoust, P. David, S. Giroux, T. Labillois, G. Laflamme, J. Morel, C. Morin, Y. Morin, S. Nadon, G. Parsons, C. Poirier, L. Swain, J. Tourigny et P. Whitridge. Christine Larabie a co-ordonné la traduction, correction d'épreuves, et la production du recueil. Neil Pecore a collaboré à la préparation du manuscrit.

Le neuvième symposium annuel et international de Statistique Canada sera tenu à Ottawa du 2 au 4 novembre 1992. Le titre sera "Conception et analyse des enquêtes longitudinales".

Comité du Symposium 91

Juillet 1992

LA SÉRIE DES SYMPOSIUMS DE STATISTIQUE CANADA

1984 - L'analyse des données d'enquête
1985 - Les statistiques sur les petites régions
1986 - Les données manquantes dans les enquêtes
1987 - Les utilisations statistiques des données administratives
1988 - Les répercussions de la technologie de pointe sur les enquêtes
1989 - L'analyse des données dans le temps
1990 - Mesure et amélioration de la qualité des données
1991 - Questions spatiales liées aux statistiques

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

Pour commander des copies additionelles du recueil du Symposium 91: "Questions spatiales liées aux statistiques", utilisez le bon de commande sur cette page. Un nombre limité de copies des receuils des symposiums 1987, 1988, 1989 et 1990 sont aussi disponibles. Pour commander, envoyez cette formule à l'addresse suivante:

RECUEIL DU SYMPOSIUM 91
STATISTIQUE CANADA
DIVISION DES OPÉRATIONS FINANCIÈRES
ÉDIFICE R.H. COATS, 6e ÉTAGE
PARC TUNNEY
OTTAWA (ONTARIO)
K1A 0T6
CANADA

Veuillez inclure le paiement avec votre commande (chèque ou mandat, en dollars canadiens ou l'équivalent, à l'ordre du "Receveur général du Canada - Recueil du Symposium 91").

RECUEIL DU SYMPOSIUM: NUMÉROS DISPONIBLES

1987 -	Les utilisations statistiques des données administratives - ANGLAIS	____ @ $10 CHACUN
1987 -	Les utilisations statistiques des données administratives - FRANÇAIS	____ @ $10 CHACUN
1987 -	ENSEMBLE DE 1 ANGLAIS ET 1 FRANÇAIS	____ @ $12 L'ENSEMBLE
1988 -	Les répercussions de la technologie de pointe sur les enquêtes - BILINGUE	____ @ $10 CHACUN
1989 -	L'analyse des données dans le temps - BILINGUE	____ @ $20 CHACUN
1990 -	Mesure et amélioration de la qualité des données - ANGLAIS	____ @ $15 CHACUN
1990 -	Mesure et amélioration de la qualité des données - FRANÇAIS	____ @ $15 CHACUN
1990 -	ENSEMBLE DE 1 ANGLAIS ET 1 FRANÇAIS	____ @ $25 L'ENSEMBLE
1991 -	Questions spatiales liées aux statistiques - ANGLAIS	____ @ $20 CHACUN
1991 -	Questions spatiales liées aux statistiques - FRANÇAIS	____ @ $20 CHACUN
1991 -	ENSEMBLE DE 1 ANGLAIS ET 1 FRANÇAIS	____ @ $35 L'ENSEMBLE

S.V.P. AJOUTEZ $2 PAR LIVRE POUR LES FRAIS DE LIVRAISON $ ____

MONTANT TOTAL DE LA COMMANDE $ ____
(les prix incluent la TPS; numéro d'enregistrement de la TPS: R121491807)

S.V.P. INCLURE VOTRE ADDRESSE POSTALE COMPLÈTE AVEC VOTRE COMMANDE !

NOM _____

ADDRESSE _____

_____

VILLE _____ PROV/ÉTAT _____ PAYS _____

CODE POSTAL _____ TÉLÉPHONE (____)_____ FAX _____

S.V.P. Notez: Chaque participant au Symposium 91 qui n'est pas un employé de Statistique Canada recevra une copie gratuite du recueil du Symposium 91.

QUESTIONS SPATIALES LIÉES AUX STATISTIQUES

TABLE DES MATIÈRES[1]

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

S.V.P. Notez: La langue originale de tous les articles du Symposium 91 était l'anglais sauf pour les suivants: Mot d'ouverture, I.P. Fellegi (bilingue); Session 4, J. Muzard (français), P. Daoust (français); Session 6, P. Ardilly (français).

MOT D'OUVERTURE

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

MOT D'OUVERTURE

I.P. Fellegi[1]

Bienvenue au Symposium 91. Le symposium sur la méthodologie est devenu un événement annuel à Statistique Canada. Depuis 1984, nous en avons organisé huit, chacun portant sur un sujet différent. Les thèmes des symposiums antérieurs ont été les suivants:

L'analyse des données d'enquêtes;
Les statistiques pour les petites régions;
Les données manquantes dans les enquêtes;
Les utilisations statistiques des données administratives;
Les répercussions de la technologie de pointe sur les enquêtes;
L'analyse des données dans le temps;
La mesure et l'amélioration de la qualité des données.

Cette année, le Symposium a pour thème «Les questions spatiales liées aux statistiques». Trois raisons nous ont incité à choisir ce thème.

La première se rapporte aux besoins de nos clients. Étant donné les grandes disparités régionales qui caractérisent le Canada, il nous a toujours fallu présenter nos produits selon une répartition géographique significative. Cependant, il existe désormais toute une gamme d'utilisations pour lesquelles la capacité d'analyse géographique est une condition préalable. De toute évidence, l'étude des questions environnementales figure au nombre de ces utilisations: il est même difficile de parler de statistiques relatives à l'environnement, si ce n'est dans un contexte géographique. L'analyse des risques pour la santé liés à l'environnement en est un autre exemple. En effet, les atlas sur la santé représentent une grande amélioration au chapitre de l'analyse explorative des données. En commercialisation, on utilise avec une souplesse toujours plus grande des données à référence spatiale. Et ainsi de suite.

La deuxième raison a trait à la technologie. Au cours des dix dernières années, des progrès considérables ont été réalisés dans le domaine des systèmes d'information géographique, ou SIG. Les systèmes assistés par ordinateur sont de plus en plus puissants et, en même temps, plus accessibles. Il est maintenant possible de saisir, de stocker, de réviser, de traiter, de manipuler, de consulter, d'extraire, d'analyser et d'afficher des quantités importantes de données spatiales. En outre, la qualité des données que produisent ces systèmes peut être améliorée grâce aux méthodes perfectionnées de collecte des données et aux capacités de révision des systèmes. Fait plus important encore, un nombre croissant de nos clients ont accès directement ou indirectement à des SIG.

La troisième raison pour laquelle «Les questions spatiales liées aux statistiques» pourraient être considérées comme un sujet d'actualité est d'ordre méthodologique. Les chercheurs et les praticiens, y compris ceux de Statistique Canada, s'intéressent de plus en plus au domaine de la statistique géographique. Cela implique l'application de méthodes spéciales de manipulation et d'analyse des données réparties géographiquement quand on pense à l'étude de questions comme l'autocorrélation spatiale et à l'utilisation de concepts tels que «modèles de diffusion spatiale», «fractales» et «statistique directionnelle». À l'heure des SIG, on redécouvre la statistique géographique; la technologie évoluée et l'utilisation plus répandue des systèmes experts facilitent grandement l'application de ces méthodes.

---

[1] I.P. Fellegi, Statisticien en chef, Statistique Canada, Parc Tunney, 26-A, Immeuble R.H.-Coats, Ottawa (Ontario), Canada K1A 0T6.

La dimension spatiale devient de plus en plus une partie intégrante de notre travail à Statistique Canada. Afin de recueillir l'information efficacement, nous pouvons utiliser les données spatiales pour planifier et contrôler les opérations de collecte. Pour être utile, l'information que nous fournissons doit être présentée selon une répartition géographique variée, avec souplesse, de façon opportune et selon un mode de visualisation convivial. Les données spatiales nous permettent en outre de produire de meilleures estimations et de faire de meilleures analyses et ainsi d'améliorer la justesse de nos résultats.

Ces réalisations et ces préoccupations sont au centre de notre plus importante activité qu'est le recensement. La planification et l'organisation seules du processus de collecte, qui visent à garantir une couverture efficace et intégrale de la population, requièrent l'entreposage et l'utilisation d'une quantité considérable de données spatiales. Celles-ci servent à délimiter les secteurs de dénombrement, à préparer des cartes, à calculer des coûts, à établir les différents taux à la pièce pour nos recenseurs en fonction du secteur de dénombrement dont ils sont chargés, et à créer des méthodes de contrôle de la qualité des données comme le registre des adresses employé dans le cadre du Recensement de la population, et les plans de cantons employés dans le cadre du Recensement de l'agriculture. Au cours des années, on a également eu recours aux données spatiales pour le traitement des données recueillies, y compris leur saisie, leur vérification, leur imputation et leur pondération. Dans le cas qui nous préoccupe, les données spatiales servent à contrôler la qualité de ces processus et à subdiviser la population en groupes en vue de la pondération et de l'imputation. Un autre outil géographique permet d'accroître grandement la valeur des données de recensement. Ces dernières sont liées à des codes de régions géographiques, à un niveau assez détaillé, par un processus appelé géocodage qui lie les coordonnées du système de projection transverse de Mercator au centroïde des secteurs de dénombrement et des côtés d'îlot. Une fois leur traitement terminé, les données de recensement sont largement exploitées, ce qui est en grande partie attribuable à la valeur des données géographiques qui les accompagnent. Il nous est par ailleurs possible de diffuser les données de recensement sous divers formats utiles comme l'illustre la série d'atlas métropolitains qui s'est révélée un produit populaire parmi la gamme des produits du recensement. Un nouveau produit, qui fait actuellement l'objet d'un test pilote au moyen de données du recensement de 1986, sera offert au nombre des produits du recensement de 1991. Il s'agit de données accompagnées d'une carte à codage numérique du Canada segmentée en très petites unités de base et liées par un système d'information géographique complexe.

Dans le passé, les efforts de nos spécialistes des systèmes d'information géographique et de nos géographes ont surtout été centrés sur des innovations à l'appui du recensement. Leurs réalisations sont nombreuses. Au nombre de celles-ci, et non la moindre, mentionnons la définition automatisée des limites de plus de 40% des secteurs de dénombrement dans les grands centres urbains (à l'aide d'une technique appelée délimitation de districts assistée par ordinateur) et la production automatisée de cartes de première qualité (à l'aide d'un programme de production de cartes par ordinateur) pour plus de la moitié des secteurs de dénombrement pour le recensement de 1991.

Ces dernières années, les techniques que nous avons mises au point dans le domaine de l'information géographique ont commencé à être appliquées plus largement au Bureau. Des données provenant de différentes sources sont raccordées grâce à l'utilisation de renseignements géographiques comme les codes postaux, ce qui donne lieu à la création d'ensembles de données améliorés. Les renseignements recueillis dans le cadre d'enquêtes ou par d'autres moyens sont affichés géographiquement. Notre système d'information sur l'environnement en est un bon exemple. Ce système contient une riche série de données tirées d'enquêtes sociales, économiques et démographiques, ainsi que des mesures physiques et des données relatives à des caractéristiques géographiques. Nous avons également développé la capacité d'exploiter des données obtenues par télédétection, principalement des données sur l'utilisation des terres agricoles, ce qui nous permet d'enrichir les renseignements obtenus grâce aux méthodes traditionnelles, c'est-à-dire les enquêtes et les sources administratives.

De plus, nous commençons à utiliser des concepts qui relèvent de la statistique géographique à des fins d'analyse de données, notamment dans les domaines de la santé et de l'environnement. Ces domaines sont en effet fortement associés à la géographie. Les progrès réalisés sur les plans de la technologie et de la méthodologie constituent certainement des forces qui contribuent à l'amélioration des résultats.

Un certain nombre de questions touchant les données spatiales sont prévues à l'ordre du jour des séances. Il existe un intérêt évident pour la création d'une interface reliant les données spatiales et les statistiques. Vu

l'intérêt que manifeste le très grand nombre de personnes qui ont choisi d'assister à la réunion, cette dernière constitue peut-être le "test de marché" idéal qui prouve que le thème choisi est en effet un thème d'actualité. Nous souhaitons la bienvenue aux participants canadiens, ainsi qu'à ceux de la France et des États-Unis, sans oublier les représentants des administrations publiques, nos représentants statistiques provinciaux et les représentants des milieux de l'industrie et de l'enseignement. Nous espérons que cet échange d'idées se révélera très fructueux.

De la part des organisateurs du Symposium, je tiens à remercier spécialement les membres du Laboratoire de recherche en statistique et probabilité de Carleton University et de l'Université d'Ottawa pour avoir cette année encore contribué à l'organisation de cet événement. Je souhaite également la bienvenue à un nouveau collaborateur, l'Association canadienne des géographes, dont l'appui enthousiaste est grandement apprécié. Nous remercions tous nos co-organisateurs pour leur aide, leurs conseils et leur soutien financier et moral.

Enfin, je souhaite la bienvenue à tous et j'espère sincèrement que votre participation à ce Symposium sera des plus enrichissantes.

DISCOURS - PROGRAMME

Recueil du Symposium 91 de Statistique Canada
Questions spatiales reliées aux statistique
Novembre 1991

LA GÉOGRAPHIE ET LES STATISTICIENS

M.F. Goodchild[1]

RÉSUMÉ

Les données spatiales posent des problèmes uniques au niveau de la statistique en raison de leurs caractéristiques inhérentes, plus particulièrement leur hétérogénéité et leurs dépendance. L'article porte sur les effets concrets de ces caractéristiques sur des éléments tels que les unités déclarantes modifiables et leurs erreurs écologiques. On y examine, d'une part l'évolution de l'intégration des données spatiales en statistique et, d'autre part, l'évolution des méthodes statistiques en géographie. L'intérêt que suscitent actuellement les systèmes d'information géographique permet de rapprocher ces deux groupes et d'examiner, de façon systématique pratique, des questions qui existent depuis longtemps. Les progrès récents au niveau de l'analyse spatiale exploratoire et de la visualisation ajoutent aux effets de la technologie numérique et augmentent la valeur de la perspective spatiale.

MOTS CLÉS: Données spatiales; statistiques spatiales; cartographie thématique; SIG.

1. INTRODUCTION

La géographie a toujours été une discipline axée sur les découvertes. À l'époque classique, elle était une forme de mathématiques, où l'on s'intéressait à mesurer et à projeter la terre et à élaborer des méthodes de navigation. Du 15e au 19e siècles, la géographie a servi à explorer et à cartographier la terre et ses praticiens étaient des navigateurs et des chercheurs en histoire naturelle. Maintenant que nous connaissons la source du Nil et que les nouvelles mesures de l'altitude du Mont Everest ne captivent plus l'imagination du public, la géographie s'est tournée vers l'étude des processus qui façonnent et qui modifient le paysage physique, vers les liens dynamiques et évolutifs entre l'humanité et l'environnement ainsi que vers celle des processus qui forment la diversité géographique de la culture humaine. La géographie est infiniment complexe: plus on étudie de près l'un quelconque de ses aspects, humains ou physiques, plus on voit de détails et plus on trouve d'éléments à expliquer et à comprendre (Mandelbrot 1967, traduction libre).

La géographie a découvert les statistiques au cours de sa révolution quantitative qui a commencé dans les années 50 et qui s'est poursuivie jusqu'à la fin des années 60. Pour la majorité des personnes, cela signifie l'application de la batterie courante de techniques statistiques, allant du test t de Student et du test F a l'analyse en composantes principales et à la corrélation canonique, appliqués aux cas observés d'un phénomène. On n'a pas accordé une grande importance au fait que ces cas étaient noyés dans un continuum spatio-temporel - l'espace et le temps ne faisaient que fournir la base de sondage.

Pour un plus petit groupe de personnes, il est devenu de plus en plus apparent que la géographie constituait un cas spécial. Certaines personnes ont découvert la statistique spatiale et ont utilisé des techniques comme l'analyse de modèles de points pour faire des inférences à propos des processus spatiaux (Getis et Boots 1978). D'autres ont étudié l'autocorrélation spatiale (Cliff et Ord 1981), les champs aléatoires, les variables

---

[1] M.F. Goodchild, Directeur, National Center for Geographic Information and Analysis, Université de la Californie, Santa Barbara (Californie) 93106, États-Unis.

régionalisées ou la géostatistique (Isaaks et Srivastava 1989) et certaines personnes ont apporté des contributions importantes. Mais ces domaines sont difficiles et, généralement, ils n'ont pas été inclus dans les sujets traités dans les cours habituels de statistique destinés aux géographes. Même aujourd'hui, on traite très peu de la statistique spatiale dans le manuel typique utilisé dans les cours destinés aux étudiants du premier cycle, qui se concentre sur l'application de la batterie normale de tests statistiques (non spatiaux) aux problèmes géographiques (voir par exemple, Barber 1988; Clark et Hosking 1986; pour une exception marquée, voir Griffith et Amrhein 1991).

Il est probable que les progrès qui se poursuivent et qui ont commencé au cours des années 80 modifieront cette situation pour le mieux. Je vais désigner ces progrès par l'expression «systèmes d'information géographique» (SIG), bien que j'utilise ce terme dans un sens beaucoup plus étendu que celui qui s'applique à l'ensemble des progiciels de SIG disponibles actuellement (voir par exemple Burrough 1986; Maguire, Goodchild et Rhind 1991). La présente communication est divisée en quatre sections. Dans la première, on détermine les domaines de la statistique qui ont le plus de rapports avec l'analyse géographique et les SIG. Dans la deuxième section, on étudie certaines questions importantes portant sur l'application de la statistique aux données géographiques dans un contexte de SIG. Vient ensuite une section qui porte sur l'utilisation des méthodes statistiques pour résoudre le problème de l'exactitude dans les données spatiales, ce qui constitue un problème-clé pour l'utilisation de la technologie des SIG. Finalement, la dernière section porte sur des questions de mise en application.

2. QUELS GENRES DE STATISTIQUES?

Il se peut que la dépendance spatiale constitue la propriété la plus frappante des données géographiques d'un point de vue statistique. Les géographes pourraient exprimer ce concept de la façon suivante: il existe un lien entre tous les endroits mais le lien est plus étroit entre des endroits rapprochés qu'entre des endroits éloignés. Il est presque impossible de concevoir une variable indépendante du point de vue spatial, une variable dont la valeur ne pourrait être prédite sur les distances les plus courtes. Le variogramme utilisé en géostatistique montre la variance empirique entre des observations comme fonction de la distance qui les sépare et on observe que cette variance augmente de façon monotone jusqu'à un «seuil» à une distance désignée par le mot «portée» puis cesse d'augmenter, pour la majorité des variables réparties géographiquement. Cette «portée» est aussi connue en géomorphologie comme la «trame» du paysage, la distance passée par laquelle le paysage cesse de fournir d'autres surprises.

Une façon de comprendre l'effet de la dépendance spatiale consiste à la considérer comme un gonflement des degrés de liberté d'un test. Il se peut qu'une observation prise à peu de distance d'une autre ne soit pas une «nouvelle» observation, s'il existe une forte dépendance spatiale pour la variable (Richardson 1990).

L'hétérogénéité spatiale (Anselin 1989) constitue une deuxième propriété qui, elle aussi, est souvent négligée. Il n'est pas rare que les coefficients des modèles varient spatialement et le fait d'essayer d'estimer une valeur «moyenne» pour une zone étudiée définie arbitrairement peut être dénué de sens. Il se peut plutôt que les géographes préfèrent utiliser des techniques telles que le filtrage spatial adaptatif, les méthodes de développement (Casetti 1972) ou le «brushing» géographique (Monmonier 1989) qui estiment explicitement la variabilité spatiale des coefficients du modèle. Dans ce contexte, l'opinion conventionnelle voulant que la zone étudiée constitue un échantillon tiré d'une population plus grande et normalement non précisée est très inappropriée.

L'existence de structures hiérarchiques constitue un troisième problème dans l'analyse des données géographiques. Des objets géographique simples peuvent faire partie d'objets complexes plus gros et il arrive souvent que les processus puissent s'appliquer quelle que soit l'échelle. Dans ce sens, il peut être très restrictif de considérer un échantillon comme un ensemble d'objets semblables avec des attributs associés. Du point de

vue des bases de données, le modèle des «tables à deux dimensions» qui est implicite dans de nombreux tests statistiques limite considérablement notre aptitude à comprendre les processus spatiaux.

Finalement, l'espace géographique est, en soi, continu et infiniment complexe; sa description complète exigerait un nombre infini d'enregistrements ou de «tuples». Pour décrire l'espace géographique en termes finis, il faut effectuer une discrétisation explicite, à l'aide de l'échantillonnage, du groupement, de la généralisation, de la modélisation ou de l'approximation. Donc, les données géographiques sont presque toujours inexactes, bien qu'il se peut que le niveau d'incertitude associé à chaque article soit parfois connu. Les objets qui peuplent une base de données spatiales - les courbes de niveau, les points échantillons, les zones sur les cartes de la couverture végétale, les secteurs de recensement - peuvent n'être que des artefacts de ce processus de discrétisation et leur rapport avec le continuum sous-jacent peut être difficile à caractériser, ou même inconnu.

Pour résumer, en géographie on a besoin de statistiques qui reconnaissent explicitement la dépendance spatiale et on a beaucoup de difficultés avec l'hypothèse d'indépendance faite dans un si grand nombre de tests statistiques. Sans eux, le processus d'inférence sera vraisemblablement trompeur. La géographie s'intéresse à des processus qui ne sont pas stationnaires du point de vue spatial et à des techniques qui peuvent estimer la variation spatiale des coefficients, puisqu'il est souvent déraisonnable de supposer l'immobilité dans les limites arbitraires de la zone étudiée. En géographie, on utilise des processus qui peuvent être appliqués quelle que soit l'échelle et avec des relations hiérarchiques complexes. Et, finalement, de nombreuses observations géographiques représentent des échantillonnages ou des discrétisations arbitraires d'une variation continue plutôt que des attributs d'objets réels.

Ces problèmes sous-jacents surgissent souvent dans des préoccupations portant sur deux problèmes particuliers: le problème de l'unité spatiale modifiable (PUSM) (Modifiable Areal Unit Problem) et celui de l'erreur écologique (EE) (Ecological Fallacy). Openshaw et Taylor (1979) ont peut-être fourni l'illustration la plus spectaculaire du PUSM, quand ils ont montré qu'il suffisait de regrouper à nouveau, selon des zones de déclaration différentes, deux variables spatiales pour que les corrélations établies entre ces dernières aillent de -0,99 à +0,99. Sans dépendance ou hétérogénéité spatiale, les zones de déclaration seraient analogues à des échantillons d'une population sous-jacente et leurs effets ne différeraient pas des effets de l'échantillonnage. Mais Openshaw et Taylor ont montré clairement que les zones de déclaration ont un effet beaucoup plus important, ce que les politiciens qui ont recours au remaniement arbitraire des circonscriptions savent depuis longtemps.

On commet une erreur écologique (EE) (Robinson 1950) quand on suppose que toutes les composantes d'un agrégat géographique possèdent les propriétés de l'agrégat dans son ensemble. Par opposition au PUSM, l'EE ne se produirait jamais si les zones de déclaration étaient toujours parfaitement homogènes. Pourtant, on la fait constamment et l'hypothèse d'homogénéité est ce qui se trouve à la base de l'utilisation des techniques de groupement (Weiss 1988) et d'un certain nombre d'autres méthodes utilisées en étude de marché. On pourrait facilement calculer et déclarer des mesures de diversité pour les zones de déclaration du recensement et cela pourrait se faire sans contrevenir de façon importante aux règles de confidentialité; l'usage qui veut qu'on ne déclare que les mesures de la tendance centrale et les totaux doit parfois avoir l'effet d'encourager les utilisateurs des données du recensement à commettre l'erreur écologique.

3. PROBLÈMES RELATIFS AUX APPLICATIONS

Jusqu'au début des années 80, la technologie numérique offrait très peu de moyens pour traiter les statistiques spatiales. Les progiciels statistiques comme SAS, SPSS ou S n'offraient que des méthodes non spatiales. On pouvait traiter des renseignements spatiaux à l'aide de progiciels de cartographie par ordinateur, mais seulement pour produire une carte. Certains progiciels comme SAS possédaient aussi des fonctions limitées de cartographie, mais les données spatiales employées pour produire des cartes ne pouvaient être utilisées pour réaliser une analyse spatiale.

C'est au début des années 80 que les systèmes d'information géographique sont devenus disponibles sur une grande échelle et ils se sont répandus rapidement dans les universités, les organismes gouvernementaux et le secteur privé. Contrairement aux progiciels de cartographie par ordinateur, leur objet principal est d'effectuer l'analyse de données spatiales, bien qu'il arrive souvent que les fonctions dont ils disposent à cette fin soient limitées à des opérations géométriques simples.

C'est aussi au cours des années 80 que les progiciels permettant de traiter les statistiques spatiales sont apparus, bien qu'ils n'aient pas eu autant d'incidence que les SIG. Certains, comme le progiciel GEOEAS, largement répandu, utilisé à des fins de géostatique et élaboré par la Environmental Protection Agency des États-Unis, ou SPACESTAT de Anselin (Anselin 1990), sont des progiciels autonomes avec des fonctions de cartographie limitées. Griffith (1988) a élaboré des procédures pour traiter des statistiques spatiales à l'aide de SAS et de MINITAB, et d'autres personnes ont élaboré des liens entre les SIG et des progiciels statistiques standard ou des modules autonomes (Ding et Fotheringham 1991). Tous ces progiciels visent à réaliser des tests standard sur des ensembles bien définis de données d'entrée.

Les SIG constituent une technologie des années 80 et leur modèle interactif sous-jacent a peu de rapports avec le traitement par lots des années 60. En soi, ils constituent une base idéale pour appuyer une approche exploratoire à l'analyse statistique. En fait, il y a récemment eu un intérêt croissant pour le concept d'analyse spatiale exploratoire, par analogie au modèle de l'analyse exploratoire des données, qui a envahi la statistique à la fin des années 70. Il semble particulièrement inapproprié d'adopter un processus de confirmation par étapes, compte tenu des difficultés imposées par les quatre caractéristiques des données spatiales déjà mentionnées. On devrait plutôt encourager les chercheurs travaillant avec des données géographiques à considérer la technologie comme une façon d'explorer un ensemble d'idées mal formulées. La visualisation est importante pour ce processus, les limites de confiance peuvent être plus utiles que les tests d'hypothèse et les techniques de randomisation peuvent être plus robustes que les tests paramétriques plus traditionnels.

3.1 Dépendance spatiale

Nous supposons que la dépendance spatiale est toujours présente dans les données géographiques et qu'elle doit être reconnue explicitement à moins que les échantillons soient distancés de plus que la portée. Si un test de dépendance spatiale (Cliff et Ord 1981) montre que cette dernière est absente, cette inférence constituera toujours une erreur statistique de deuxième espèce - l'hypothèse nulle d'une absence de dépendance spatiale aura été acceptée quand, en fait, elle est fausse. Les degrés de liberté seront toujours gonflés.

Cette position tranche par rapport à la tradition, en particulier par rapport à l'enseignement classique qui considère la dépendance spatiale comme une exception ou un problème et où l'absence de dépendance spatiale entre les résidus est souvent utilisée comme diagnostic d'un modèle entièrement spécifié. Malheureusement, les modèles de dépendance spatiale sont complexes, en partie à cause de la nature simultanée de la dépendance spatiale qui s'oppose à la nature unidirectionnelle de la dépendance temporelle. Il se peut que la visualisation constitue la seule façon de communiquer des renseignements sur la dépendance spatiale, parce que les paramètres des modèles de dépendance spatiale ont très peu de signification sur le plan intuitif.

On dispose de deux familles de techniques pour traiter la dépendance spatiale, selon la représentation de l'espace. L'expression «statistiques spatiales» se rapporte normalement aux techniques qui supposent un espace peuplé d'objets bien définis qui interagissent comme des touts (Arbia 1989), bien qu'il puisse en fait s'agir d'objets définis arbitrairement qui servent de zone de déclaration pour un continuum sous-jacent. Les mesures de la dépendance spatiale comprennent les mesures de Moran et Geary de l'autocorrélation spatiale (Cliff et Ord 1981) et se concentrent sur le niveau de dépendance entre des objets voisins. Des modèles autorégressifs et à moyenne mobile sont disponibles pour modéliser les processus entre des objets. Une bonne partie du logiciel qu'on peut employer actuellement utilise des progiciels standard (Griffith 1988). Finalement, pour ces techniques, la représentation de l'espace est réduite à une simple matrice de facteurs de contiguïté, de facteurs

de proximité ou de facteurs de connectivité entre des objets et les résultats sont invariants pour toute transformation spatiale qui préserve cette matrice.

Le terme «géostatistique» est utilisé pour des techniques qui modélisent une distribution sous-jacente continue par tous les échantillons de points et d'îlots. L'interpolation de cette surface sous-jacente est souvent le principal objectif. Les variables sont mesurées sur des échelles continues et la dépendance spatiale est décrite à l'aide du variogramme ou du spectre. Le krigeage est peut-être la technique de géostatistique la plus connue (Isaaks et Srivastava 1989).

3.2 Autres méthodes utilisées en statistique spatiale

En plus de ces techniques, la statistique spatiale comprend une gamme de tests de divers modèles de structures. Dans le cas le plus simple, il est souvent souhaitable de faire un test pour savoir si une structure de points aurait pu être produite par un processus stochastique donné. Pour déterminer des grappes de cas en épidémiologie, par exemple, on doit disposer de tests qui permettent de faire la distinction entre des taux de maladies constants dans l'espace et d'autres qui varient dans l'espace et, dans ce dernier cas, les tests doivent pouvoir fournir des estimations de la répartition géographique des taux. Dans d'autres cas, il peut être souhaitable de faire la distinction entre des grappes qui découlent de taux qui varient spatialement et celles qui découlent de processus contagieux. La machine d'analyse géographique (Geographical Analysis Machine) d'Openshaw (Openshaw et coll. 1987) constitue un exemple stimulant d'utilisation de la technologie des SIG pour faire des tests portant sur des grappes.

De nombreux indices descriptifs ont été définis pour des structures géographiques. Il se peut que le plus connu de ces indices soit le centroïde, défini comme la moyenne des abscisses et des ordonnées d'un ensemble de points, invariant par rapport à la rotation des axes, ainsi que le point qui minimise le total des carrés des distances à l'ensemble de points. Le mouvement du centroïde constitue un indicateur utile des changements dans les répartitions géographiques. Les centroïdes fournissent aussi les seuls renseignements disponibles sur les répartitions géographiques de la population au plus bas niveau de chaque système national de zones de déclaration - le SD au Canada et le DD ou îlot aux É.-U. Au R.-U., des travaux récents effectués par Bracken et Martin (1989) fournissent des estimations de répartitions continues à de très grandes échelles en décomposant les chiffres pour des centroïdes à l'aide de fonctions noyaux précisées de façon intelligente. Les centroïdes et autres mesures de tendance centrale en géographie ont aussi une valeur normative comme endroits logiques à partir desquels on peut fournir un service à une population dispersée. Des techniques plus évoluées et plus appropriées pour déterminer l'emplacement d'installations centrales sont traitées, dans un contexte pratique, par Ghosh et Rushton (1987).

3.3 Analyse spatiale exploratoire

L'analyse exploratoire des données (AED) est déjà bien acceptée comme un modèle en analyse statistique de données non spatiales. Elle fournit un ensemble de techniques pour ajouter à l'intuition naturelle du chercheur et pour élaborer des hypothèses qui peuvent, par la suite, être vérifiées de façon plus objective et délibérée. Il semble donc raisonnable de supposer que l'analyse spatiale exploratoire ou ASE peut jouer un rôle semblable. Dans un petit nombre d'articles on a déjà étudié les possibilités de l'ASE. En fait, les indices préliminaires laissent supposer que l'ASE diffère considérablement de l'AED et qu'elle peut offrir beaucoup plus de possibilités.

Nous définissons ici l'ASE, en termes comparatifs, comme un ensemble de techniques conçues pour apporter des améliorations ou ajouter à l'aptitude d'un chercheur à tirer des conclusions après l'étude de données dans leur contexte spatial. Les genres les plus valables d'ASE seraient donc ceux conçus afin d'aider le chercheur dans des domaines du raisonnement spatial où l'intuition humaine est peu fiable. Certaines de ces analyses sont évidentes et ont déjà été utilisées pour justifier les SIG dans le passé: la difficulté à combiner des données provenant de différentes sources spatiales, surmontée dans un SIG à l'aide de l'opération de recouvrement; ou

la difficulté à combiner des données de formes différentes - position (structure), attributs, son et texte - quand elle est limitée par les possibilités d'un écran sur lequel est affiché une carte. Les raisons invoquées pour le recouvrement s'appliquent également dans le cas de la détection du changement dans une série d'images: l'oeil n'est pas un bon instrument pour enregistrer plusieurs types de données et pour les distinguer, un SIG peut donc se révéler un outil très efficace pour détecter un changement dans une série chronologique spatio-temporelle. L'intuition humaine non plus n'est pas bonne pour mesurer des objets spatiaux - des documents publiés de longue date en cartographie décrivent le problème qu'impose l'obligation de tenir compte de la perception par l'oeil humain de la taille d'un objet.

Les outils de l'ASE peuvent également diminuer la difficulté qu'a l'oeil à intégrer sous une surface, pour effectuer des estimations de la population totale dans une région arbitraire tirée d'une carte de courbes de niveau, par exemple. L'opération inverse aussi est difficile - l'estimation d'une surface de densité à partir d'une carte de points. Les personnes réussissent mal à faire des raisonnements d'une échelle à l'autre ou à effectuer l'essai de structures par rapport à des modèles de processus statistiques - il est facile de trouver de nombreux cas où l'intuition a amené des personnes à penser, à tort, qu'une structure aléatoire de points possédait une certaine tendance systématique. Et finalement, l'oeil n'est pas un bon outil pour séparer les tendances régionales des tendances locales ou pour détecter des valeurs spatiales aberrantes, particulièrement dans des données à plusieurs variables.

Le système SPIDER de Haslett et autres systèmes semblables (Haslett et coll. 1991) utilisent plusieurs fenêtres sur un écran pour montrer des données dans différentes perspectives - carte, série chronologique, nuage de points, histogramme - puis pour les coupler logiquement afin que les opérations effectuées à l'aide de la souris dans une fenêtre soient reflétées dans les autres. Le fait de désigner une région sur une carte entraîne la mise en évidence du point correspondant dans un nuage de points. L'animation de séries chronologiques constitue aussi un moyen simple mais efficace d'aider les êtres humains à obtenir une compréhension intuitive des processus spatiaux.

4. LE PROBLÈME DE L'EXACTITUDE

Nous avons déjà fait remarquer que presque toutes les données spatiales sont fondamentalement incertaines et inexactes. En même temps, il existe une tradition bien établie en cartographie qui vise à établir des normes portant sur l'exactitude des objets qui figurent sur les cartes, particulièrement l'exactitude de la position. De nombreux organismes réglementent la distance en dedans de laquelle l'emplacement véritable d'un point doit se trouver dans 90% des cas - la norme de précision cartographique circulaire (NPCC) - et la fixent à une fraction de millimètre sur le produit cartographique fini. Mais, pour de nombreux genres de données géographiques, il est impossible de mesurer la NPCC parce que les objets représentés sont des artefacts du processus de discrétisation et qu'ils ne peuvent être repérés sur le terrain. La NPCC porte aussi sur l'exactitude des points et elle ne peut être appliquée à des objets complexes, qu'il s'agisse de lignes ou de surfaces, qui sont plus qu'une collection de points discrets.

Beaucoup de recherches ont été consacrées à l'élaboration de modèles des erreurs pour des bases de données spatiales. Nous définissons un modèle des erreurs comme un processus stochastique qui peut reproduire les distorsions présentes dans les données spatiales par suite de toutes les sources d'incertitude. La variation entre les répétitions pourrait être interprétée comme la variation entre diverses représentations, dans des bases de données spatiales, de la même réalité géographique en raison d'une variation entre les observateurs, les interprètes, les instruments de mesure ou les numériseurs. Le plus simple de tels modèles des erreurs serait l'équivalent cartographique de la distribution gaussienne pour les mesures scalaires - la distribution que l'on supposerait si l'on ne connaissait rien à propos de la nature exacte du processus d'erreur, autre que cette erreur est la combinaison additive d'un nombre très grand de contributions indépendantes.

Dans un certain nombre d'universités, on réalise actuellement de la recherche en modélisation des erreurs. Toutes ces universités emploient la même approche fondamentale : l'incertitude dans la base de données peut être décrite par un modèle des erreurs et des paramètres qui y sont associés; il est souhaitable de connaître les effets de cette incertitude sous forme de limites de confiance sur les produits obtenus à la suite du traitement de la base de données et la mathématique de l'analyse des erreurs est complexe et devrait donc, dans une large mesure, être dissimulée à l'utilisateur. À Newcastle, Openshaw et son équipe (Carver 1991) travaillent avec un modèle de bande des erreurs simple qui fournit des limites à la déviation spatiale de chaque ligne dans la base de données. À Utrecht, Burrough et son équipe (Heuvelink, Burrough et Stein 1989) utilisent la simulation de Monte-Carlo et des développements en série de Taylor pour propager l'erreur dans le traitement de trames de données continues. À Santa Barbara (Goodchild, Sun et Yang 1992), nous avons utilisé le système GRASS afin de simuler la propagation des erreurs dans des opérations sur des cartes de la couverture végétale multinomiales discrètes, basées sur un modèle spatialement autorégressif de l'incertitude.

5. PROBLÈMES RELATIFS À LA MISE EN APPLICATION

Si les SIG offrent la possibilité d'utiliser de façon plus efficace les statistiques spatiales dans l'analyse de données spatiales, alors de quels progrès avons-nous besoin pour que cela se produise? Jusqu'ici les progrès ont été limités et on trouve encore très peu de documents décrivant des exemples de recherche efficace en statistique spatiale qui emploient des SIG. En même temps, nous disposons d'une gamme étendue de logiciels statistiques qui possèdent très peu de fonctions spatiales.

L'industrie des SIG est considérable et florissante - les estimations de son importance vont de plusieurs centaines de millions à plusieurs milliards de dollars. Mais son marché se retrouve surtout dans des domaines qui exigent très peu d'analyse - la gestion, par les administrations locales, par les organismes et par les sociétés de services publics et d'exploitation de richesses naturelles, de vastes ensembles d'installations réparties dans des régions géographiques. Ainsi, les progrès récents dans le logiciel des SIG tendent à avoir augmenté beaucoup plus leurs possibilités comme systèmes de bases de données que leur puissance analytique.

Il semble exister trois modèles pour l'intégration des SIG et des techniques d'analyse statistique. Le premier, que l'on appellera ici «intégration complète», exigerait l'addition de fonctions analytiques aux SIG eux-mêmes, il semble peu probable que cela se produise sans une réorientation importante des priorités du marché. Le deuxième, que nous appellerons ici «couplage relâché» est ce dont nous disposons actuellement. Les données provenant des SIG sont communiquées à des programmes analytiques, soit des progiciels, soit des modules autonomes par l'intermédiaire de tables à deux dimensions d'enregistrements en caractères ASCII. Dans ce mode, toutes les structures de niveau plus élevé des données spatiales - les rapports hiérarchiques, les contiguïtés et les géométries complexes - sont perdues et il est difficile (et peut-être impossible) de retransmettre au SIG les résultats de l'analyse pour cartographie ou stockage. Le troisième mode est appelé «couplage étroit» et il se produit quand le SIG et le progiciel statistique utilisent des modèles de données communs, préservant ainsi la structure complète des données spatiales. Malheureusement, aucun des progiciels statistiques que l'on peut se procurer facilement à l'heure actuelle ne reconnaît de structures plus complexes que la simple table à deux dimensions d'enregistrements. Par contre, les SIG tendent à représenter des structures spatiales complexes à l'aide de diverses mises en application du modèle de données relationnelles.

6. PRIORITÉS

Si l'on se fie à l'évolution actuelle des SIG, des statistiques spatiales et de l'ASE, l'avenir est loin d'être sombre. Nous verrons probablement apparaître des progiciels très intéressants, pour l'analyse des données sociales et économiques, pour la planification des transports, pour la gestion des installations de crise, pour l'épidémiologie et pour le choix des sites d'implantation, qui mettent en application un grand nombre des méthodes d'analyse dont nous disposons actuellement, de façon beaucoup plus efficace, par l'intermédiaire de la technologie des SIG.

Les techniques des multimédias et des hypermédias ouvrent des perspectives totalement nouvelles sur les données spatiales et invitent de nouvelles formes de visualisation et d'inférence qui dépassent de beaucoup la carte familière sur support papier. En même temps, l'intuition peut être très trompeuse et l'analyse statistique dans un contexte spatial est loin d'être simple et directe.

Compte tenu des possibilités, le fait de savoir où commencer et comment s'y prendre pose un problème réel. Les statistiques et l'analyse spatiales comprennent une gamme étendue de méthodes et on dispose de peu de principes d'organisation pour simplifier cet ensemble. Nous pouvons, dans une certaine mesure, faire la distinction entre l'analyse et la modélisation et entre les modèles exploratoires et corroboratifs, mais la distinction la plus efficace, dans la perspective des SIG, se trouve dans le modèle de données - l'ensemble d'objets qui est créé à partir du processus de discrétisation de la variation géographique continue.

Les statistiques spatiales reposent sur un nombre remarquablement faible de modèles de données, moins, en fait, que l'ensemble qui peut actuellement être utilisé par les SIG dont on dispose. Il se peut que le plus simple de ces modèles de données soit l'ensemble de points non différencié utilisé, par exemple, en épidémiologie et en analyse de modèles de points. Un autre de ces modèles est la table d'attributs plus la matrice de facteurs de proximité utilisés pour mesurer la dépendance spatiale et dans les modèles autorégressifs et à moyenne mobile. Les modèles basés sur des attributs pour deux ensembles d'objets plus un tableau rectangulaire d'interactions ou de facteurs de proximité utilisés, par exemple, pour modéliser les flux et les interactions spatiales constituent des modèles plus complexes. Un autre de ces modèles est l'ensemble d'échantillons de points tirés d'une distribution continue et un autre est la trame de cellules - ces deux ensembles sont utilisés couramment dans l'étude de processus physiques et environnementaux.

Les SIG ont connu du succès en partie parce qu'ils ont réussi à intégrer un bon nombre de ces modèles de données dans un cadre commun. Mais, alors que la diversité des modèles peut être un des principaux avantages des SIG, elle constitue un obstacle important à l'élaboration de statistiques spatiales et à leur intégration avec les SIG. La majorité des efforts d'intégration au niveau du logiciel, qui ont été couronnés de succès, sont basés sur l'existence d'un modèle de données commun simple et les progiciels statistiques ont été établis de cette façon autour de la table simple à deux dimensions. Il se pourrait que la principale contribution des SIG à la statistique spatiale soit le fait que, dans ces systèmes, on insiste sur une reconnaissance explicite des modèles de données comme leur principe d'organisation fondamental.

REMERCIEMENTS

Le National Center for Geographic Information and Analysis est appuyé par la National Science Foundation, subvention SES 88-10917.

BIBLIOGRAPHIE

Anselin, L. (1989). What is special about spatial data? Report 89-4, Santa Barbara, CA: *National Center for Geographic Information and Analysis*.

Anselin, L. (1990). SPACESTAT: a program for the statistical analysis of spatial data, Santa Barbara, CA: Département de géographie, Université de Californie.

Arbia, G. (1989). *Spatial Data Configuration in Statistical Analysis of Regional Economic and Related Problems*, Dordrecht: Kluwer.

Barber, G.M. (1988). *Elementary Statistics for Geographers*, New York: Guilford.

Bracken, I., et Martin, D. (1989). The generation of spatial population distributions from census centroid data. *Environment and Planning A*, 21, 537-543.

Burrough, P.A. (1986). *Principles of Geographical Information Systems for Land Resources Assessment*, Oxford: Clarendon.

Carver, S. (1991). Adding error handling functionality to the GIS toolkit. *Proceedings, EGIS 91*, Brussells 1, 187-194.

Casetti, E. (1972). Generating models by the expansion method: applications to geographical research. *Geographical Analysis*, 4, 1, 81-91.

Clark, W.A.V., et Hosking, P.L. (1986). *Statistical Methods for Geographers*, New York: Wiley.

Cliff, A.D., et Ord, J.K. (1981). *Spatial Processes: Models and Applications*, London: Pion.

Ding, Y., et Fotheringham, A.S. (1991). The integration of spatial analysis and GIS: The development of the Statcas module for ARC/INFO. Report 91-5, Santa Barbara, CA: National Center for Geographic Information and Analysis.

Getis, A., et Boots, B.N. (1978). *Models of Spatial Processes: An Approach to the Study of Point, Line and Area Patterns*, Cambridge: Cambridge University Press.

Ghosh, A., et Rushton, G. (1987). *Spatial Analysis and Location-Allocation Models*, New York: Van Nostrand Reinhold.

Goodchild, M.F., Sun, G., et Yang, S. (1992). Development and test of an error model for categorical data. *International Journal of Geographical Information Systems*, 6 (sous presse).

Griffith, D.A. (1988). Estimating spatial autoregressive model parameters with commercial statistical packages. *Geographical Analysis*, 20, 176-186.

Griffith, D.A., et Amrhein, C.G. (1991). Statistical analysis for geographers, Englewood Cliffs, N.J.: Prentice Hall.

Haslett, J., Bradley, R., Craig, P., Unwin, A., et autres (1991). Dynamic graphics for exploring spatial data with application to locating global and local anomalies. *American Statistician*, 45, 3, 234-242.

Heuvelink, G.B.M., Burrough, P.A., et Stein, A. (1989). Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*, 3, 303-322.

Isaaks, E.H., et Srivastava, R.M. (1989). *Applied Geostatistics*, Oxford: Oxford University Press.

Maguire, D.J., Goodchild, M.F., et Rhind, D.W. (1991). *Geographical Information Systems: Principles and Applications*. London: Longman Scientific and Technical.

Mandelbrot, B.B. (1967). How long is the coast of Britain? Statistical self similarity and fractional dimension. *Science*, 156, 636-638.

Monmonier, M. (1989). Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis*, 21 (1), 81-84.

Openshaw, S., Charlton, M., Wymer, C., et Craft, C. (1987). A Mark I Geographical Analysis Machine for the automated analysis for point data sets. *International Journal of Geographical Information Systems*, 1, 335-358.

Openshaw, S., et Taylor, P.J. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem, *Statistical Applications in the Spatial Sciences*, édité par N. Wrigley, London: Pion, 127-144.

Richardson, S. (1990). Some remarks on the testing of association between spatial process, *Spatial Statistics: Past, Present and Future*, édité par D. Griffith, Ann Arbor, Michigan: Image, 277-309.

Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351-357.

Weiss, M.J. (1988). *The Clustering of America*, New York: Harper and Row.

SESSION 1

Perspectives géographiques sur la modélisation des données

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

ESTIMATION COMPOSITE PAR ESPACE D'ÉTATS POUR LES PETITES RÉGIONS

A.C. Singh et H.J. Mantel[1]

RÉSUMÉ

Lorsqu'on effectue de l'estimation pour petites régions, il est courant de renforcer l'estimation en «empruntant» à d'autres petites régions, puisque les estimations directes tirées de l'enquête présentent souvent une très grande variabilité d'échantillonnage. On crée souvent un estimateur composite, qui est une combinaison linéaire de l'estimateur d'enquête et d'un estimateur synthétique ou d'un estimateur de régression. Dans des études récentes, on a tenté de renforcer l'estimation en utilisant le facteur temps dans le cas d'enquêtes à passages répétés; on a pour cela utilisé des modèles chronologiques pour les paramètres des petites régions et les erreurs d'enquête. Nous proposons un estimateur composite par espace d'états pour lequel le filtre de Kalman utilisé pour l'ajustement des séries chronologiques a été modifié de manière qu'il ne soit plus nécessaire de modéliser les erreurs d'enquête. Nous ajouterons plutôt un terme de covariance à l'équation de filtrage pour tenir compte de l'autocorrélation des erreurs d'enquête. Nous proposons la technique d'échantillonnage répété jackknife pour obtenir une estimation non paramétrique de ce terme de covariance. Nous analyserons ensuite l'application possible de ces techniques à l'enquête sur la population active du Canada.

MOTS CLÉS: Enquêtes à passages répétés; modèles structurels; lissage spatial et temporel; filtre de Kalman avec jackknife.

1. INTRODUCTION

Considérons une enquête périodique. Nous nous intéresserons aux totaux par domaine relatifs au plus récent passage de l'enquête. Nous supposerons que ces domaines sont des domaines mineurs non prévus au sens de Purcell et Kish (1980). Le problème de l'estimation pour petites régions vient de ce que l'échantillon obtenu pour un domaine donné peut être petit ou même inexistant. L'estimateur calculé à partir de l'échantillon est par conséquent peu fiable ou peut-être même non défini. Il existe plusieurs stratégies pour résoudre ce problème, mais l'idée principale en est toujours de compléter les données de l'échantillon par de l'information supplémentaire. Cet apport d'information peut s'effectuer au moyen d'un modèle explicite ou implicite qui nous permettrait d'utiliser une certaine quantité d'information corrélée. Il est aussi possible d'utiliser des données provenant d'autres petites régions ou d'autres passages de l'enquête, toujours à l'aide d'un modèle. Au prix d'un certain biais, ces méthodes conduisent généralement à une plus grande stabilité des estimations relatives aux petites régions.

Si l'on utilise des données provenant d'autres petites régions similaires à l'intérieur d'une région plus vaste, une des méthodes qu'il est possible d'utiliser est celle de l'estimation synthétique. Dans cette méthode, on ajuste un modèle de régression en tenant compte de l'ensemble des données et l'estimation relative à la petite région n'est rien d'autre que l'espérance du modèle estimé. Si le modèle n'est pas très bon, ce procédé peut créer un biais important pour les estimateurs relatifs aux petites régions. On tente d'équilibrer le biais que peuvent introduire les estimateurs synthétiques et l'instabilité des estimateurs d'enquête en utilisant des estimateurs composites, qui sont des combinaisons convexes d'estimateurs d'enquête et d'estimateurs synthétiques. Ils peuvent être évalués de façon systématique, à l'aide d'une extension du modèle, ou à la pièce. Si on utilise des informations se

[1] A.C. Singh et H.J. Mantel, Division des méthodes d'enquête sociale, 16-A, Édifice R.H. Coats, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

rapportant au même domaine, mais recueillies à d'autres occasions dans le passé, on peut utiliser des estimations de changement fondées sur des échantillons communs pour corriger les estimations obtenues lors de passages antérieurs de l'enquête; le nouvel estimateur est alors une combinaison convexe de cet estimateur corrigé et de l'estimateur d'enquête. Par récursivité, il est possible d'utiliser toutes les données relatives aux passages précédents de l'enquête. Il y a aussi d'autres variations, dans lesquelles on suppose que les paramètres (θ_{kt}, où k dénote la petite région et t le temps) sont aléatoires en k et en t dans un modèle de superpopulation. Pour estimer un θ_{kt} donné, on renforce l'estimation en «empruntant» à d'autres régions et à d'autres passages de l'enquête en supposant que les θ_{kt} sont liés selon k et selon t par un modèle linéaire mixte convenable. Un des avantages importants de cette méthode fondée sur le caractère aléatoire des θ_{kt} est qu'elle n'exige pas d'unités d'échantillonnage communes (ou superposées) dans le temps.

Dans le présent article, nous examinerons d'abord brièvement dans les sections 2 et 3 diverses méthodes d'estimation pour petites régions. Nous mettrons l'accent sur l'estimation composite. Nous utiliserons le terme «composite» dans un sens général qui signifie que l'estimateur est une combinaison linéaire convexe de l'estimateur sous étude et d'un estimateur synthétique fondé sur les données d'autres régions, d'autres passages de l'enquête, ou des deux à la fois. On utilisera le terme «lissage spatial» pour dénoter des corrections apportées à l'estimation sous étude à partir d'une modélisation des données provenant d'autres régions à la même période ou période de temps. Le terme «lissage temporel» sera utilisé pour les corrections fondées sur une modélisation des données de passages antérieurs de l'enquête et se rapportant à la même région. La section 2 étudie le cas où les θ_{kt} ne sont pas aléatoires, et la section 3 celui où ils le sont. Dans la section 3, nous examinerons aussi une stratégie de modélisation par espace d'états pour l'estimation relative à de petites régions. Cette méthode utilise un espace d'états comme cadre servant à relier entre elles des variables θ_{kt} aléatoires. Dans ce contexte, nous proposons dans la section 4 un estimateur appelé «estimateur composite par espace d'états». Celui-ci exige qu'on modifie le filtre de Kalman pour pouvoir l'appliquer au cas où les erreurs d'enquête sont corrélées à cause de la répétition des unités d'échantillonnage dans le temps. Nous proposons aussi d'effectuer cette modification en employant la technique d'échantillonnage répété jackknife pour estimer les termes de covariance pertinents dans l'équation de filtrage modifiée. La section 5 présente certaines perspectives d'avenir pour la recherche, y compris une application à l'Enquête sur la population active du Canada.

2. ESTIMATION POUR PETITES RÉGIONS (θ_{kt} non aléatoires)

2.1 Lissage spatial pour des θ_{kt} non aléatoires

Pour un temps donné t, on se sert d'un modèle de régression pour décrire les θ_{kt} selon k (k=1, ...,K), à l'intérieur d'une région plus vaste, en fonction de variables corrélées connues et de certains paramètres inconnus. On utilise les estimateurs directs pour petites régions (ex: estimateurs d'expansion ou estimateurs de stratification a posteriori) pour estimer les paramètres. L'estimateur synthétique, qui n'est que la fonction de régression évaluée aux valeurs connues de la variable corrélée et aux valeurs estimées des paramètres, peut comporter un biais important si le modèle n'a pas été choisi avec soin. Gonzalez (1973) est le texte de référence principal sur l'estimation synthétique.

En pratique, on pourrait préférer un estimateur composite, c'est-à-dire une combinaison linéaire d'estimateurs directs et synthétiques se rapportant à une petite région. L'objectif serait d'équilibrer le biais possible de l'estimateur synthétique et l'instabilité de l'estimateur direct. L'estimateur de Drew, Singh et Choudhry (1982), qui dépend de la taille de l'échantillon, est un exemple d'estimateur composite.

2.2 Lissage temporel pour des θ_{kt} non aléatoires

Pour une petite région fixe k, il se peut qu'on puisse lier entre eux les différents estimateurs directs se rapportant à la petite région en fonction du temps t. Dans le cas d'enquêtes dont les passages se chevauchent sans se recouvrir exactement, on peut corriger les estimations antérieures en établissant des estimations de changement fondées sur des unités d'échantillonnage communes. Cet estimateur constitue alors une information supplémentaire pour améliorer l'estimateur étudié θ_{kt} au temps T. Cette idée a été utilisée par Jessen (1942) et Patterson (1950), entre autres (voir le compte rendu de Binder et Hidiroglou 1988). Gurney et Daly ont mis au point une méthode multidimensionnelle générale qui consiste à utiliser le concept d'estimations

«élémentaires». Dans un modèle linéaire multivarié, si les valeurs de la matrice des covariances de l'erreur peuvent être déterminées, on peut appliquer cette matrice au vecteur $\{ \theta_{kt} : t=1, \ldots, T \}$ pour obtenir l'ELSBVM (estimateur linéaire sans biais à variance minimum). En pratique, la détermination et l'inversion de la matrice des covariances peuvent conduire à une instabilité des estimations résultantes. Les calculs peuvent être grandement simplifiés si on induit une structure d'autocorrélation particulière lorsqu'on suppose que les θ_{kt} sont aléatoires (voir section 3). Gurney et Daly (1965), à la suite de Hansen, Hurwitz et Madow (1953) ont aussi proposé un compromis à la méthode de l'ELSBVM, dans lequel on calcule de façon récursive un estimateur composite qui combine l'estimateur sous étude et l'estimateur évalué au temps t qui précède (corrigé pour tenir compte du changement).

2.3 Lissage spatial et temporel pour des θ_{kt} non aléatoires

On peut aussi utiliser la méthode multivariée générale de Gurney et Daly (1965) mentionnée ci-dessus lorsque k et t varient tous les deux. Cependant, il sera alors encore plus difficile de déterminer et d'inverser de façon fiable la matrice des covariances.

Purcell et Kish (1980) proposent un autre type de lissage spatio-temporel appelé SPREE (structure preserving estimation; estimation avec préservation de la structure). Ce type de lissage convient aux données de fréquence. On utilise des données historiques fiables (provenant d'un recensement, par exemple) pour construire un tableau multidimensionnel de chiffres absolus ou de proportions, dans lequel une des dimensions correspond à la petite région, une autre aux facteurs pour lesquels on désire avoir les chiffres courants se rapportant aux petites régions, et où les autres dimensions correspondent à des facteurs associés. C'est ce qu'on appelle la structure d'association. On utilise l'information courante pour obtenir des estimations fiables de certains des tableaux marginaux; c'est ce qui s'appelle la structure d'allocation. On réconcilie ensuite la structure d'association et la structure d'allocation par ajustement proportionnel itératif et on élimine les facteurs associés pour en arriver à un tableau à deux dimensions mettant en relation les petites régions et le facteur étudié.

3. ESTIMATION POUR PETITES RÉGIONS (θ_{kt} aléatoires)

3.1 Lissage spatial pour des θ_{kt} aléatoires

Le fait de considérer les θ_{kt} comme aléatoires a comme principal avantage de fournir un cadre qui permet d'en arriver à un compromis rationnel entre le biais possible des estimateurs synthétiques et l'instabilité des estimateurs directs. Fay et Herriot (1979), Särndal (1984), Battese, Harter et Fuller (1988), Särndal et Hidiroglou (1989), Pfeffermann et Barnard (1991), Datta et Gosh (1991) et Gosh et Rao (1991) comptent parmi les textes de référence les plus importants sur le sujet. Nous suivrons les lignes générales de Fay et Herriot mais, plutôt que d'adopter un point de vue bayesien, nous choisirons la méthode du meilleur prédicteur linéaire sans biais (MPLSB) pour les modèles linéaires mixtes. Nous supposerons que l'information corrélée, s'il y en a, se trouve au niveau des petites régions. Si on peut trouver cette information au niveau des unités d'échantillonnage, alors il se peut qu'on puisse élaborer des estimateurs plus efficaces que ceux que nous étudierons dans le présent article.

Pour un temps t donné, le système de Fay-Herriot est donné par les équations suivantes:

$$\underset{\sim}{\theta}_t = F_t \underset{\sim}{\beta}_t + \underset{\sim}{a}_t \ , \ \underset{\sim}{y}_t = \underset{\sim}{\theta}_t + \underset{\sim}{\varepsilon}_t \qquad (3.1a)$$

$$\text{c.-à-d.} \quad \underset{\sim}{y}_t = F_t \underset{\sim}{\beta}_t + \underset{\sim}{a}_t + \underset{\sim}{\varepsilon}_t \qquad (3.1b)$$

où les $\underset{\sim}{y}_t$ sont les estimateurs directs, $\underset{\sim}{a}_t \sim (\underset{\sim}{0}, W_t)$, $\underset{\sim}{\varepsilon}_t \sim (\underset{\sim}{0}, V_t)$ et où $\underset{\sim}{\varepsilon}_t$ et $\underset{\sim}{a}_t$ ne sont pas corrélés. Il découle alors de l'annexe que le MPLSB de $\underset{\sim}{\theta}_t$ est donné par:

$$\hat{\underset{\sim}{\theta}}_t = F_t \hat{\underset{\sim}{\beta}}_t + \hat{\underset{\sim}{a}}_t \qquad (3.2a)$$

où

$$\hat{\underset{\sim}{\beta}}_t = (F_t' U_t^{-1} F_t)^{-1} F_t' U_t^{-1} \underset{\sim}{y}_t \quad , U_t = W_t + V_t$$
$$\hat{\underset{\sim}{a}}_t = W_t U_t^{-1} (\underset{\sim}{y}_t - F_t \hat{\underset{\sim}{\beta}}_t) . \tag{3.2b}$$

Il faut noter que, dans 3.2a, $\hat{\underset{\sim}{\theta}}_t$ est aussi un estimateur composite, c'est-à-dire qu'il peut s'écrire comme combinaison convexe de l'estimateur direct $\underset{\sim}{y}_t$ et de l'estimateur synthétique $F_t \hat{\underset{\sim}{\beta}}_t$.

3.2 Lissage temporel pour des θ_{kt} aléatoires

Certains des principaux textes de référence sur cette question sont Blight et Scott (1973), Scott et Smith (1974), Jones (1980), Bell et Hillmer (1987), Binder et Dick (1989), Choudhry et Rao (1989), Pfeffermann et Burck (1990) et Pfeffermann (1991). Bien qu'un grand nombre des articles mentionnés ci-dessus n'abordent pas directement le problème de l'estimation pour petites régions, l'idée sous-jacente de ces articles est essentiellement la même: utiliser des séries chronologiques pour estimer des moyennes de population à partir des données d'enquêtes à passages répétés. Dans le présent article, nous considérons uniquement la modélisation structurelle par séries chronologiques et le cadre fourni par l'espace d'états associé. Pour une région k donnée, la modélisation de la série chronologique des estimateurs directs pour petites régions $\{y_{kt} : t = 1,\ldots,T\}$ n'est qu'un cas particulier de la modélisation de la série chronologique à plusieurs variables $\{\underset{\sim}{y}_t : t = 1,\ldots,T\}$ qui sera étudiée dans la prochaine sous-section. Nous ne donnerons donc pas de détails supplémentaires sur ce cas.

3.3 Lissage spatial et temporel pour des θ_{kt} aléatoires

Le lissage spatial et temporel dans le cas où les θ_{kt} sont aléatoires est le sujet principal que nous voulons étudier dans le présent article. Nous voudrions renforcer l'estimation à la fois selon k et selon t en utilisant des modèles structurels de séries chronologiques. Un article récent de Pfeffermann et Burck (1990) fournit une excellente formulation d'une classe générale de modèles pour l'estimation relative à de petites régions. Cette classe comprend le lissage spatial et temporel. Dans le présent article, nous examinons une sous-classe importante qui généralise l'estimation de Fay-Herriot.

Considérons le système de Fay-Herriot, donné par les équations (3.1) ci-dessus. Dans la modélisation structurelle par séries chronologiques, l'autocorrélation apparaît lorsqu'on fait varier $\underset{\sim}{\beta}_t$ et $\underset{\sim}{a}_t$ avec le temps. Si on pose $\underset{\sim}{\alpha}_t^T = (\underset{\sim}{\beta}_t^T \ \underset{\sim}{a}_t^T)$ et $H_t = (F_t \ I)$, le modèle général que nous voulons étudier a la forme suivante:

$$\text{équation de mesure:} \quad \underset{\sim}{y}_t = H_t \underset{\sim}{\alpha}_t + \underset{\sim}{\varepsilon}_t , \quad \underset{\sim}{\varepsilon}_t \sim (\underset{\sim}{0}, V_t) \tag{3.3a}$$

$$\text{équation de transition:} \quad \underset{\sim}{\alpha}_t = G_t \underset{\sim}{\alpha}_{t-1} + \underset{\sim}{\zeta}_t , \quad \underset{\sim}{\zeta}_t \sim (\underset{\sim}{0}, \Gamma_t) \qquad t = 1,\ldots,T \tag{3.3b}$$

où les $\underset{\sim}{\zeta}_t$ sont indépendants d'un temps t à un autre, ainsi que de tous les $\underset{\sim}{\varepsilon}_t$. Dans l'espace d'états tel qu'on le définit d'habitude, les $\underset{\sim}{\varepsilon}_t$ sont indépendants d'un temps t à un autre; toutefois, pour les enquêtes à passages répétés, il est plus réaliste de supposer qu'ils sont dépendants par rapport au temps. Dans les prochaines sous-sections, nous diviserons donc l'étude en un cas corrélé et un cas non corrélé. Nous supposerons aussi que les matrices H_t, G_t, V_t et Γ_t sont connues. En pratique, il est généralement nécessaire d'estimer V_t et Γ_t.

3.3.1 Erreurs d'enquête $\underset{\sim}{\varepsilon}_t$ non corrélées

Le cas où les $\underset{\sim}{\varepsilon}_t$ sont indépendants (c'est-à-dire le cas où la sélection des échantillons se fait de façon indépendante pour les différents passages de l'enquête) est traité dans Singh, Mantel et Thomas (1991). On y présente aussi un certain nombre de résultats empiriques. Lorsqu'il n'y a pas de corrélation entre les $\underset{\sim}{\varepsilon}_t$, les calculs liés au filtre de Kalman demeurent réalisables en pratique et on peut utiliser cette méthode pour calculer le MPLSB d'$\underset{\sim}{\alpha}_T$ à partir des données $\underset{\sim}{y}_1,\ldots,\underset{\sim}{y}_t$ (voir par exemple Duncan et Horn (1972)). Le filtre de Kalman est une méthode récursive qui combine de façon optimale les données $\underset{\sim}{y}_t$ avec le prédicteur $\hat{\underset{\sim}{\alpha}}_{t-1}$, où $\hat{\underset{\sim}{\alpha}}_{t-1}$ est le MPLSB de $\underset{\sim}{\alpha}_{t-1}$, calculé à partir de $\underset{\sim}{y}_1,\ldots,\underset{\sim}{y}_{t-1}$. De façon plus précise, on a:

$$\hat{\underset{\sim}{\alpha}}_t = \hat{\underset{\sim}{\alpha}}_{t|t-1} + P_{t|t-1}H_t'(V_t + H_tP_{t|t-1}H_t')^{-1}(\underset{\sim}{y}_t - H_t\hat{\underset{\sim}{\alpha}}_{t|t-1}) \qquad (3.4)$$

où $\hat{\underset{\sim}{\alpha}}_{t|t-1} = G_t\hat{\underset{\sim}{\alpha}}_{t-1}$ et $P_{t|t-1}$ est l'erreur quadratique moyenne (EQM) de $\hat{\underset{\sim}{\alpha}}_{t|t-1}$ considéré comme estimation de $\underset{\sim}{\alpha}_t$. Des équations correspondantes permettent la mise à jour de P_t, qui est l'EQM de $\hat{\underset{\sim}{\alpha}}_t$ considéré comme estimation de $\underset{\sim}{\alpha}_t$, et la mise à jour de $P_{t|t-1}$. À partir de $\hat{\underset{\sim}{\alpha}}_t$, il est facile d'obtenir le MPLSB des paramètres $\underset{\sim}{\theta}_t$ se rapportant aux petites régions. Notons que, dans l'équation (3.4), $\hat{\underset{\sim}{\alpha}}_t$ est également un estimateur composite qui pourrait s'écrire comme combinaison convexe de l'estimateur direct $\underset{\sim}{y}_t$ et de $H_t\,\hat{\underset{\sim}{\alpha}}_{t|t-1}$.

L'estimation des paramètres de la matrice des covariances Γ_t peut se faire par la méthode du maximum de vraisemblance, si on suppose que les termes d'erreur sont distribués normalement. On peut se référer à Pfeffermann et Burck (1991), par exemple, pour l'utilisation de l'algorithme de Newton-Raphson dans le but d'obtenir les estimations de vraisemblance maximum. On peut aussi utiliser l'algorithme EM (voir par exemple Shumway et Stoffer (1982)). Toutefois, lorsque les G_t sont connus, on peut aussi employer la méthode des moments, qui ne nécessite aucune hypothèse de normalité des termes d'erreur. Cette méthode ressemble à celle utilisée dans Fay et Herriot (1979) et est utilisée dans Singh, Mantel et Thomas (1991). Lorsqu'on remplace les paramètres du modèle de calcul du MPLSB par des valeurs estimées, le MPLSB devient le MPLSBE (meilleur prédicteur linéaire sans biais empirique). L'EQM du MPLSBE est un peu plus grande que celle du MPLSB et il est possible, sous certaines conditions, de corriger en conséquence l'erreur quadratique estimée, selon la méthode indiquée par Prasad et Rao (1990).

3.3.2 Erreurs d'enquête $\underset{\sim}{e}_t$ corrélées

Le cas où les erreurs d'enquête sont corrélées est le plus réaliste. Une des façons de l'aborder est de supposer qu'on peut représenter les erreurs d'enquête par un modèle ARMM (voir par exemple Binder et Dick (1989), Pfeffermann et Burck (1990) et Pfeffermann (1991)). On augmente de façon convenable le vecteur d'états de manière que les nouvelles erreurs de mesure deviennent non corrélées. On peut maintenant appliquer le filtre de Kalman de la façon habituelle pour obtenir les MPLSB requis. On peut appeler «approche paramétrique» cette façon de traiter le cas où les erreurs d'enquête sont corrélées. Il serait commode de disposer d'une méthode non paramétrique pour les cas où il est difficile de formuler un modèle qui représente correctement les $\underset{\sim}{e}_t$. Cette méthode pourrait aussi servir à vérifier la validité des estimations obtenues à l'aide de l'approche paramétrique. Dans la section suivante, nous proposons justement une méthode de ce genre, fondée sur un croisement du filtre de Kalman et de la technique du jackknife.

4. ESTIMATION COMPOSITE D'ESPACE D'ÉTATS PAR LE FILTRE DE KALMAN AVEC JACKKNIFE

Nous décrirons maintenant une méthode d'estimation composite dans les cas où il y a corrélation des erreurs d'enquête. Cette méthode ne comporte pas de représentation des erreurs d'enquête par des séries chronologiques. Lorsque les $\underset{\sim}{e}_t$ sont corrélés dans le temps, on peut modifier l'équation de mise à jour du filtre de Kalman (3.4) pour qu'elle devienne:

$$\hat{\underset{\sim}{\alpha}}_t = \hat{\underset{\sim}{\alpha}}_{t|t-1} + (P_{t|t-1}H_t' - C_t)(V_t + H_tP_{t|t-1}H_t' - H_tC_t - C_t'H_t')^{-1}(\underset{\sim}{y}_t - H_t\hat{\underset{\sim}{\alpha}}_{t|t-1}) \qquad (4.1)$$

où C_t est la covariance de $\hat{\underset{\sim}{\alpha}}_{t|t-1} - \underset{\sim}{\alpha}_t$ et de $\underset{\sim}{e}_t$. Ici, l'estimateur composite $\hat{\underset{\sim}{\alpha}}_t$ est la combinaison optimale de $\underset{\sim}{y}_t$ et de $\hat{\underset{\sim}{\alpha}}_{t|t-1}$. Le $\hat{\underset{\sim}{\alpha}}_t$ obtenu de cette façon n'est pas en général le MPLSB de $\underset{\sim}{\alpha}_t$. Cependant, il tient tout de même compte de toutes les données, a l'avantage pratique de pouvoir être évalué par des calculs réalisables et devrait constituer un estimateur passablement efficace de $\underset{\sim}{\alpha}_T$. Ce type d'estimation, qu'on pourrait appeler «estimation composite par espace d'états», n'est pas sans rappeler le type d'estimation composite introduit par Hansen, Hurwitz et Madow (1953) et que les études empiriques de Gurney et Daly (1965) et Wolter (1979) ont révélé être passablement efficace.

La méthode proposée pour le cas où les erreurs d'enquête sont corrélées est fondée sur l'observation qu'on peut estimer sans biais les covariances qui nous intéressent en employant des méthodes qui dépendent uniquement du plan d'enquête. Pour voir cela, définissons $\underset{\sim}{\alpha}_t^* = (\underset{\sim}{\alpha}_0, \ldots, \underset{\sim}{\alpha}_t)$ et remarquons que:

$$
\begin{aligned}
& E[(\hat{\underset{\sim}{\alpha}}_{t|t-1} - \underset{\sim}{\alpha}_t)(\underset{\sim}{y}_t - H_t \underset{\sim}{\alpha}_t)] \\
& \quad = E_{\underset{\sim}{\alpha}_t^*} E\left[\left(\hat{\underset{\sim}{\alpha}}_{t|t-1} - E(\hat{\underset{\sim}{\alpha}}_{t|t-1} \mid \underset{\sim}{\alpha}_t) \right) \left(\underset{\sim}{y}_t - H_t \underset{\sim}{\alpha}_t \right) \mid \underset{\sim}{\alpha}_t^* \right] \\
& \quad + E_{\underset{\sim}{\alpha}_t^*} E\left[\left(E(\hat{\underset{\sim}{\alpha}}_{t|t-1} \mid \underset{\sim}{\alpha}_t) - \underset{\sim}{\alpha}_t \right) \left(\underset{\sim}{y}_t - H_t \underset{\sim}{\alpha}_t \right) \mid \underset{\sim}{\alpha}_t^* \right].
\end{aligned}
\tag{4.2}
$$

Le deuxième terme de l'équation ci-dessus est identiquement égal à 0 puisque $E\left[\underset{\sim}{y}_t - H_t \underset{\sim}{\alpha}_t \mid \underset{\sim}{\alpha}_t^* \right] = 0$ et que le premier terme est constant pour $\underset{\sim}{\alpha}_t^*$ donné. Il suffit donc d'estimer $\mathrm{Cov}(\hat{\underset{\sim}{\alpha}}_{t|t-1}, \underset{\sim}{y}_t \mid \underset{\sim}{\alpha}_t^*)$, ce qui peut se faire à l'aide d'une méthode d'échantillonnage répété comme la technique du jackknife, convenablement redéfinie pour les enquêtes complexes. Pour des enquêtes à plusieurs degrés d'échantillonnage, on applique généralement la technique du jackknife en retirant de l'échantillon les unités primaires d'échantillonnage (UPÉ) et en observant les effets obtenus sur les estimations. Dans le contexte d'une enquête à passages répétés, retirer une UPÉ de l'échantillon signifie ne pas tenir compte de cette unité ou de l'UPÉ correspondante à chacun des passages de l'enquête. Cela exige, notamment, un appariement ou une identification des UPÉ dans le temps. Par exemple, dans une enquête par panel avec renouvellement, on peut apparier les UPÉ qui se trouvent dans l'échantillon à la fois au temps t et au temps $t-1$. Cependant, lorsque le moment viendra pour une UPÉ de sortir de l'échantillon, il faudra l'apparier avec une UPÉ de remplacement. Cela implique aussi que le nombre d'UPÉ doit être le même à tous les passages de l'enquête. Les mêmes restrictions d'ordre pratique s'appliqueraient également à d'autres méthodes d'échantillonnage répété.

Le procédé itératif correspondant au filtre de Kalman avec jackknife (FKJ) consiste donc à calculer, pour chaque UPÉ, quelle influence le fait de retirer cette UPÉ de l'échantillon a sur $\underset{\sim}{y}_t$. On combine ensuite cette information avec les calculs déjà effectués sur l'influence du retrait de chaque UPÉ sur $\hat{\underset{\sim}{\alpha}}_{t|t-1}$ pour estimer la matrice C_t à l'aide de la technique du jackknife et pour en déduire l'effet du retrait de chaque unité d'échantillonnage sur $\hat{\underset{\sim}{\alpha}}_{t+1|t}$.

On peut estimer les paramètres de variance du FKJ, comme l'ont fait Singh, Mantel et Thomas (1991), par la méthode des moments appliquée aux G_t connus. Cependant, cette démarche exige de l'information supplémentaire sur $\mathrm{Cov}(\underset{\sim}{\varepsilon}_t, \underset{\sim}{\varepsilon}_{t-1})$. Cette information peut elle aussi être estimée par la technique du jackknife puisqu'on suppose que les UPÉ sont liées dans le temps. Cependant, il est possible qu'on puisse utiliser une estimation de $\mathrm{Cov}(\underset{\sim}{\varepsilon}_t, \underset{\sim}{\varepsilon}_{t-1})$ qui soit fondée sur le plan de sondage de l'enquête elle-même.

5. TRAVAUX À VENIR

Le filtre de Kalman avec jackknife, proposé ici, sera étudié dans un contexte réaliste à l'aide d'une étude de simulation, afin de le comparer à d'autres méthodes décrites dans le présent article, et en particulier à l'approche par modélisation.

Un bon exemple d'application possible du filtre de Kalman avec jackknife est l'enquête sur la population active du Canada (EPA). Le plan de sondage de cette enquête est décrit dans Statistique Canada (1990). En résumé, on divise chaque province en régions économiques. Chaque région économique est couverte par trois bases: une base autoreprésentative, une base non autoreprésentative et une base de secteurs spéciaux, qui comprend régions éloignées, établissements institutionnels et bases militaires.

La base autoreprésentative comprend les régions urbaines assez grandes pour que la taille espérée de l'échantillon soit d'au moins cinquante ménages. Chacune de ces régions urbaines ferait partie de l'échantillon. Ces régions se subdivisent à leur tour en bases régulières et bases d'appartements. Pour une base régulière, les unités primaires d'échantillonnage sont des grappes constituées de groupes d'îlots ou de côtés d'îlots contenant

un nombre convenable de logements. Pour une base d'appartements, les unités primaires d'échantillonnage sont des immeubles. À l'intérieur d'une UPÉ, les logements sont sélectionnés par échantillonnage systématique. Chaque mois, selon un plan de renouvellement, environ un sixième des logements sélectionnés sont retirés de l'échantillon et remplacés, de sorte qu'aucun logement ne pourra faire partie de l'échantillon plus de six mois de suite. Quand on a épuisé tous les logements d'une UPÉ, c'est l'UPÉ elle-même qui est retirée de l'échantillon et remplacée par une autre selon un procédé de renouvellement.

On subdivise la base non autoreprésentative en strates, qui sont classées rurales, urbaines ou mixtes. Dans ces strates, on définit des UPÉ selon la contiguïté géographique et selon d'autres facteurs. Chaque UPÉ sélectionnée est à nouveau subdivisée en grappes ou groupes à partir desquels on effectue un échantillonnage systématique de logements. On emploie aussi un plan de renouvellement semblable à celui utilisé pour les régions autoreprésentatives.

Au cours d'une application de la technique du jackknife à l'EPA, nous supprimerions les unités d'échantillonnage du premier degré. Le nombre d'unités d'échantillonnage de premier degré à l'intérieur d'une strate donnée est généralement constant dans le temps et, dans la plupart des cas, lorsqu'une unité disparaît de l'échantillon à cause du processus de renouvellement, il serait possible de l'apparier à une unité de remplacement bien définie. Il faudrait utiliser des méthodes spéciales dans les cas où cette affirmation se révélerait fausse. Des cas de ce genre pourraient se produire lorsqu'il y a redéfinition de l'enquête, ce qui arrive tous les dix ans environ, et après une correction de la taille de l'échantillon.

Pfeffermann et Bleuer (1992) ont déjà beaucoup travaillé à l'élaboration d'un type particulier de modèle d'espace d'états pour les données obtenues à partir de l'enquête par panel avec renouvellement qu'est l'EPA. Il serait utile de disposer d'une comparaison entre les deux approches (modélisation et technique du jackknife) dans ce contexte précis. On pourrait procéder à une étude de simulation qui utiliserait les données de l'EPA. Il faudrait d'abord construire une pseudo-population longitudinale dans laquelle les individus et les ménages demeureraient pendant une longue période. Puis on simulerait l'échantillonnage de l'EPA dans cette pseudo-population. Pour des raisons de simplicité et de faisabilité, l'étude de simulation pourrait se limiter à une province et exclure les secteurs spéciaux.

Des travaux additionnels pourraient aussi être consacrés au perfectionnement du filtre de Kalman avec jackknife. Pfeffermann et Burck (1990) ont pensé appliquer des contraintes aux estimations pour petites régions de manière que certains totaux de référence donnés soient égaux aux estimations directes des chiffres de population correspondants. Des contraintes de ce genre aident à rendre la méthode plus robuste en la protégeant contre une défaillance possible du modèle tout en obligeant les estimations pour petites régions à être cohérentes avec les estimations publiées se rapportant à des régions plus vastes. Des modifications de ce genre seraient aussi facilement applicables au filtre de Kalman avec jackknife.

Les méthodes que nous avons décrites utilisent de l'information supplémentaire au niveau des domaines, pour autant qu'elle soit disponible. Si par contre l'information supplémentaire est disponible au niveau des unités, il peut y avoir une perte d'information considérable lorsque celle-ci est agrégée au niveau des domaines. Il serait donc intéressant de mettre au point des méthodes d'estimation pour petites régions qui feraient usage de l'information supplémentaire au niveau des unités. Une autre question liée aux données supplémentaires est l'explication des erreurs de mesure possibles. Cette question deviendrait pertinente si, par exemple, l'information corrélée provenait d'une liste et que cette liste ne s'appliquait plus aujourd'hui ou était de qualité douteuse.

REMERCIEMENTS

Nous voudrions remercier Danny Pfeffermann et Jon Rao pour les conversations utiles qu'ils ont eues avec nous. La recherche du premier auteur du présent article a été rendue possible en partie par une subvention du Conseil de recherches en sciences naturelles et ingénierie du Canada accordée à l'Université Carleton d'Ottawa. Nous tenons aussi à remercier Christine Larabie pour son aide dans le traitement du manuscrit.

ANNEXE: lissage spatial pour des $\underset{\sim}{\theta}_i$ aléatoires

Considérons le modèle à effets aléatoires:

$$\underset{\sim}{y}_{k\times 1} = \underset{\sim}{\theta}_{k\times 1} + \underset{\sim}{\varepsilon}_{k\times 1}\ , \quad \underset{\sim}{\varepsilon} \sim (\underset{\sim}{0}, V) \qquad\qquad \underset{\sim}{\theta}_{k\times 1} = X_{k\times r}\underset{\sim}{\beta}_{r\times 1} + \underset{\sim}{a}_{k\times 1}\ , \quad \underset{\sim}{a} \sim (\underset{\sim}{0}, W)$$

où $\underset{\sim}{\varepsilon}$ et $\underset{\sim}{a}$ sont indépendants, $\text{rang}(X) = r$ et $r<k$.

Lemme 1 (Rao 1973, p. 234, Pfeffermann 1984). Si la moyenne de $\underset{\sim}{a}$ est inconnue, alors le MPLSB de $\underset{\sim}{\theta}$ est aussi le MELSB (meilleur estimateur linéaire sans biais) de $\underset{\sim}{\theta}$, lorsqu'on considère celui-ci comme constant. De plus, l'EQM est aussi la même.

Supposons maintenant que $\underset{\sim}{a}$ est de moyenne $\underset{\sim}{0}$. Alors la moyenne de $\underset{\sim}{\theta}$ se trouve dans un sous-espace à r dimensions de $\mathbf{R}^k$. Le MPLSB de $\underset{\sim}{\theta}$ sera différent du MELSB de $\underset{\sim}{\theta}$ lorsqu'on considère $\underset{\sim}{\theta}$ comme constant. Il s'ensuit selon Rao (1973, p.267), comme aussi selon Harville (1976), qu'on obtient le MPLSB de $\underset{\sim}{a}$ à partir de $\underset{\sim}{y}$ en traitant $E(\underset{\sim}{a} \mid \underset{\sim}{y})$ comme une fonction de régression linéaire et en substituant à tous les paramètres inconnus de la fonction linéaire leurs MELSB. On a donc:

Lemme 2 ($\underset{\sim}{\beta}$ connu). $\quad \underset{\sim}{\hat{a}} = \text{MPLSB de } \underset{\sim}{a} = E(\underset{\sim}{a}) + \text{Cov}(\underset{\sim}{a}, \underset{\sim}{y})\text{Var}(\underset{\sim}{y})^{-1}(\underset{\sim}{y} - E(\underset{\sim}{y}))$

$$= \underset{\sim}{0} + WU^{-1}(\underset{\sim}{y} - X\underset{\sim}{\beta})$$

où $U = V + W$, et

Lemme 3 ($\underset{\sim}{\beta}$ inconnu). $\quad \underset{\sim}{\hat{a}} = \text{MPLSB de } \underset{\sim}{a} = WU^{-1}(\underset{\sim}{y} - X\underset{\sim}{\hat{\beta}})$

où $\underset{\sim}{\hat{\beta}} = (X'U^{-1}X)^{-1}X'U^{-1}\underset{\sim}{y}$ est le MELSB de $\underset{\sim}{\beta}$.

BIBLIOGRAPHIE

Battese, G.E., Harter, R.M., et Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, *Journal of the American Statistical Association*, 83, 28-36.

Bell, W.R., et Hillmer, S.C. (1987). Time series methods for survey estimation, *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 83-92.

Binder, D.A., et Dick, J.P. (1989). Enquêtes répétées - Modélisation et estimation, *Techniques d'enquête*, 15, 1, 31-48.

Binder, D.A., et Hidiriglou, M.A. (1988). Sampling in time, *Handbook of Statistics*, 6, (Eds. P.R. Krishnaiah and C.R. Rao), Elsevier Science Publishers, 187-211.

Blight, B.J.N., et Scott, A.J. (1973). A stochastic model for repeated surveys, *Journal of the Royal Statistical Society*, Series B, 35, 61-68.

Choudhry, G.H., et Rao, J.N.K. (1989). Estimation de données régionales à l'aide de modèles qui combinent des séries chronologiques et des données transversales, *Recueil du Symposium de Statistique Canada sur l'analyse des données dans le temps*, (Eds. A.C. Singh et P. Whitridge), 71-80.

Datta, G.S., et Ghosh, M. (1991). Bayesian prediction in linear models: applications to small area estimation, *Annals of Statistics*, 19, 1748-1770.

Drew, J.D., Singh, M.P., et Choudhry, G.H. (1982). Évaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active du Canada, *Techniques d'enquête*, 8, 19-52.

Duncan, D.B., et Horn, S.D. (1972). Linear dynamic recursive estimation from the viewpoint of regression analysis, *Journal of the American Statistical Association*, 67, 815-821.

Fay, R.E., et Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association*, 74, 269-277.

Ghosh, M., et Rao, J.N.K. (1991). Small area estimation: an appraisal. Submitted for publication.

Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates, *Proceedings of the American Statistical Association, Social Statistics Section*, 33-36.

Gurney, M., et Daly, J.F. (1965). A multivariate approach to estimation in periodic sample surveys, *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 247-257.

Hansen, M.H., Hurwitz, W.N., et Madow, W.G. (1953), *Sample Survey Methods and Theory*, 2. New York: John Wiley & Sons.

Harville, D.A. (1976). Extension of the Gauss Markov theorem to include the estimation of random effects, *Annals of Statistics*, 4, 384-396.

Jessen, R.J. (1942). Statistical investigation of a farm survey for obtaining farm facts, *Iowa Agricultural Station Research Bulletin*, 304, 54-59.

Jones, R.G. (1980). Best linear unbiased estimators for repeated surveys, *Journal of the Royal Statistical Society*, Series B, 42, 221-226.

Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units, *Journal of the Royal Statistical Society*, Series B, 12, 241-255.

Pfeffermann, D. (1984). On extensions of the Gauss-Markov theorem to the case of stochastic regression coefficients, *Journal of the Royal Statistical Society*, Series B, 46, 139-148.

Pfeffermann, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys, *Journal of Business and Economic Statistics*, 9, 163-177.

Pfeffermann, D., et Barnard, C.H. (1991). Some new estimators for small area means with application to the assessment of farmland values, *Journal of Business and Economic Statistics*, 9, 73-84.

Pfeffermann, D. et Bleuer, S.R. (1992). Robust joint modelling of Canadian labour force series of small areas. Submitted for publication.

Pfeffermann, D., et Burck, L. (1990). Estimation robuste pour petits domaines par la combinaison de données chronologiques et transversales, *Techniques d'enquête*, 16, 2, 229-249.

Prasad, N.G.N., et Rao, J.N.K. (1990). The estimation of mean squared error of small-area estimators, *Journal of the American Statistical Association*, 85, 163-171.

Purcell, N.J., et Kish, L. (1980). Postcensal estimates for local areas (or domains), *International Statistical Review*, 48, 3-18.

Rao, C.R. (1973). *Linear Statistical Inference and its Applications, second edition*, New York: John Wiley.

Särndal, C.E. (1984). Design-consistent versus Model-dependent estimators for small domains, *Journal of the American Statistical Association*, 79, 624-631.

Särndal, C.E., et Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis, *Journal of the American Statistical Association*, 84, 255-275.

Scott, A.J., et Smith, T.M.F. (1974). Analysis of repeated surveys using time series methods, *Journal of the American Statistical Association*, 69, 674-678.

Shumway, R.H., et Stoffer, D.S. (1982). An approach to time series smoothing and forecasting using the EM algorithm, *Journal of Time Series Analysis*, 3, 253-264.

Singh, A.C., Mantel, H.J., et Thomas, B.W. (1991). Time series generalizations of Fay-Herriot estimator for small areas, *Proceedings of the American Statistical Association, Section on Survey Research Methods*, to appear.

Statistique Canada (1990). *Méthodologie de l'enquête sur la population active du Canada*, (auteurs: Singh, M.P., Drew, J.D., Gambino, J.G., et Mayda, F.) cat. n° 71-526.

Wolter, K.M. (1979). Composite estimation in finite populations, *Journal of the American Statistical Association*, 74, 604-613.

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

LES EFFETS DE GROUPEMENTS DIFFÉRENTS SUR LES ESTIMATIONS LOCALES DU SOUS-DÉNOMBREMENT

H. Hogan et C.T. Isaki[1]

RÉSUMÉ

L'Enquête post-censitaire (EP) de 1990 avait pour but de produire des chiffres du recensement redressés en fonction du sous-dénombrement ou du surdénombrement pour les États et les petites régions. L'échantillon prévu pour cette enquête était trop petit pour que l'on puisse produire des estimations directes pour de nombreux États et encore moins pour de plus petites régions. L'EP a donc été conçue de manière à produire des estimations synthétiques. On a calculé des estimations pour des strates formées a posteriori. La stratification a posteriori impliquait que l'on groupe des régions de divers États. On a ensuite lissé ces estimations au moyen d'un modèle statistique. On a constaté que les données servant à l'estimation et à la modélisation produisaient des estimations locales du sous-dénombrement passablement différentes selon la méthode utilisée pour grouper les régions.

MOTS CLÉS: Recensement; estimation pour petites régions; sélection de modèle; estimation synthétique.

1. INTRODUCTION

L'Enquête post-censitaire (EP) de 1990 avait pour but de produire des chiffres du recensement redressés en fonction du sous-dénombrement ou du surdénombrement pour les États et les petites régions. Elle consistait en deux parties. La première partie faisait intervenir un échantillon de la population, dit échantillon *P*. La proportion de l'échantillon *P* qui était incluse dans le recensement servait à estimer la proportion de la population totale qui avait été recensée. On détermine la proportion de l'échantillon qui a été recensée en appariant les membres de l'échantillon *P* aux enregistrements du recensement. Il faut également un échantillon des enregistrements du recensement pour estimer le nombre d'enregistrements erronés. Cet échantillon est désigné comme l'échantillon *D* et il forme l'autre partie de l'EP. On compare directement les enregistrements de l'échantillon *D* à ceux du recensement pour déterminer le degré de répétition. On réinterviewe aussi les membres de cet échantillon afin d'évaluer le nombre d'enregistrements fictifs, le nombre de personnes qui ont été recensées mais qui sont nées après le jour du recensement, *etc.* L'EP faisait intervenir le modèle de système dual pour estimer la population réelle. Se référer à Hogan (1991) pour une description de l'EP et de l'estimateur de système dual.

La série d'estimations initiales devait être produite pour le 17 mai 1991. Ces estimations ont permis de produire des chiffres du recensement redressés pour le 15 juillet 1991. Finalement, le Secretary of Commerce a décidé de ne pas corriger les chiffres du recensement à l'aide des résultats de l'EP. Depuis ce temps, nous explorons d'autres méthodes d'estimation du sous-dénombrement. Notre recherche consiste notamment à reprendre quelques-unes des opérations d'appariement et de codage initiales. Cette communication ne contient pas les résultats de ces recherches; elle traite plutôt un aspect de la question, à savoir l'effet de groupements géographiques différents sur les estimations du sous-dénombrement.

---

[1] H. Hogan et C. Isaki, Statistical Research Division, Bureau of the Census, Washington (DC) 20233. Les opinions exprimées dans cette communication sont celles des auteurs et ne reflètent pas nécessairement la position du Bureau of the Census.

Dans les deux sections qui suivent, nous décrivons le plan de sondage et la méthode d'estimation originaux de l'EP ainsi que les techniques de lissage préétablies. La section 4 expose deux autres méthodes de lissage et la section 5 compare les effets des diverses méthodes sur les estimations du sous-dénombrement pour les États, les comtés et les localités.

2. ÉCHANTILLONNAGE ET ESTIMATION

L'EP a été conçue pour produire des estimations synthétiques de la population de petites régions. La population est divisée en strates formées a posteriori. On calcule des estimations "post-censitaires" de la population réelle pour chacune de ces strates. Le rapport entre l'estimation "post-censitaire" de la population réelle et le chiffre du recensement pour chaque strate est appelé facteur de redressement et il sert de base à l'estimation synthétique.

Les strates sont formées de manière à ce que les éléments d'une strate aient à peu près les mêmes chances d'être recensés. Les variables de stratification utilisées dans l'EP étaient la division de recensement, le genre d'agglomération et la taille de l'agglomération, l'origine ethnique, l'âge, le sexe et le mode d'occupation (propriétaire/locataire). Voir le tableau 1. On a créé en plus douze strates selon l'âge et le sexe pour les Amérindiens qui vivaient dans les réserves. En faisant une classification combinée avec toutes les variables, on obtient un grand nombre de strates qui représentent chacune une très petite population. Dans ces circonstances, on fusionne de nombreuses cases. Par exemple, on ne tient pas compte du mode d'occupation à l'extérieur des villes centrales et il n'y a pas non plus de classe particulière pour les Noirs qui vivent dans les régions rurales de la Nouvelle-Angleterre. En somme, nous avons formé 116 strates "a posteriori" que nous avons croisées avec 12 groupes d'âge-sexe pour obtenir en tout 1 392 strates.

Tableau 1. Variables utilisées pour la stratification a posteriori

Origine raciale:	Noir, Non-noir d'origine hispanique, Asiatique et personne originaire des Îles du Pacifique, et tous les autres
Âge:	0-9, 10-19, 20-29, 30-44, 45-64, 65+
Sexe:	Homme, femme
Division de recensement:	Nouvelle-Angleterre, Atlantique centre, Atlantique sud, Centre sud-est, Centre sud-ouest, Centre nord-est, Centre nord-ouest, Rocheuses, Pacifique
Genre d'agglomération et taille:	* Ville centrale de zone statistique métropolitaine majeure (ZSMM) * Ville centrale de grande zone statistique métropolitaine (ZSM) (comptant au moins une ville de 250 000 habitants ou plus) * Ville centrale de petite ZSM * Ville non centrale de ZSMM ou de grande ZSM * Ville non centrale de petite ZSM * Municipalités de 10 000 habitants ou plus qui ne font pas partie d'une ZSM * Toutes les autres
Mode d'occupation:	Propriétaire ou locataire

L'unité primaire d'échantillonnage dans l'EP de 1990 était la grappe d'îlots, qui pouvait être composée d'un seul îlot ou d'un ensemble d'îlots. On a prélevé un échantillon d'environ 5 300 grappes pour l'enquête. Les mêmes îlots servaient pour l'échantillon ***P*** et l'échantillon ***D***. L'échantillon ***D*** comprenait tous les enregistrements du recensement, exacts comme erronés, des îlots échantillonnés. L'échantillon ***P*** comprenait toutes les personnes qui vivaient dans les unités de logement ou les logements collectifs non institutionnels des îlots échantillonnés au moment de l'interview de l'EP, soit environ 172 000 unités. Des estimations de système dual ont été calculées pour chacune des 1 392 strates formées a posteriori.

3. LISSAGE ET REDRESSEMENT JUSQU'AU NIVEAU DE L'ÎLOT

On prévoyait qu'un grand nombre des facteurs de redressement des 1 392 strates allaient avoir un coefficient de variation trop élevé pour les besoins des opérations de redressement. On a donc appliqué une méthode de régression dans le but de réduire la variabilité. Par cette méthode, on a pu prédire un facteur de redressement pour chaque strate. On a ensuite combiné le facteur prédit avec le facteur observé pour obtenir un facteur lissé. L'utilisation d'un modèle tentait d'"emprunter de l'information" à de nombreuses strates, un peu à la manière d'une méthode empirique de Bayes ou d'un modèle des composantes de la variance.

Soit

$$Y = X\beta + w + e,$$

où

Y	=	vecteur des facteurs de redressement observés pour chaque strate
X	=	matrice des variables explicatives
β	=	vecteur des paramètres de régression
w	=	erreur de modèle, distribuée par hypothèse selon une loi $N(0, \sigma^2 I)$
e	=	erreur d'échantillonnage, distribuée par hypothèse selon une loi $N(0, V)$, où V est la matrice des covariances de l'erreur d'échantillonnage.

On a supposé que les termes d'erreur, w et e, étaient indépendants et que V était connue. Les observations étaient les facteurs de redressement des 1 392 strates formées a posteriori. On a fait des ajustements de modèle distincts pour les quatre régions de recensement et pour les strates d'Amérindiens vivant dans des réserves.

Les variables ayant servi à constituer les strates -- de même que certaines mesures de difficulté du dénombrement -- ont été utilisées comme prédicteurs. Des mesures de difficulté ont été utilisées par les facteurs suivants: l'origine raciale (Noir, Asiatique), les éléments d'origine hispanique, le groupe d'âge, le mode d'occupation, la division de recensement ainsi que le genre d'agglomération et la taille de l'agglomération. On reconnaissait la possibilité d'interactions entre l'origine raciale et la taille de l'agglomération, entre l'âge, le sexe et l'origine raciale et entre l'âge, le sexe et le mode d'occupation. Ces variables étaient exprimées sous forme d'indicateur ou, si des classes étaient combinées, sous forme de proportion. Plusieurs variables servaient d'indice pour mesurer la difficulté de réalisation des opérations de recensement. La proportion de personnes qui avaient retourné leur questionnaire du recensement par la poste servait à mesurer le degré de coopération de la population. La proportion de substitutions de personnes dans le recensement servait à mesurer l'importance de l'imputation dans les opérations de recensement. Une autre variable indiquait la proportion des ménages qui avaient été recensés selon la formule classique du porte-à-porte, méthode principalement en usage dans les régions rurales éloignées.

Lorsqu'il s'est agi de déterminer les variables explicatives qui allaient constituer X, les indicateurs pour l'origine raciale, l'âge et le mode d'occupation, étaient toujours incluses dans le modèle tandis qu'on choisissait les autres variables en fonction de leur efficacité prédictive. Le choix des variables explicatives s'est fait au moyen de la régression relative au meilleur sous-ensemble (Furnival et Wilson 1974). On a préféré cette méthode à d'autres plus subjectives pour se conformer à la condition de spécification préalable. Pour tout sous-ensemble de variables prédictives, X, on a estimé β et σ^2 par une méthode itérative. Autrement dit, étant donné une valeur estimée de σ^2, nous pouvons calculer $\hat{\Sigma} = (V + \hat{\sigma}^2 I)$ et l'estimation par les moindres carrés généralisés:

$$\hat{\beta} = (X' \hat{\Sigma}^{-1} X)^{-1} (X' \hat{\Sigma}^{-1} Y).$$

Nous réestimons ensuite σ^2 par la méthode du maximum de vraisemblance. On doit répéter le processus jusqu'à ce qu'il y ait convergence des estimations. Les facteurs de redressement lissés, $\hat{y}$, sont calculés au moyen de la formule

$$\hat{y} = X\hat{\beta} + \hat{\sigma}^2 I \hat{\Sigma}^{-1} (Y - X\hat{\beta}).$$

Si V ne renfermait pas de covariances, l'opération ci-dessus équivaudrait à rajouter une partie du résidu à l'estimation de régression, laquelle partie serait proportionnelle à la variance de modèle et inversement proportionnelle à la variance d'échantillonnage. Mais comme V renferme des covariances, le facteur lissé peut

se trouver hors de l'intervalle formé par les valeurs du facteur de redressement observé et du facteur calculé par régression. Finalement, on a pondéré les facteurs lissés de manière que, pour chaque total de région estimé, le sous-dénombrement lissé égale le sous-dénombrement estimé directement (c.-à-d. en n'utilisant que Y).

Les résultats de tests antérieurs et des considérations théoriques laissaient supposer que les variances estimées de l'échantillon étaient plus élevées lorsque les facteurs de redressement estimés étaient élevés ou bien très faibles. Si les variances estimées à partir de l'échantillon avaient eu un lien uniquement avec les facteurs de redressement réels, l'ajustement par les moindres carrés généralisés et le lissage en auraient tenu compte convenablement. Or, il semblait plutôt exister un rapport entre les erreurs d'échantillonnage des variances estimées et celles des facteurs de redressement estimés. Par conséquent, il aurait pu y avoir sous-pondération ou surpondération de certains facteurs. Dans ces conditions, et aussi dans une tentative d'"emprunter de l'information" afin d'améliorer la variance estimée de l'erreur d'échantillonnage, nous avons lissé au préalable les variances à l'aide du modèle suivant:

$$n_i v_i/(1 + CV_i^2) = b_0 + b_1 W_i + b_2 AI_{1i} + b_3 AI_{2i} + b_4 Min_i,$$

où

v_i	=	variance vraie du facteur de redressement brut
n_i	=	nombre de membres de l'échantillon P inclus dans la strate i
CV_i	=	coefficient de variation des poids des individus de l'échantillon P
W_i	=	approximation brute du facteur de redressement obtenue par régression et dont la valeur doit être au moins égale à 1.00
AI_{1i}	=	indicateur d'âge pour les 0 - 19 ans
AI_{2i}	=	indicateur d'âge pour les 20 - 44 ans
Min_i	=	variable indiquant la proportion de membres d'une minorité dans la strate i.

Le terme "W_i" représente la corrélation entre la variance vraie et le facteur de redressement vrai. On le calcule en utilisant les mêmes variables prédictives et la même méthode de régression -- technique du meilleur sous-ensemble -- que celles employées pour le calcul des estimations finales, à la différence qu'il n'y a pas d'itération et que l'on se sert, évidemment, des variances estimées de l'échantillon.

On a ajusté le modèle de variance par région à l'aide des moindres carrés, les poids étant inversement proportionnels à la racine carrée de n_i. On attribuait une valeur nulle aux coefficients qui avaient une statistique t inférieure à deux, puis on ajustait à nouveau le modèle. Celui-ci semblait avoir une bonne efficacité prédictive en ce qui regarde les variances peu élevées. Cependant, dans le cas des variances élevées, il prédisait des valeurs beaucoup moindres. L'utilisation des variances de ce modèle dans l'analyse de régression des facteurs de redressement aurait eu pour effet de "surpondérer" les valeurs extrêmes. Afin d'atténuer le problème, on allait exclure du modèle tout point (ou toute valeur) dont le résidu studentisé était supérieur à quatre (4) et on allait utiliser la variance estimée de l'échantillon. La détection des valeurs aberrantes s'est faite en deux itérations et on a calculé les covariances en se servant des corrélations originales et des variances lissées au préalable. Pour une description plus détaillée du processus de lissage, voir Isaki et coll. (1991).

Enfin, on a réparti géographiquement les chiffres estimés du sous-dénombrement à l'intérieur de chaque strate en multipliant, pour chaque strate de chaque région, le facteur de redressement de la strate par le chiffre du recensement. Les chiffres du recensement relatifs aux groupes qui étaient exclus de la base de l'EP -- la population vivant en institution par exemple -- n'ont pas été modifiés.

4. AUTRES MÉTHODES DE LISSAGE

Par suite de l'ajustement des modèles régionaux aux données des 1 380 strates formées a posteriori (si l'on fait abstraction des 12 strates d'Amérindiens), nous avons voulu répéter l'expérience en nous servant d'un moins grand nombre de strates. De fait, nous avons conservé les 115 classes de strate initiales (Amérindiens exclus) mais nous avons réduit de douze à six le nombre de groupes d'âge-sexe:

- Hommes et femmes de 0 à 9 ans
- Hommes et femmes de 10 à 19 ans

- Hommes de 20 à 44 ans
- Femmes de 20 à 44 ans
- Hommes et femmes de 45 à 64 ans
- Hommes et femmes de 65 ans et plus.

La règle était que l'on fusionne les groupes qui avaient des niveaux de sous-dénombrement comparables.

L'utilisation d'un moins grand nombre de strates offre plusieurs avantages. En effet, un échantillon plus grand pour chaque strate permet de mieux estimer la matrice des variances-covariances. De plus, avec un moins grand nombre de strates, on peut faire un ajustement de modèle au niveau national et de ce fait, "emprunter de l'information" à diverses régions. Nous étions d'avis que ce deuxième avantage était particulièrement intéressant dans le cas de certains groupes minoritaires pour lesquels il y avait relativement peu d'observations (strates formées a posteriori) dans certaines des régions. Par exemple, les strates d'Asiatiques et de personnes originaires des Îles du Pacifique dans le Nord-est, les strates de personnes d'origine hispanique dans le Nord-est et le Midwest, et les strates de Noirs dans l'Ouest.

Le premier modèle que nous avons ajusté dans ces nouvelles conditions utilise les 690 strates formées a posteriori (115 x 6). On l'appelle le modèle national. Les variables explicatives qui entrent dans ce modèle sont sensiblement les mêmes que celles utilisées dans chaque modèle régional sauf que nous n'avons pas à définir de variables "obligatoires". De plus, nous avons créé plusieurs variables régionales explicites de manière à faire ressortir les différences qui étaient implicites dans l'ajustement des modèles régionaux.

Le processus d'élaboration du modèle national et des facteurs de redressement lissés correspondants est semblable à celui décrit précédemment pour les régions, sauf à deux points de vue. Primo, comme nous l'avons mentionné plus haut, les variables explicatives destinées à l'analyse de régression ont été choisies au moyen de la technique du meilleur sous-ensemble et *seule* l'ordonnée à l'origine devait, au départ, être incluse dans le modèle. Une comparaison de la variance (σ^2) calculée par la technique du meilleur sous-ensemble et de celle calculée au moyen de toutes les variables explicatives disponibles a révélé qu'il n'y avait pas de biais appréciable dans notre estimation de σ^2. Secundo, au lieu de rajuster les facteurs lissés en fonction d'un seul total pour les États-Unis, nous les avons rajustés en fonction de totaux pour les minorités et la majorité.

Nous avons tout d'abord lissé les variances d'échantillonnage estimées en nous servant du modèle décrit dans la section 3. Cette opération a permis de découvrir 10 valeurs aberrantes parmi les variances brutes. Les dix valeurs aberrantes avaient un résidu studentisé positif supérieur à 4.0. L'opération de détection des valeurs aberrantes était interrompue après deux itérations. Avec la technique du meilleur sous-ensemble, nous avons obtenu $\hat{\sigma}^2$ = .000285 et 14 variables explicatives ont été choisies en tout. Pour ce qui a trait au lissage des facteurs de redressement, les coefficients de rajustement pour les minorités et la majorité aux États-Unis étaient de 1.00692 et de 1.00017 respectivement. De plus, dans une comparaison avec un prédicteur, six facteurs de redressement bruts avaient une valeur qui était à plus de trois erreurs types de la valeur du prédicteur et ils ont été tronqués en conséquence. La variance estimée ($\hat{\sigma}^2$) et le nombre de variables explicatives utilisées dans la régression sont les deux points par lesquels le modèle national se distingue le plus des modèles régionaux. La valeur moyenne de $\hat{\sigma}^2$ pour les quatre modèles régionaux était environ .00055 et le nombre moyen de variables explicatives était 20. Avec des variances d'échantillonnage vraisemblablement moindres et les valeurs $\hat{\sigma}^2$ qui en découlent, le modèle national amène une diminution notable de l'erreur type des facteurs lissés.

Une autre méthode consistait à diviser les 115 classes de strate en deux groupes -- 66 strates de membres de la majorité et 49 strates de membres de minorités -- puis à ajuster des modèles différents à chaque groupe. Cette méthode suppose implicitement que le comportement d'une minorité ressemble plus à celui des autres minorités qu'à celui de la majorité. C'est ce qu'on appelle le modèle fractionné.

Nous avons procédé de la même manière que pour le modèle national en ajustant séparément les facteurs de redressement pour les minorités et la majorité. Nous avons rajusté les facteurs de redressement lissés en fonction des totaux respectifs pour les États-Unis calculés à l'aide des facteurs de redressement bruts. Lorsque nous avons appliqué la série combinée de facteurs de redressement lissés et que nous avons estimé la covariance de ces facteurs, nous avons pu poser l'hypothèse que les facteurs pour les minorités et la majorité étaient non corrélés. Cette hypothèse ne paraissait pas trop audacieuse car moins de 5% des corrélations brutes étaient

supérieures à .25 en valeur absolue. En ce qui a trait au modèle pour les minorités, le lissage des variances de l'échantillon a révélé cinq valeurs aberrantes. La technique du meilleur sous-ensemble a permis de choisir 11 variables explicatives pour le modèle de régression et a permis d'obtenir $\hat{\sigma}^2$ = .000558. La comparaison des facteurs de redressement bruts avec le prédicteur a donné quatre facteurs tronqués. En comparant les résultats du modèle national et ceux du modèle fractionné, nous avons constaté que les erreurs types des facteurs de redressement pour les minorités étaient moins élevées dans le second cas. Le coefficient de rajustement était de 1.00658, donc très comparable à celui calculé selon le modèle national.

En ce qui concerne le modèle pour la majorité, le lissage des variances a révélé six valeurs aberrantes. La technique du meilleur sous-ensemble a permis de choisir onze variables explicatives, avec une variance estimée ($\hat{\sigma}^2$) de .000118, soit environ la moitié de l'estimation calculée selon le modèle national. Deux facteurs de redressement bruts ont été tronqués. Une comparaison des résultats du modèle national avec ceux du modèle pour la majorité a permis de constater que les erreurs types des facteurs lissés étaient plus élevées dans le premier cas. Le coefficient de rajustement était de 1.00036 et, là aussi, très comparable à celui calculé selon le modèle national. La différence entre les estimations de variance ($\hat{\sigma}^2$) calculées selon l'une et l'autre portions du modèle fractionné donne à penser qu'il serait plus utile pour le modèle national d'utiliser une structure d'erreur qui suppose des variances distinctes pour chaque catégorie de facteurs de redressement (facteurs pour les minorités et facteurs pour la majorité). La question est présentement à l'étude.

Le tableau 2 donne les estimations du sous-dénombrement calculées selon les quatre modèles pour les 116 classes de strate: les estimations de division (non lissées), les estimations régionales (1 380), les estimations nationales (690) et les estimations du modèle fractionné (minorités/majorité). On peut se rendre compte des avantages des deux derniers modèles en examinant quelques-unes des classes de strate. Si on considère tout d'abord les groupes non minoritaires, l'estimation régionale pour les villes centrales de Nouvelle-Angleterre indique, étonnamment, un taux de *surdénombrement* de 1.16%. Avec les deux derniers modèles, on obtient pratiquement une erreur nette nulle. De même, pour les "autres régions" du Centre nord-est, le modèle régional indique un taux de surdénombrement de près de 1%. Avec le modèle national, on obtient une erreur nette nulle, tandis que le modèle pour la majorité prédit un taux de sous-dénombrement faible.

Les avantages relatifs des deux derniers modèles sont plus appréciables lorsqu'on considère les strates de membres de minorités. Quelle que soit la région, les strates de membres de minorités sont toujours moins nombreuses que les strates de membres de la majorité et l'estimation qui a trait aux premières est caractérisée par une variance relativement élevée. En outre, comme les strates de membres de minorités sont subdivisées en classes (ex.: Noirs, Non-noirs d'origine hispanique et, parfois, Asiatiques), la possibilité d'"emprunter de l'information" est souvent limitée, du moins en ce qui concerne les membres d'un même groupe ethnique.

Dans le Nord-est par exemple, il n'y avait que deux strates de personnes d'origine hispanique, plus trois strates qui comptaient à la fois des Noirs et des personnes d'origine hispanique. Les deux strates de personnes d'origine hispanique étaient celles des villes centrales de New York City et des autres grandes zones statistiques métropolitaines (ZSM). Les estimations directes (estimations de division) du sous-dénombrement pour ces deux strates étaient 4.0 et 9.91% respectivement. Les erreurs types correspondantes étaient élevées: 3.81 et 6.07 respectivement. Le lissage selon le modèle régional a ramené ces estimations à 1.73% et 2.01%, soit en deçà de la moyenne nationale. Le modèle national et le modèle pour les minorités ont pu "emprunter de l'information" à d'autres régions où vivent des personnes d'origine hispanique et ont produit des estimations de 5.5 et 5.2% (modèle national) et de 6.5 et 6.6% (modèle pour les minorités). Ces modèles ont permis de redresser des chiffres encore plus aberrants dans le Midwest. L'estimation "directe" du sous-dénombrement pour les personnes d'origine hispanique dans les grandes ZSM (y compris Chicago et Détroit) du Centre nord-est était de 0.38%. Le modèle régional a produit un taux de *surdénombrement* de 1.6%. Bien que de tels chiffres ne soient pas invraisemblables, les estimations calculées à l'aide du modèle national et du modèle pour les minorités (4.0 et 3.5% respectivement) semblent plus plausibles aux yeux de tous.

Prenons la région Ouest comme dernier exemple. L'estimation "directe" (estimation de division) du sous-dénombrement pour les Noirs dans les villes non centrales des ZSM était de 14.3%, soit l'un des taux les plus élevés. Le lissage selon le modèle régional a porté ce taux à 16.4%. Le modèle national et le modèle pour les minorités ont ramené ces estimations à un niveau plus raisonnable, soit à 6.6 et à 8.8% respectivement.

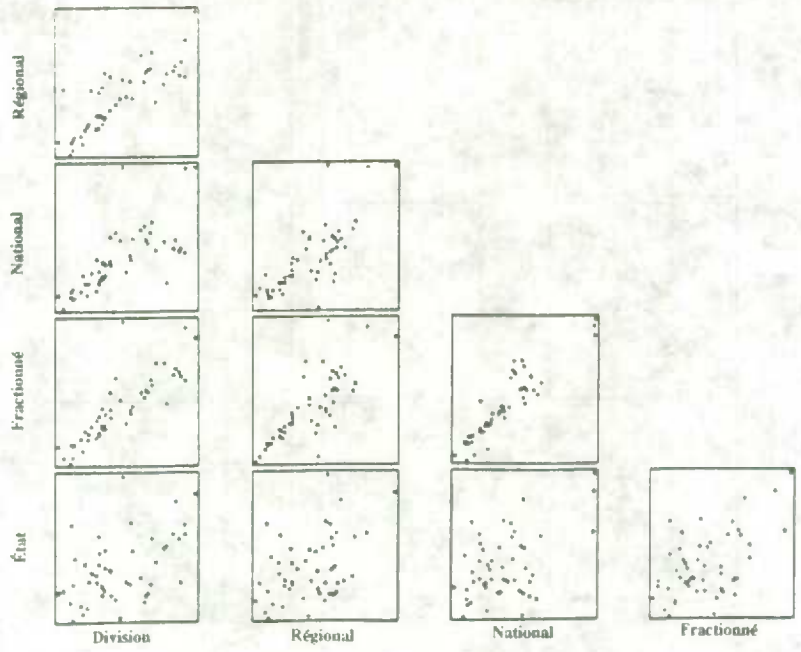
Cependant, bon nombre des observations aberrantes sont précisément dues au fait que les strates formées a posteriori représentent des groupes numériquement faibles. La question est de savoir si cela fait une grande différence lorsque les strates correspondent à des divisions administratives.

5. EFFET SUR LES ESTIMATIONS POUR L'ÉTAT, LE COMTÉ ET L'AGGLOMÉRATION

L'échantillon de l'EP a été réparti de façon à étayer la série "régionale" d'estimations synthétiques décrite dans la section 4. À l'origine, on voulait éviter de répartir l'échantillon afin d'obtenir des estimations directes d'État qui auraient une variance acceptable. Or, il est possible de produire des estimations directes pour un État en se servant uniquement de l'échantillon de cet État. C'est ce qu'on appelle le modèle de l'État. Les estimations calculées à l'aide de ce modèle sont caractérisées par une forte variance d'échantillonnage ce qui leur enlève toute utilité. Néanmoins, il est bon de comparer ces estimations à celles obtenues indirectement à l'aide des quatre séries de facteurs de redressement du tableau 2.

La figure 1 contient des représentations graphiques du taux de sous-dénombrement dans les États selon cinq modèles. Plusieurs constatations se dégagent de ces graphiques. Le modèle de l'État a peu de rapport avec les quatre autres modèles. Cela peut très bien être l'indice de différences de taux de sous-dénombrement entre des États qui ne sont pas touchés par le processus d'agrégation. Nous croyons que cette différence est plutôt attribuable à la variance d'échantillonnage des estimations directes. En effet, l'écart entre les estimations directes d'État et les estimations de division était à peine supérieur à l'erreur d'échantillonnage pour trois États: Montana, Idaho et Washington. Comme ces États forment, avec la Californie, une partie de la région Ouest, il est fort possible que les données reflètent des différences de taux de sous-dénombrement réel.

Figure 1: Taux de sous-dénombrement dans les États, selon cinq modèles



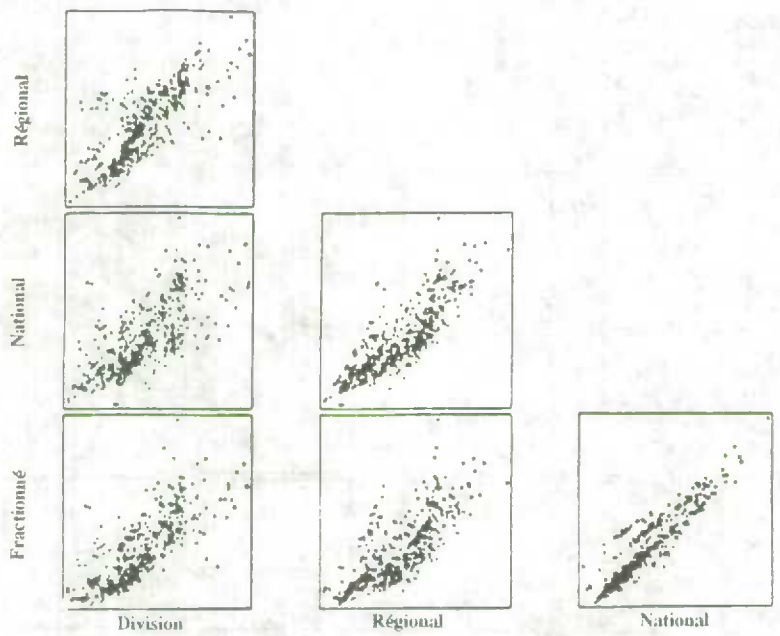
Le modèle de la division et le modèle régional ont plus de rapport avec les autres modèles. Malgré cela, beaucoup de points se trouvent hors de la diagonale. C'est dans la division de l'Atlantique sud qu'on trouvait les estimations non lissées les plus élevées. Les divisions Centre sud-est et Centre sud-ouest, pour leur part, avaient un taux de sous-dénombrement beaucoup moins élevé que les autres. Le lissage selon le modèle régional a eu pour effet de placer ces trois divisions sur le même pied, le taux pour le Centre sud-ouest étant désormais un peu plus élevé que les autres. Autrement dit, l'opération de lissage a réduit de 0.9 point le taux de sous-dénombrement pour Atlantique sud et a haussé de 1.2 et de 0.8 le taux pour les divisions Centre sud-est et Centre sud-ouest respectivement. L'estimation lissée était de 427 000 inférieure à l'estimation non lissée pour

Atlantique sud et de 191 000 supérieure pour Centre sud-est; en ce qui concerne la division Centre sud-ouest, l'estimation lissée dépassait de 222 000 l'estimation non lissée.

On retrouve les mêmes différences lorsqu'on compare le modèle national avec le modèle fractionné. De façon générale, les États se trouvent sur la diagonale. Cependant, on observe très bien au centre un nuage de points qui sont hors de la diagonale et l'estimation calculée à l'aide du modèle fractionné est visiblement plus élevée que celle du modèle national. Ces points sont en fait les huit États de la division Atlantique sud. D'ailleurs, l'indicateur de cette division a été inclus dans le modèle de régression pour les groupes majoritaires mais non dans le modèle régional pour le Sud ou le modèle national. Les points qui correspondent à une estimation du sous-dénombrement moins élevée dans le modèle fractionné que dans le modèle national ne forment pas une entité régionale particulière: le Connecticut, le Massachusetts et le Rhode Island font partie de ce groupe, mais l'Oregon et le Washington aussi. Les trois points qui correspondent à une estimation du sous-dénombrement élevée selon les deux modèles sont la Californie, Hawaii et le Nouveau-Mexique. Les deux estimations pour Hawaii sont presque équivalentes (3.57 et 3.59%). Ces résultats tranchent avec l'estimation calculée à l'aide du modèle régional (2.47%).

La figure 2 contient des représentations graphiques du taux de sous-dénombrement calculé selon les quatre modèles du tableau 2 pour les 458 comtés de plus de 10 000 habitants. Comme il n'y a pas d'estimations de comté (estimations directes), nous ne comparons que quatre modèles. De toute évidence, les résultats des quatre modèles sont fortement corrélés, mais il y a des exceptions notables. Ainsi, le modèle de la division ne s'accorde pas aussi bien avec les trois autres modèles que ceux-ci entre eux. Là encore, si on examine les points qui sont hors de la diagonale, on peut voir qu'un grand nombre des différences entre les modèles sont imputables aux comtés de la division Atlantique sud.

Figure 2: Taux de sous-dénombrement dans les comtés, selon quatre modèles

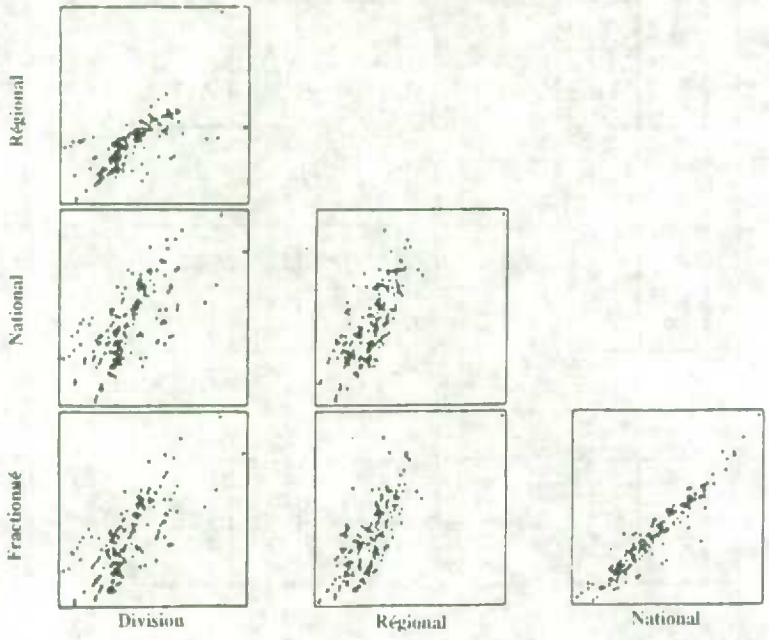


La figure 3 contient les représentations graphiques du taux de sous-dénombrement estimé pour les 195 agglomérations de plus de 10 000 habitants. La valeur aberrante extrême observée avec le modèle régional est la ville d'Inglewood, en Californie. Cette ville a un taux de sous-dénombrement estimé de 11.15%. Le modèle national et le modèle fractionné réduisent considérablement ce taux, le ramenant à 6.94 et à 6.05% respectivement.

Cette communication avait pour but d'étudier l'effet de groupements différents sur les estimations du sous-dénombrement. De façon générale, nous avons obtenu des résultats semblables. Il est vrai que nous avons

toujours utilisé la même série de données. Cependant, pour certains États, certains comtés et certaines agglomérations, les taux de sous-dénombrement sont très différents selon les modèles. Le modèle national et le modèle fractionné (avec ses deux volets: modèle pour les minorités et modèle pour la majorité) semblent tous deux présenter des avantages appréciables par rapport au modèle régional. Les deux font disparaître bon nombre des anomalies que l'on trouve dans le modèle régional. Il est toutefois plus difficile de dire lequel du modèle national ou du modèle fractionné est supérieur à l'autre. La capacité du modèle fractionné de détecter les différences entre divisions à l'intérieur d'un groupe ethnique, lorsqu'il y en a, semble être un avantage appréciable. Malgré cela, la supériorité de l'un des modèles par rapport à l'autre n'est pas encore démontrée.

Figure 3: Percent Undercount of Places by Four Models



6. REMERCIEMENTS

Nous tenons à remercier Elizabeth Huang et Julie Tsay, qui ont calculé les facteurs lissés dans tous les modèles utilisés, ainsi que Fred Navarro, qui a produit les estimations du sous-dénombrement.

BIBLIOGRAPHIE

Furnival, G.M., et Wilson, R.W. (1974). Regression by leaps and bounds, *Technometrics*, 16, 499-512.

Hogan, H. (1991). The 1990 Post-Enumeration Survey: An Overview, *Proceedings of the Section on Survey Methods Research of the American Statistical Association,* 518-523.

Isaki, C.T., Huang, E.T., et Tsay, J.H. (1991). Smoothing adjustment factors from the 1990 PES, paper presented to the Section on Survey Methods Research, American Statistical Association meeting, Atlanta, Georgia.

Isaki, C.T., Schultz, L.K., Diffendal, G.J., et Huang, E.T. (1988). On estimating census undercount in small areas, *Journal of Official Statistics,* 4:2, 2, 95-112.

Tableau 2: Estimations de système dual de l'Enquête post-censitaire de 1990
Taux de sous-dénombrement (%) par classe de strate

	Division	Régional	National	Fractionné	Division			Régional			National			Fractionné		
	Groupes non minoritaires				Noirs	Hisp.	Asiat.	Noirs	Hisp.	Asiat.	Noirs	Hisp.	Asiat.	Noirs	Hisp.	Asiat.
Nord Est																
Nouvelle-Angleterre																
Villes centrales	-1.74	-1.16	0.26	-0.07	5.69			4.25			5.76			3.95		
Villes non centrales de ZSM	0.61	0.19	0.80	0.57												
Autres agglomérations 10 000+	0.54	0.59	1.22	1.03	5.88 *			5.39 *			4.60 *			5.75 *		
Autres zones	1.68	1.79	1.79	2.00												
Atlantique centre																
Villes centrales de New York City																
Locataire	2.06	0.87	1.81	2.72	6.44	4.00	9.47	7.76	1.73	10.50	7.29	5.48	4.84	8.43	6.54	6.33
Propriétaire	-2.64	-0.23	-0.33	-0.76	-2.86			-0.15			1.07			2.27		
Villes centrales d'autres grandes ZSM																
Locataire	-6.41	-0.37	2.26	2.26	10.78	9.91		7.74	2.01		8.02	5.23		9.36	6.59	
Propriétaire	-2.93	-0.19	-0.45	-0.65	2.66			-0.03			1.68			3.12		
Villes centrales de petites ZSM	2.05	0.07	1.02	0.67	17.92			9.34			6.55			8.38		
Villes non centrales de la ZSMM de NYC	5.03	0.42	1.12	0.52	5.63			6.73			4.67			5.34		
Villes non centrales de grandes ZSM	-0.80	0.36	0.70	0.48												
Villes non centrales de petites ZSM	-0.78	-0.09	0.41	0.29	5.88 *			5.39 *			4.60 *			5.75 *		
Autres agglomérations 10 000+	1.36	0.41	1.39	0.81												
Autres zones	0.43	0.70	0.98	0.87												
Sud																
Atlantique sud																
Villes centrales de grandes ZSM																
Locataire	11.49	5.00	3.78	4.05	10.46			9.33			7.64			7.95		
Propriétaire	1.09	1.72	-0.17	0.84	1.68	2.77		0.95	4.92		1.55	3.87		2.13	3.90	
Villes centrales de petites ZSM	2.84	2.74	1.86	2.18	4.93			4.00			5.34			5.67		
Villes non centrales de grandes ZSM	0.93	0.44	0.52	1.83	4.17	13.79		1.97	5.13		4.83	3.55		5.30		
Villes non centrales de petites ZSM	3.50	2.80	1.38	1.97	0.27			3.59			4.14			4.52	3.69	
Autres agglomérations 10 000+	1.23	1.51	0.79	1.77	-1.71			1.60			3.06			3.66		
Autres zones	3.25	2.71	1.23	1.79	5.68			2.64			2.09			3.36		

* Indique que des cellules ont été combinées à l'intérieur de la division

	Division	Régional	National	Fractionné	Division			Régional			National			Split		
	Groupes non minoritaires				Noirs	Hisp.	Asiat.	Noirs	Hisp.	Asiat.	Noirs	Hisp.	Asiat.	Black	Hisp.	Asian
Centre sud-est																
Villes centrales de grandes ZSM																
Locataire	2.17	4.80	2.81	2.60	6.46			5.81			5.22			6.11		
Propriétaire	3.19	2.56	0.45	0.06												
Villes centrales de petites ZSM	0.90	2.58	0.80	0.62												
Villes non centrales de ZSM	1.42	2.31	1.11	0.83	4.82			2.26			2.89			3.85		
Autres agglomérations 10 000+	-6.02	1.84	1.01	0.64												
Autres zones	-0.95	1.65	0.09	0.16												
Centre sud-ouest																
Villes centrales de Houston, Dallas, Ft Worth																
Locataire	6.24	4.60	3.31	2.82												
Propriétaire	0.56	1.49	0.34	0.18	8.09	8.96		6.64	7.11		5.60	5.03		5.98	5.44	
Villes centrales d'autres grandes ZSM																
Locataire	1.34	3.23	3.06	2.76												
Propriétaire	-1.16	0.69	-0.10	0.04	4.54	3.18		4.82	3.76		4.86	3.27		5.31	3.70	
Villes centrales de petites ZSM	-3.16	2.48	1.06	0.83												
Villes non centrales de ZSM	2.07	2.28	1.21	0.99												
Autres agglomérations 10,000+	1.19	1.25	0.65	0.49	1.66	2.36		2.28	5.11		3.18	2.48		4.35	2.56	
Autres zones	1.72	1.96	1.09	0.97												
Midwest																
Centre nord-est																
Villes centrales de Chicago, Détroit																
Locataire	2.76	5.17	3.32	2.31	6.76	0.38		5.77	-1.61		7.64	4.04		7.02	3.52	
Propriétaire	-0.05	1.12	0.34	-0.25	0.42			1.98			1.79			2.16		
Villes centrales d'autres grandes ZSM																
Locataire	1.56	1.04	1.95	2.30	4.03			4.49			7.64			7.33		
Propriétaire	-1.24	-0.15	0.84	0.41	7.09			0.64			1.92			2.77		
Villes centrales de petites ZSM	1.76	2.09	1.35	0.71	4.61			5.44			4.59					
Villes non centrales de grandes ZSM	0.84	0.59	0.64	0.62										5.06		
Villes non centrales de petites ZSM	0.96	0.64	0.55	0.62	3.99 *			4.66 *			4.71 *			5.03 *		
Autres agglomérations 10,000+	0.42	0.20	0.78	0.62												
Autres zones	-1.64	-0.99	0.01	0.65												

	Division	Régional	Nation	Fractionné	Division			Régional			National			Fractionné		
	Groupes non minoritaires				Noirs	Hisp.	Asiat.	Noirs	Hisp.	Asiat.	Noirs	Hisp.	Asiat.	Noirs	Hisp.	Asiat.
Centre nord-ouest																
Villes centrales de grandes ZSM																
Locataire	5.20	2.47	3.05	2.83	5.47			5.44			5.27			5.45		
Propriétaire	-0.53	-0.33	-0.05	-0.26												
Villes centrales de petites ZSM	1.82	1.90	1.37	0.81	4.85			7.23			4.59			5.67		
Villes non centrales de grandes ZSM	1.09	0.71	0.92	0.85												
Villes non centrales de petites ZSM	0.22	1.64	0.75	0.95	3.99 *			4.66 *			4.71 *			5.03 *		
Autres agglomérations 10 000+	0.83	0.75	1.00	0.73												
Autres zones	0.78	0.31	1.05	1.45												
Ouest																
Rocheuses																
Villes centrales de grandes ZSM																
Locataire	4.65	5.03	3.69	2.86	1.48			4.61			4.08			4.29		
Propriétaire	1.24	0.98	0.23	0.04												
Villes centrales de petites ZSM	2.88	1.52	1.57	1.08												
Villes non centrales de ZSM	0.60	0.75	0.58	0.79	7.39 *			7.80 *			4.65 *			4.68 *		
Autres agglomérations 10,000+	1.22	1.45	0.99	0.81												
Autres zones	3.00	3.22	2.46	2.80												
Pacifique																
Locataire																
Villes centrales de Los Angeles, Long	6.44	4.75	4.05	3.09	7.38	10.14	6.29	6.83	7.87	6.50	8.26	7.95	7.44	9.46	8.81	7.80
Beach	3.73	3.72	4.31	3.62												
Villes centrales d'autres grandes ZSM																
Villes centrales de petites ZSM																
Propriétaire																
Villes centrales de Los Angeles, Long	-0.35	1.39	0.62	-0.16	8.36	2.01	3.10	7.86	1.95	4.80	3.26	1.71	1.71	5.33	2.80	3.12
Beach	1.39	1.39	0.89	0.40												
Villes centrales d'autres grandes ZSM																
Villes centrales de petites ZSM																
Villes centrales de petites ZSM	0.56	0.95	1.38	0.95												
Villes non centrales de grandes ZSM	1.05	0.17	1.25	1.06	14.32	5.65	0.82	16.37	6.94	0.79	6.62	5.38	3.31	8.77	6.35	4.30
Villes non centrales de petites ZSM	2.90	3.15	2.26	1.30												
Autres agglomérations 10 000+	1.38	1.89	1.68	1.25	7.39 *		-3.22	7.80 *		0.18	4.65 *		3.32	4.68 *		3.58
Autres zones	3.15	1.92	2.56	2.51												
Réserves indiennes							12.72			12.72			12.72			12.72

SESSION 2

Considérations spatiales lors de la conception des enquêtes ou des bases de sondage

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

CONSTRUCTION DE LISTES À ARTICULATION SPATIALE POUR LES ENQUÊTES DE MÉNAGE

A. Saalfeld[1]

RÉSUMÉ

Cet article présente et compare quelques méthodes, dont certaines sont anciennes et certaines sont nouvelles, servant à ordonner des entités spatiales d'une manière qui reflète leur proximité spatiale.

L'article décrit ensuite comment les entités ainsi ordonnées peuvent constituer des listes qui serviront à effectuer un échantillonnage systématique. Ces nouvelles méthodes découlent des méthodes d'ordination des sommets ou arêtes d'un graphe acyclique connexe, un circuit eulérien étant utilisé pour définir un ordre cyclique. Les nouvelles méthodes d'ordination sont indépendantes des systèmes de coordonnées et invariantes par rotation. Ces nouvelles méthodes d'ordination ne dépendent pas d'un "groupement" préalable de l'espace, contrairement à certaines méthodes standard. Ces nouvelles ordinations dépendent plutôt de la distribution spatiale des données d'entrée et de la métrique de l'espace sous-jacent.

On appelle *ordinations par arbre* ces nouvelles ordinations. On peut les construire en temps linéaire à partir de structures de données topologiques. Elles sont entièrement caractérisées par une propriété utile de préservation de la proximité, appelée *récursivité par branche*. L'article décrit de quelle façon les techniques d'ordination par arbre peuvent servir à trouver des classements pour les éléments spatiaux de base listes utilisées en échantillonnage multiple:

- Ordination de points dans le plan (unités d'habitation).
- Ordination de surfaces (pâtés de maisons ou régions).
- Ordination de segments de droite (sections de rue).
- Ordination de segments de pseudo-droites (côtés d'îlots).
- Ordination de n'importe laquelle des entités mentionnées ci-dessus de façon à respecter les hiérarchies territoriales.

Pour chacune des entités spatiales mentionnées ci-dessus, les ordinations engendrées par une extension des méthodes d'ordination par arbre présentent d'importantes propriétés de préservation de la proximité. L'article décrit des applications à l'échantillonnage de ces nouvelles façons de classer divers types d'objets spatiaux.

MOTS CLÉS: Données spatiales; ordres linéaires; arbres; proximité.

1. INTRODUCTION

Il est utile, et quelquefois nécessaire, d'ordonner des données. L'ordinateur traditionnel en série (contrairement à certains appareils modernes qui fonctionnent en parallèle) utilise un paradigme linéaire. Celui-ci exige que les données soient lues dans un certain ordre et place ces données en mémoire dans des cases consécutives qui peuvent être récupérées par groupes ou par blocs à la fois. À son niveau le plus élémentaire, la gestion de bases de données est l'art d'organiser ou d'ordonner des données pour qu'on puisse y avoir accès le plus facilement

[1] A. Saalfeld, Statistical Research Division, U.S. Bureau of the Census, Washington DC 20233, États-Unis.

possible et les utiliser le plus efficacement possible pour effectuer certaines opérations d'intérêt particulier. Le présent article expose une nouvelle façon d'ordonner des données, qui permettra d'effectuer un ensemble d'opérations importantes avec économie et efficacité. Ces opérations comprennent les opérations générales de gestion de données informatiques ainsi que certaines opérations importantes d'échantillonnage telles que l'échantillonnage en grappes et l'échantillonnage systématique. Dans la présente section, nous passons en revue et résumons certaines définitions et concepts de base nécessaires à la description de nos techniques d'ordination.

1.1 Ordre et listes

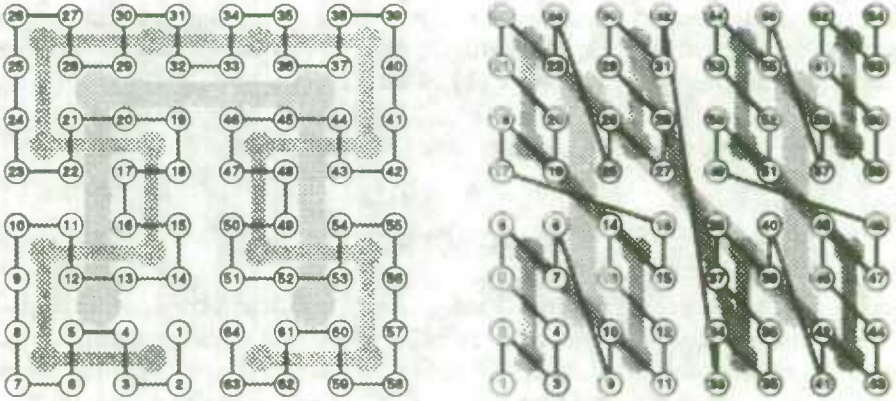
Tout au long du présent article, *ordre* ou *ordination*, sans qualificatif, signifie un ordre total ou ordre linéaire sur un ensemble fini de n éléments. Ce type d'ordre n'est rien d'autre qu'un ordonnancement des n éléments, une correspondance biunivoque entre les éléments de l'ensemble et les entiers de 1 à n, ou une liste de ces n éléments. On appellera *liste* ou *liste ordonnée* un ensemble de n éléments qui ont été ordonnés.

Dans une liste de n éléments, l'élément de rang $(i + 1)$ est dit le *successeur* de l'élément de rang i ; tout élément sauf le $n^{ième}$ ou dernier élément a un unique successeur. De la même façon, tout élément sauf le premier a un unique *prédécesseur*. Nous pouvons construire une *liste cyclique* ou définir un *ordre cyclique* à partir d'une liste en nommant le premier élément comme successeur du dernier (et le dernier élément comme prédécesseur du premier). Les listes cycliques sont souvent utiles parce qu'elles n'ont pas d'élément distingué qui exige un traitement spécial. Par exemple, si on utilise une liste cyclique, il est possible de commencer *n'importe où* dans la liste et d'en énumérer tous les éléments de façon exhaustive par un passage répété au successeur jusqu'au retour à l'élément de départ choisi.

1.1.1 Limites des méthodes d'ordination linéaire

Une représentation à une seule dimension ne peut refléter avec exactitude toute la complexité d'une structure bidimensionnelle. Par exemple, il n'existe aucune application bijective continue d'une droite dans un plan qui ait un inverse continu. De la même façon, il ne peut exister d'application d'un réseau rectangulaire complet de points sur un ensemble de points de la droite qui préserve la distance. Par contre, il est possible de trouver des applications bijectives des entiers 1 à 4^m sur un réseau rectangulaire 2^m par 2^m, qui préservent la contiguïté (une de ces applications s'appelle courbe de Hilbert et est illustrée à la figure 1). Il est clair que l'application inverse ne pourra jamais préserver *toutes* les contiguïtés, puisque chaque point intérieur du réseau a 4 voisins et que chaque entier intérieur sur la droite en a deux. Les applications de la section 3 indiquent certaines autres propriétés qui ne sont pas toujours préservées avec d'autres types de données spatiales lorsqu'on emploie un ordre linéaire.

Figure 1: Ordres récursifs par quadrant:
courbe de Hilbert (à gauche) et clé de Peano



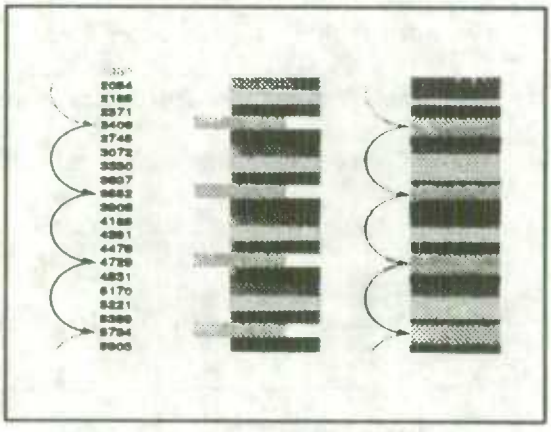
1.2 Interrogation spatiale

On attribue souvent un ordre ou clé primaire aux points dans l'espace à deux ou plusieurs dimensions afin de faciliter leur entreposage dans des bases de données et de rendre plus simple leur récupération. Certaines courbes fractales, comme la clé de Peano et la courbe de Hilbert (Faloutsos et Rong 1989), (Faloutsos et Roseman 1989) se sont avérées très utiles pour l'interrogation portant sur l'étendue et les recherches du plus proche voisin. Ces courbes sont des exemples d'une classe plus vaste d'ordres, dits *récursifs par quadrant* (Mark 1990). Un ordre est dit récursif par quadrant si, dans toute décomposition récursive en sous-quadrants d'une région rectangulaire, les points de n'importe quel sous-quadrant sont toujours numérotés de façon consécutive. Les points de n'importe quel sous-quadrant sont toujours énumérés de façon exhaustive avant de sortir du sous-quadrant. Nous verrons à la section 2.2.4 que les ordres récursifs par quadrant sont un cas spécial d'une classe plus générale d'ordres, dits *récursifs par branche*.

1.3 Échantillonnage systématique

On entend généralement par *échantillonnage systématique* la sélection, dans une liste, d'un sous-ensemble obtenu en choisissant des éléments à intervalles réguliers (la longueur de cet intervalle étant désignée comme l'*intervalle de sélection*) (Kish 1965). Les éléments peuvent être *pondérés* pour corriger leur probabilité de sélection (voir figure 2).

Figure 2: Échantillonnage systématique à partir de listes



Si tous les éléments sont de poids 1, alors un intervalle de sélection de k produira un taux d'échantillonnage de $1/k$. On peut considérer l'échantillon comme étant muni de l'ordre induit par la sélection systématique de ses éléments à l'aide de sauts successifs dans la liste.

Si on attribue à un ensemble de points du plan un ordre récursif par quadrant, alors un procédé de sélection systématique échantillonnera chaque sous-quadrant, peu importe sa taille, avec une proportion égale au taux d'échantillonnage global, à une unité près. Cette propriété de complétude de la représentation a été notée et utilisée par Wolter et Harter dans le cas de l'ordre induit par une clé de Peano (Wolter et coll. 1989).

1.4 Graphes et cartes

Le dessin des lignes de n'importe quelle carte géographique possède une structure sous-jacente de *graphe*[2]. Nous utiliserons les définitions combinatoires habituelles de la théorie des graphes, telles qu'on peut les trouver dans le texte classique de Harary (Harary 1969). Un *graphe* $G = (V, E)$ consiste en un ensemble fini non-vide

---

[2] Pour certaines applications, il peut être préférable et même nécessaire de considérer les lignes d'une carte géographique comme un *pseudo-graphe*, structure qui peut comporter plusieurs arêtes entre deux sommets. Pour les applications que nous étudions ici, cette distinction est sans importance.

V de *sommets* et en un ensemble E de couples non ordonnés de sommets appelés *arêtes*. Un sommet v et une arête $\{u, w\}$ sont dits *incidents* si et seulement si $v = u$ ou $v = w$. Le *degré* d'un sommet est le nombre d'arêtes incidentes à ce sommet. Une *promenade* dans le graphe G est une suite $(v_1v_2v_3...v_k)$ de sommets v_i, non nécessairement distincts, telle que, pour chaque $j = 1,2,...,(k - 1)$, $\{v_j, v_{j+1}\}$ est une arête de G. Un *circuit* est une promenade $(v_1v_2v_3...v_k)$ telle que $v_1 = v_k$. Un *chemin* est une promenade où aucune arête n'est répétée. Un *cycle* est un chemin $(v_1v_2v_3...v_k)$, où $k \geq 3$, tel que $v_1 = v_k$. Un *arbre* est un graphe sans cycles. Un arbre tel que nous l'avons défini est quelquefois appelé *arbre libre* pour le distinguer d'une *arborescence* qui elle, possède un sommet distingué appelé *racine*.

1.5 Arbres

Nous énonçons quelques propriétés des arbres qui les rendent plus faciles à utiliser que les graphes en général. Nous indiquerons dans les sections 3.3 et 3.4 de quelle façon les problèmes d'ordination des composantes d'un graphe peuvent être transformés en problèmes d'ordination des composantes d'un arbre dérivé. Les informaticiens ont mis au point plusieurs façons d'ordonner les sommets d'arborescences plongées dans le plan (Aho 1985). Nous examinerons de nouveaux ordres pour les arbres libres.

1.5.1 Propriétés combinatoires des arbres

Nous dressons une liste de quelques propriétés importantes des arbres:

Propriété 1 *Tout arbre à n sommets a exactement $(n - 1)$ arêtes.*

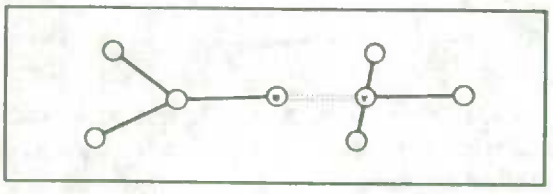
Propriété 2 *Un graphe connexe à n sommets et $(n - 1)$ arêtes est un arbre.*

Propriété 3 *L'addition à un arbre d'une nouvelle arête (entre deux sommets existants) crée toujours un cycle.*

Propriété 4 *La suppression d'une arête dans un arbre produit toujours deux composantes connexes.*

Propriété 5 *Entre deux sommets quelconques d'un arbre, il y a toujours un et un seul chemin.*

Figure 3: La suppression d'une arête crée deux branches



Nous disons que chaque arête détermine *deux branches* (voir figure 3), qui sont les deux composantes connexes du graphe résultant de la suppression de cette arête. Ces deux branches sont elles-mêmes des arbres. Le sommet u est toujours contenu dans une des deux branches déterminées par l'arête $\{u, v\}$, et le sommet v est toujours contenu dans l'autre.

1.5.2 Plongements d'arbres dans le plan

Il n'est pas toujours possible de dessiner un graphe dans le plan à l'aide de segments de droite qui ne se croisent pas, mais un arbre peut toujours être représenté ou "réalisé" comme un dessin formé de segments de droite dans le plan. De plus, supposons que, pour tout sommet d'un arbre, nous attribuons de façon arbitraire un ordre cyclique aux arêtes incidentes à ce sommet. Alors, il est toujours possible de dessiner l'arbre dans le plan à l'aide de segments de droite de telle façon que l'ordre horaire des arêtes incidentes à tout sommet soit l'ordre arbitrairement attribué au départ à ces arêtes.

2. CIRCUITS EULÉRIENS ET ORDRES INDUITS PAR DES ARBRES

2.1 Circuits eulériens

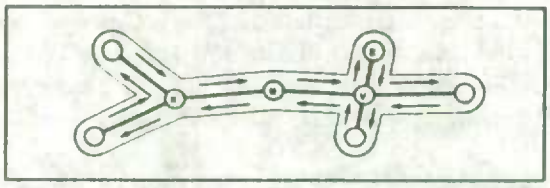
Un circuit eulérien dans un arbre est un circuit particulier, bien équilibré, qui passe par chaque arête exactement deux fois. Nous présentons deux descriptions d'un circuit eulérien dans un arbre. *Chaque description dépend de l'attribution d'un ordre cyclique aux arêtes incidentes à chaque sommet.*

2.1.1 Version géométrique

Dessinons l'arbre de telle façon que l'ordre cyclique attribué à chaque sommet soit l'ordre des arêtes dans le sens des aiguilles d'une montre. Faisons commencer un circuit à partir d'un sommet quelconque x. Suivons l'une quelconque des arêtes incidentes $\{x, u\}$ en direction de u. Une fois arrivé à u, repartons en suivant l'arête $\{u, v\}$ qui suit $\{x, u\}$ dans l'ordre horaire des arêtes incidentes à u. Une fois arrivés à v, repartons en suivant l'arête $\{v, z\}$ qui suit $\{u, v\}$ dans l'ordre horaire des arêtes incidentes à v, etc., jusqu'à ce que nous retournions finalement à x par l'arête qui précède $\{x, u\}$ dans l'ordre horaire (voir figure 4).

Le circuit que nous venons de décrire passera deux fois par chaque arête, une fois dans chaque direction, et visitera chaque sommet un nombre de fois égal au degré de ce sommet.

Figure 4: Représentation géométrique d'un circuit eulérien



On peut aussi visualiser ce circuit de la façon suivante : imaginons que l'arbre lui-même est un mur vu du dessus. Marchons le long de ce mur en le touchant continuellement de la main droite. Nous finirons par revenir à notre point de départ après avoir touché les deux côtés de chacun des pans du mur. Donc, si quelqu'un avait marché sur le haut du mur en nous suivant constamment, il aurait passé exactement deux fois par chacune des arêtes de l'arbre.

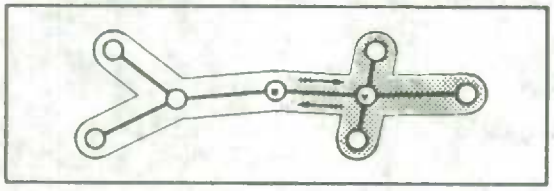
2.1.2 Version combinatoire

Un circuit eulérien sur un arbre de n sommets est une promenade $(v_1 v_2 v_3 \ldots v_{2n-1})$, où les v_i sont des sommets, évidemment pas tous distincts, tels que l'on ait :

1. $v_1 = v_{2n-1}$.
2. Pour $i = 1, 2, \ldots, 2n - 3$, $\{v_{i+1}, v_{i+2}\}$ est le successeur de $\{v_i, v_{i+1}\}$ dans l'ordre cyclique des arêtes incidentes au sommet v_{i+1}.
3. $\{v_1, v_2\}$ est le successeur de $\{v_{2n-2}, v_1\}$ dans l'ordre cyclique des arêtes incidentes au sommet v_1.
4. Pour chaque arête $\{u, v\}$ de l'arbre, la suite uv et la suite vu apparaissent chacune exactement une fois dans la promenade $(v_1 v_2 \ldots v_{2n-1})$.
5. Chaque sommet sauf v_1 apparaît un nombre de fois égal à son degré.
6. Le sommet v_1 apparaît une fois de plus que son degré.

Remarquons de plus que si uv apparaît avant vu dans le circuit eulérien, alors la sous-promenade qui va de uv à vu en commençant et en finissant à v couvre complètement tous les sommets et toutes les arêtes de la branche-v déterminée par l'arête $\{u, v\}$. De plus, cette sous-promenade ne touche à rien d'autre que la branche-v de l'arbre (voir figure 5).

Figure 5: Une sous-promenade qui couvre une branche entière



De même, le reste du circuit (soit la partie avant uv et après vu) couvre complètement chaque arête et chaque sommet de la branche-u résultant de la suppression de l'arête $\{u, v\}$. Il est clair qu'aucune arête et aucun sommet de la branche-v ne peut apparaître dans la branche-u.

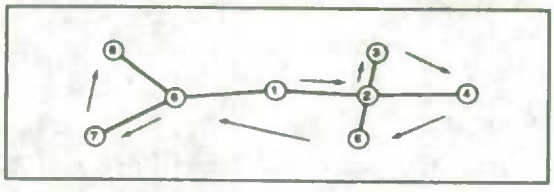
2.1.3 Ordination eulérienne par arbre

Tout circuit eulérien visite chacun des sommets et chacune des arêtes au moins une fois. Supposons que nous voulions numéroter nos n sommets à l'aide des entiers de 1 à n. On appellera *ordination eulérienne par arbre* (OEA) des sommets un procédé qui consiste à visiter les sommets dans l'ordre induit par un circuit eulérien, en attribuant soit le plus petit entier non attribué ou aucun entier à chaque *visite* de chaque sommet de telle façon qu'un entier soit attribué à exactement une des deux visites de chaque sommet.

On appellera *ordination eulérienne par arbre* (OEA) des arêtes un procédé qui consiste à visiter les arêtes dans l'ordre d'un circuit eulérien et à attribuer soit le plus petit entier non attribué ou aucun entier à chaque *visite* de chaque arête de telle façon qu'un entier soit attribué à exactement une des deux visites de chaque arête.

Garey et Johnson (Garey 1979) et d'autres auteurs (Preparata 1985), (Edelsbrunner 1987) décrivent une de ces ordinations eulériennes par arbre des sommets d'un arbre recouvrant de longueur minimum (ARLM), dont le point de départ se situe n'importe où dans le circuit eulérien et pour laquelle le plus petit entier non attribué est attribué à la *première* visite de chacun des sommets (voir figure 6).

Figure 6: OEA avec attribution à la première visite



L'ordre qu'ils obtiennent, comme d'ailleurs toute ordination eulérienne par arbre d'un ARLM, sera toujours une approximation à un facteur 2 près du circuit euclidien du commis-voyageur, puisque le circuit eulérien lui-même n'a jamais plus de deux fois la longueur du circuit euclidien du commis-voyageur.

Nous décrivons maintenant une méthode simple pour engendrer toutes les autres ordinations eulériennes par arbre des sommets d'un arbre.

2.2 Ordinations par arbre des sommets

Supposons que nous avons un arbre *et un circuit eulérien* $(v_1 v_2 \ldots v_{2n-1})$ sur cet arbre (ou, de façon équivalente, un plongement de l'arbre dans le plan ainsi qu'un sommet et une arête de départ). Alors, pour ordonner les sommets, nous procéderons comme suit.

2.2.1 Mise en place : pondération des visites de sommets

Pour $i = 1, 2, 3, \ldots, (2n - 1)$, considérons chaque v_i qui apparaît dans le circuit $(v_1 v_2 \ldots v_{2n-1})$ comme une visite de sommet.

Pour $i = 1, 2, 3, \ldots, (2n - 1)$, attribuons une pondération non négative w_i à la $i^{\text{ième}}$ visite de telle façon que la somme des pondérations des visites à un sommet donné quelconque v soit un :

$$\text{Pour tout } v \in V, \sum_{(i|v_i=v)} w_i = 1.$$

Une telle pondération est appelée *pondération unitaire.*

Un exemple important de pondération unitaire est celle qui attribue le même poids à toutes les visites d'un même sommet. Comme chaque sommet v_i est visité $\deg(v_i)$ fois[3], ce poids uniforme est donné de façon exacte par :

$$w_i = \frac{1}{\deg(v_i)}, \quad \text{pour } i = 1, 2, \ldots, (2n - 2);$$

$$\text{et} \quad w_{2n-1} = 0.$$

2.2.2 Construction de l'intervalle d'échantillonnage

À mesure que nous parcourons le circuit eulérien, nous commençons à cumuler des poids (exactement comme lorsqu'on construit une liste pondérée pour un échantillonnage systématique).

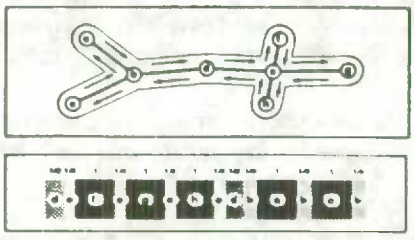
$$\text{Soit} \quad W_0 = 0,$$

$$\text{et} \quad W_j = W_{j-1} + w_j.$$

2.2.3 Une illustration : la pondération uniforme

Nous illustrons les poids cumul au cours du procédé de pondération uniforme en nous servant de la promenade (*defegehedbabcbd*), représentée à la figure 7.

Figure 7: Un circuit et ses poids cumulés



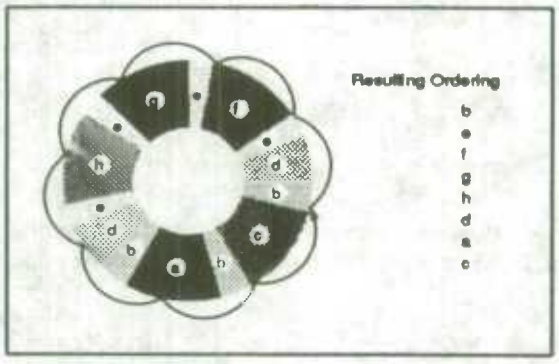
[3] Comme le circuit eulérien est cyclique, nous voulons compter v_1 et v_{2n-1} comme une seule et même visite et n'attribuer qu'à un des deux sommets le poids approprié. Nous avons choisi d'attribuer le poids uniforme à v_1.

Le total cumulatif des poids est exactement égal à n et le poids total correspondant à chacun des sommets est exactement égal à un. Nous pouvons numéroter les sommets en faisant des sauts de longueur 1 dans l'intervalle des poids cumulés, ce qui équivaut à attribuer un numéro à un sommet chaque fois qu'une visite de sommet nous fait égaler ou dépasser le prochain entier non attribué :

Si $\lfloor W_{j-1} \rfloor < \lfloor W_j \rfloor$, nous attribuerons au sommet v_j le numéro $\lfloor W_j \rfloor$.

Comme certains sommets apparaissent à plusieurs endroits dans notre intervalle de poids cumulés, on pourrait craindre que notre procédé de numérotation n'attribue plus d'un numéro à un même sommet. Une attribution de ce genre est toutefois impossible.

Figure 8: Classement cyclique par pondération uniforme



Comme le circuit eulérien est cyclique, nous pouvons rendre cycliques notre intervalle des poids cumulés des visites de sommets ainsi que le classement qui en résulte, en rendant la construction de cet intervalle indépendante du point de départ du circuit. C'est ce qui est illustré à la figure 8. Le théorème suivant, démontré par Saalfeld (1991), garantit que ce procédé de numérotation fournit une ordination des sommets.

Théorème 2.1 *En parcourant un circuit eulérien d'un arbre, construisons un intervalle (cyclique) séparé, d'une longueur totale de n unités, en attribuant un poids non négatif arbitraire à chaque visite de sommet de telle façon que le poids total de toutes les visites d'un sommet donné quelconque soit un. Alors chaque sommet sera touché exactement une fois si on saute d'une unité à la fois dans l'intervalle cyclique de longueur n.*

2.2.4 Récursivité par branche

Dans toute la présente section, nous ferons du premier sommet le successeur du dernier et pourrons ainsi considérer que les ordres engendrés par notre méthode d'ordination sont cycliques.

Corollaire 2.2. *Pour toute OEA cyclique, les numéros des sommets d'***une branche quelconque** *forment toujours un intervalle complet, c'est-à-dire un ensemble d'entiers consécutifs.*

Nous dirons qu'une ordination cyclique des sommets dans laquelle les numéros des sommets d'une branche forment toujours un ensemble d'entiers consécutifs est une *ordination récursive par branche*. Le corollaire 2.2 affirme que toute ordination eulérienne par arbre est récursive par branche. La réciproque est également vraie.

Théorème 2.3 *Toute ordination cyclique récursive par branche des sommets d'un arbre est une ordination eulérienne par arbre résultant d'un circuit eulérien de l'arbre et d'une pondération unitaire des visites de sommets correspondant à ce circuit eulérien.*

La récursivité par branche est une très forte propriété de préservation de la proximité, lorsque la proximité est mesurée par la longueur des chaînes reliant deux sommets dans l'arbre ou le graphe. Les branches de l'arbre pourront correspondre à des grappes de données lorsque l'arbre qui a été construit est un arbre recouvrant de longueur minimum. Toutes les ordinations récursives par quadrant d'un ensemble de points du plan peuvent

être réalisées comme les ordinations induites sur des sous-ensembles feuilles par des ordinations récursives par branche (c'est-à-dire *eulériennes par arbre*) de la quadripartition de ces points.

2.3 Ordinations par arbre des arêtes

Nos méthodes d'ordination des sommets valent également pour l'ordination des arêtes. Le théorème 2.1 sur les sommets a sa contrepartie exacte pour les arêtes:

Théorème 2.4 *En parcourant un circuit eulérien d'un arbre, construisons un intervalle (cyclique) séparé, d'une longueur totale de* $(n - 1)$ *unités, en attribuant un poids non négatif arbitraire à chaque visite d'arête de telle façon que le poids total de toutes les visites d'une arête donnée quelconque soit un. Alors chaque arête sera touchée exactement une fois si on saute d'une unité à la fois dans l'intervalle cyclique de longueur* $(n - 1)$.

2.3.1 Pondération uniforme des arêtes

Dans un procédé de pondération *uniforme* des arêtes, plutôt que des sommets, chaque arête recevrait exactement le poids ½ (puisque chaque arête est visitée exactement deux fois dans le circuit eulérien). Mais donner un poids de ½ à chaque arête équivaut à sauter une arête sur deux dans notre méthode de sélection. Nous avons donc les corollaires suivants du théorème 2.4:

Corollaire 2.5 *En parcourant un circuit eulérien d'un arbre, attribuons un numéro d'arête à toutes les deux visites d'arête exactement. Alors chaque arête recevra un seul numéro.*

Corollaire 2.6 *La distance de chaîne entre deux arêtes qui ont reçu des numéros consécutifs dans un procédé de pondération uniforme n'est jamais supérieure à 2.*

2.3.2 Récursivité par branche

Comme pour les sommets, chaque ordre induit par un arbre sur les arêtes est récursif par branche dans le même sens:

Corollaire 2.7 *Pour toute OEA cyclique des arêtes, les numéros des arêtes d'***une branche quelconque** *forment toujours un intervalle complet, c'est-à-dire un ensemble d'entiers consécutifs.*

Et réciproquement:

Théorème 2.8 *Toute ordination cyclique récursive par branche des arêtes d'un arbre est une ordination eulérienne par arbre résultant d'un circuit eulérien de l'arbre et d'une pondération unitaire des visites d'arêtes correspondant à ce circuit eulérien.*

3. CONSTRUCTION DE LISTES

Dans la présente section, nous adapterons nos techniques d'ordination par arbre à l'ordination d'objets situés dans l'espace. Dans chaque cas, notre démarche consistera à transformer le problème d'ordination en un problème d'ordination par arbre, puis à résoudre le problème d'ordination par arbre à l'aide d'une pondération uniforme des sommets ou des arêtes, selon le cas.

3.1 Ordination de points du plan: ménages

Si nous attribuons des coordonnées aux ménages, en les voyant comme des points du plan, il est possible d'ordonner ces points de la façon suivante. Nous savons comment ordonner les sommets d'un arbre. Nous pouvons donc transformer les points en sommets en construisant un arbre (c'est-à-dire en ajoutant des arêtes); une construction qui vient naturellement à l'esprit est celle d'un arbre recouvrant euclidien de longueur minimum (ARELM). L'ARELM est unique si les points sont en position générale ou si aucune des distances entre deux

points n'est égale à une autre. Les étapes de la transformation d'un problème d'ordination de points dans l'espace en un problème d'ordination des sommets d'un arbre sont donc:

ALGORITHME ORDONNER_POINTS_PLAN

1. **Construire l'arbre recouvrant euclidien de longueur minimale.**
2. **Parcourir un circuit eulérien en ordonnant les sommets.**

Nous pouvons construire un arbre recouvrant euclidien de longueur minimum en temps $O(n \log n)$ (Aho 1985), en ordonnant les arêtes incidentes à chaque sommet, dans le sens horaire, à mesure qu'elles sont insérées. Le plongement de l'arbre dans le plan nous donne en prime la version géométrique du circuit eulérien (c'est-à-dire l'ordre habituel des arêtes autour d'un sommet, dans le sens des aiguilles d'une montre), ce qui nous permet de parcourir le circuit eulérien et d'ordonner les sommets en un temps supplémentaire d'ordre $O(n)$.

3.1.1 Une application à l'échantillonnage en grappes

L'échantillonnage en grappes est une stratégie de sondage qui consiste à choisir de petits groupes (grappes) de points voisins plutôt que de sélectionner des points individuels distribués de façon aléatoire. La corrélation intra-grappe réduit peut-être l'efficacité de cette stratégie du pur point de vue de la théorie de l'échantillonnage, mais cet inconvénient est souvent compensé par les avantages économiques d'une réduction des frais de déplacement des enquêteurs.

Toutefois, un obstacle important à la sélection de grappes à partir de listes est le fait que la proximité sur la liste ne garantit pas la proximité sur le terrain. La sélection de points à partir d'une liste qui a été ordonnée à l'aide de notre algorithme d'ordination par pondération uniforme d'un ARLM des points d'une liste garantira une très bonne préservation de la proximité. On a le théorème suivant:

Théorème 3.1 *Ordonnons des points du plan en construisant leur ARLM et en appliquant l'algorithme d'ordination des sommets par pondération uniforme. La distance de chaîne entre deux points consécutifs dans cet ordre est alors de six au maximum et la distance de chaîne moyenne est inférieure à deux.*

Démonstration: Le degré de tout sommet d'un ARLM est inférieur ou égal à six. L'algorithme d'ordination des sommets par pondération uniforme cumulera donc un poids d'au moins 1/6 à chaque visite de sommet. De plus le degré moyen des sommets de n'importe quel arbre est $\frac{2n-2}{n}$.

3.2 Ordination de points dans des espaces de dimensions supérieures

Pour appliquer les méthodes de la section 3.1 à des points situés dans des espaces à plus de deux dimensions, nous devons d'abord résoudre deux problèmes: 1. construire un ARLM en dimensions supérieures; 2. définir un circuit eulérien à plus de deux dimensions.

Il existe des algorithmes élémentaires, réalisables en temps $O(n^2)$ pour construire un ARLM en dimensions supérieures (Aho 1974). On connaît aussi certains algorithmes exacts dont la complexité est légèrement sous-quadratique (Yao 1982).

La construction d'un circuit eulérien en dimensions supérieures n'est pas aussi simple. Elle exige qu'on fixe un ordre pour les arêtes incidentes à chaque sommet. Il est possible de projeter les arêtes sur un sous-espace quelconque à deux dimensions, puis d'ordonner dans le sens horaire la projection des arêtes sur ce plan. Une autre approche, d'application plus générale, a été proposée par Herbert Edelsbrunner (Edelsbrunner 1990). Elle consiste à appliquer la configuration des arêtes incidentes à un sommet sur des points d'une sphère dont la dimension est inférieure de un à celle de l'espace où l'ARLM est plongé, puis à appliquer récursivement le procédé d'ordination à ces points de la sphère (c'est-à-dire à construire leur ARLM et à les ordonner dans un espace de dimension inférieure).

Quoi qu'il en soit, si nous n'avons besoin que d'une ordination quelconque des arêtes incidentes à chaque sommet, nous pouvons en trouver une en temps $O(n \log n)$. Résumons les étapes nécessaires à la conversion du problème en un problème d'ordination par arbre.

ALGORITHME ORDONNER_POINTS_À_N_DIMENSIONS

1. **Construire l'ARLM.**
2. **Donner un ordre cyclique aux arêtes incidentes à chaque sommet.**
3. **Parcourir un circuit eulérien en ordonnant les sommets.**

3.2.1 Une application à la stratification d'un échantillon

La stratification d'un échantillon est la division de l'univers en groupes d'individus qui ont plusieurs caractéristiques similaires. Ces caractéristiques devraient toujours être comparables dans une certaine mesure (l'étude de l'incomparabilité relative constitue ce qu'on appelle quelquefois le problème d'échelle). On stratifie souvent en traitant les observations comme des n-tuplets des n caractéristiques (c'est-à-dire comme des points de l'espace à n dimensions) et en trouvant un hyperplan ou une famille d'hyperplans qui assurent une séparation optimale des points par des demi-espaces ou cellules à n dimensions délimitées par ces hyperplans. Une façon plus élémentaire de voir la stratification (qui simplifierait aussi de beaucoup les calculs) pourrait être de découper un ARLM des points de l'échantillon en branches de séparation maximale. Après numérotation à l'aide de méthodes d'ordination récursive par branche, ce découpage équivaut à diviser la liste en sous-listes (et donc à stratifier!). Au Bureau of the Census, nous nous proposons de comparer les résultats obtenus de ces méthodes d'ordination par arbre à ceux obtenus à l'aide des algorithmes de stratification courants, plus complexes.

3.3 Ordination des sommets d'un graphe quelconque

Si nous sommes seulement intéressés à l'ordination des sommets d'un graphe, nous pouvons considérer le graphe comme un arbre avec trop d'arêtes. Nous taillerons donc le graphe en enlevant les arêtes les moins utiles, jusqu'à ce que nous en ayons fait un arbre. Si des coûts sont associés aux arêtes, il se peut que nous voulions minimiser le coût de l'arbre résultant. Nous savons exactement combien d'arêtes nous devons supprimer. Nous rejetterons une arête pourvu que cela ne sépare pas le graphe en deux composantes connexes et qu'il nous reste toujours au moins $(n - 1)$ arêtes. Résumons les étapes nécessaires à la conversion du problème en un problème d'ordination par arbre.

ALGORITHME ORDONNER_SOMMETS_GRAPHE

1. **Construire un arbre recouvrant (de longueur minimum).**
2. **Donner un ordre cyclique aux arêtes incidentes à chaque sommet.**
3. **Parcourir un circuit eulérien en ordonnant les sommets.**

3.4 Ordination des arêtes d'un graphe quelconque

Dans la section 3.3, nous avons considéré que notre graphe avait trop d'arêtes et nous en avons supprimé quelques-unes. Pour ordonner les arêtes de nos graphes, nous considérerons que notre graphe a trop peu de sommets pour être un arbre et nous ajouterons des sommets au graphe en effectuant une subdivision des sommets qui crée de nouveaux sommets sans changer le nombre d'arêtes (voir figure 9). Encore une fois, nous utilisons notre connaissance de la relation entre le nombre de sommets et d'arêtes dans un arbre pour déterminer à quel moment nous devons arrêter de subdiviser des sommets. Résumons les étapes nécessaires à la conversion du problème en un problème d'ordination par arbre.

ALGORITHME ORDONNER_ARÊTES_GRAPHE

1. **Subdiviser les sommets.**
2. **Donner un ordre cyclique aux arêtes incidentes à chaque sommet.**
3. **Parcourir un circuit eulérien en ordonnant les arêtes.**

Nous devons ensuite ordonner les arêtes de l'arbre autour de chaque sommet subdivisé. On pourra alors attribuer un ordre cyclique aux arêtes de l'arbre en attribuant des numéros consécutifs à une arête sur deux dans un parcours d'un circuit eulérien des arêtes correspondantes de l'arbre dérivé.

Comme nous pouvons certainement subdiviser les sommets en temps $O(n \log n)$ en utilisant une version modifiée de la recherche primaire en profondeur, et comme nous pouvons aussi trouver une façon acceptable

d'ordonner les arêtes incidentes à chacun des sommets subdivisés avec la même complexité temporelle d'exécution, nous pouvons effectuer avec efficacité l'ordination suivante des arêtes d'un graphe connexe quelconque :

Corollaire 3.2 *Il est possible d'attribuer un ordre cyclique aux arêtes d'un graphe connexe quelconque en temps $O(n \log n)$ de telle façon que deux arêtes recevant des numéros consécutifs dans cette ordination ne soient jamais à une distance de chaîne supérieure à deux dans le graphe.*

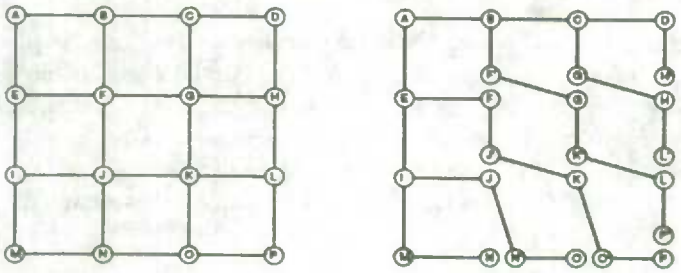
3.5 Ordination de segments de droite dans des réseaux à deux dimensions

Ce problème n'est rien d'autre que le problème d'ordination des arêtes d'un graphe. Il y a toutefois moins de décisions à prendre que dans le cas général, parce que le circuit eulérien est fourni par la structure géométrique. Le mot *réseau* implique aussi que l'information topologique du graphe permet une génération en temps linéaire de l'ordre recherché. Les étapes nécessaires à la conversion d'un problème d'ordination des arêtes d'un réseau connexe en un problème d'ordination des arêtes d'un arbre sont :

ALGORITHME ORDONNER_SEGMENTS_DE_RÉSEAU_2-D

1. **Subdiviser les sommets.**
2. **Parcourir un circuit eulérien en ordonnant les sommets.**

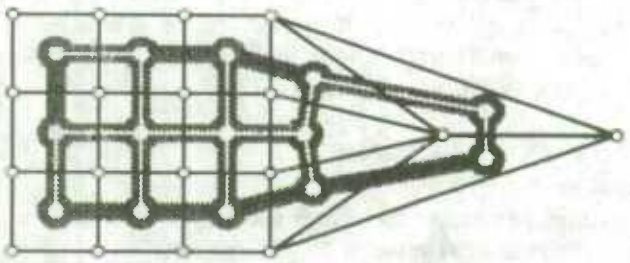
Figure 9: Subdivision de sommets dans un réseau à 2 dimensions



3.6 Ordination de régions du plan

Tout graphe ou pseudographe du plan possède un pseudographe dual qui est lui-même un pseudographe. Chaque région du plan correspond à un sommet du nouveau pseudographe; deux sommets du nouveau pseudographe sont adjacents (ont une arête en commun) si et seulement si les régions ont une frontière ou côté en commun. Ce dual est appelé *pseudographe d'adjacence*; la méthode pour réduire un pseudographe à un arbre est la même que pour un graphe : éliminer certaines arêtes.

Figure 10: Graphe planaire, graphe dual et arbre recouvrant du dual



Résumons les étapes nécessaires à la conversion du problème en un problème d'ordination par arbre.

ALGORITHME ORDONNER_RÉGIONS

1. **Construire le pseudographe d'adjacence.**
2. **Trouver un arbre recouvrant minimal.**
3. **Parcourir un circuit eulérien en ordonnant les sommets.**

3.6.1 Application au numérotage par bloc

Considérons le problème qui consiste à numéroter les régions d'une carte de telle façon que deux régions qui reçoivent des numéros consécutifs soient adjacentes. Il est bien connu que ce ne sont pas tous les arrangements de blocs qui peuvent être numérotés de cette manière. En fait, lorsqu'on formule le problème dans le graphe dual, le numérotage par bloc n'est rien d'autre que la recherche d'un chemin hamiltonien dans le graphe d'adjacence (c'est-à-dire un chemin qui passe exactement une fois par chacun des sommets). Même le problème qui consiste à décider si un tel chemin existe pour un graphe planaire arbitraire est NP-complet.

La suppression d'arêtes de façon à minimiser le degré maximum des sommets de l'arbre résultant permet de garantir que la distance de chaîne entre des blocs ayant reçu des numéros consécutifs ne sera pas supérieure au degré maximum de l'arbre élagué.

3.6.2 Échantillonnage multiple

L'échantillonnage se fait souvent par étapes. On choisira d'abord certaines régions; puis les ménages des régions sélectionnées seront soumis à un sous-échantillonnage. L'échantillonnage en grappes de régions, soit la possibilité de choisir des ensembles de régions voisines est un facteur important dans la réduction des frais de déplacement et autres frais d'exploitation liés aux sondages. La réalisation d'un *échantillonnage de régions en grappes non compactes* suppose la sélection de régions rapprochées mais non adjacentes. L'échantillonnage en grappes non compactes est une tentative pour profiter des avantages liés à une réduction des frais de déplacement sans que se manifestent les effets négatifs de corrélations élevées. La numérotation des régions à l'aide d'un arbre obtenu par élagage du graphe d'adjacence des régions est une technique fiable pour créer des grappes non compactes de régions. Un autre moyen de réaliser un échantillonnage en grappes de points situés dans des régions déterminées à l'avance consiste, lorsqu'on construit un arbre recouvrant de coût minimum, à attribuer des coûts très élevés aux arêtes qui pourraient traverser les frontières entre deux régions.

BIBLIOGRAPHIE

Aho, A., Hopcroft, J., et Ullman, J. (1974). *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, MA.

Aho, A., Hopcroft, J., et Ullman, J. (1985). *Data Structures and Algorithms*, Addison-Wesley, Reading, MA.

Edelsbrunner, H. (1987). *Algorithms in Combinatorial Geometry*, New York: Springer-Verlag.

Edelsbrunner, H. (1990). Communication personnelle.

Faloutsos, C., et Rong Y. (1989). Spatial access methods using fractals: Algorithms and performance evaluation, University of Maryland Computer Science Technical Report Series, UMIACS-TR-89-31, CS-TR-2214.

Faloutsos, C., et Roseman, S. (1989). Fractals for secondary key retrieval, University of Maryland Computer Science Technical Report Series, UMIACS-TR-89-47, CS-TR-2242.

Garey, M.R., et Johnson, D.S. (1979). *Computers and Intractability, A Guide to the Theory of NP-Completeness*, W.H. Freeman, New York.

Harary, F. (1969). *Graph Theory*, Addison-Wesley, Reading, MA.

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

Mark, D.M. (1990). Neighbor-based properties of some orderings of two-dimensinal space. *Geographical Analysis*, Avril, 22, 2, 145-157.

Preparata, F., et Shamos, M. (1985). *Computational Geometry, An Introduction*, New York: Springer-Verlag.

Saalfeld, A. (1990). Canonical cyclic orders for points in the plane, submitted to *Journal of Computational Geometry: Theory and Applications*, Elsevier.

Saalfeld, A. (1991). New proximity-preserving orderings for spatial data. *Proceedings of the Tenth International Symposium on Computer-Assisted Cartography (AutoCarto 10)*, ACSM/ASPRS, Baltimore, MD, 59-76.

Wolter, K., et Harter, R. (1989). Mise à jour des échantillons basée sur les valeurs de Peano, *Recueil du Symposium 1989 sur l'analyse des données dans le temps*, Statistique Canada, Ottawa, Canada, (Éds. A.C. Singh et P. Whitridge), 21-31.

Yao, A.C-C. (1982). On constructing minimum spanning trees in k-dimensional spaces and related problems, *SIAM Journal of Computing*, Novembre, 11, 4, 721-736.

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

AUTOMATISATION DU DÉVELOPPEMENT DE BASES DE SONDAGE ARÉOLAIRES UTILISANT DES DONNÉES NUMÉRIQUES AFFICHÉES SUR L'ÉCRAN D'UN POSTE DE TRAVAIL GRAPHIQUE

J.J. Cotter et C. Mazur[1]

RÉSUMÉ

Le National Agricultural Statistics Service utilise des bases de sondage aréolaires pour réaliser des enquêtes visant à recueillir des données agricoles aux États-Unis. Pour créer ces bases aréolaires, on répartit le terrain selon des catégories d'utilisation du sol puis on délimite des parcelles de terre à des fins d'échantillonnage. Actuellement, on utilise des produits sur support papier comme l'imagerie par satellite, la photographie à haute altitude et les produits cartographiques de la United States Geological Survey. Cet automne, une base de sondage sera créée à l'aide du système CASS (Computer Assisted Stratification and Sampling) qui nous permettra d'automatiser ce processus en affichant des données numériques obtenues par satellite et des données sous forme de cartes numériques sur l'écran d'un poste de travail graphique. Le nouveau système permettra au cartographe de créer une base de sondage plus précise, car ce dernier pourra accéder à des données plus récentes tout en nous permettant de compléter le travail beaucoup plus rapidement.

MOTS CLÉS: Stratification; échantillonnage à l'aide de bases aréolaires; imagerie numérique par satellite; graphique linéaire numérique.

1. INTRODUCTION

Depuis 1954, le National Agricultural Statistics Service (NASS) élabore, utilise et analyse des bases de sondage aréolaires comme véhicule pour réaliser des enquêtes visant à recueillir des renseignements sur la superficie en culture, sur le coût de production, sur les dépenses agricoles, sur le rendement céréalier et sur la production céréalière, sur le nombre d'animaux et sur d'autres données agricoles. Une base de sondage, dans le cas d'une région comme un État ou un comté, est composée d'une collection ou d'une liste de toutes les parcelles de terre dans la région qui nous intéresse. Ces parcelles de terre peuvent être définies en fonction de facteurs comme la propriété ou simplement basées sur des limites faciles à identifier, cette dernière méthode étant utilisée par le NASS. Les bases de sondage aréolaires fournissent une couverture complète, toutes les superficies de terrain étant représentées dans une enquête probabiliste avec une probabilité connue d'être choisies.

La procédure utilisée actuellement pour élaborer les bases de sondage est lente, à forte intensité de main-d'oeuvre et coûteuse. L'élaboration d'une base de sondage aréolaire pour un seul état peut exiger 11 000 heures de travail et coûter plus de $150 000.

Dans la présente communication, nous décrirons brièvement les documents et les procédures utilisés actuellement pour élaborer une base de sondage aréolaire. Viendra ensuite une description d'un projet de recherche, sur une méthode qui commence maintenant à être utilisée de façon opérationnelle, afin d'élaborer des bases de sondage aréolaires à l'aide de données numériques. Pour plus de renseignements sur l'élaboration de bases de sondage aréolaires, consultez la bibliographie.

---

[1] Les auteurs sont respectivement chef de la Technology Research Section, Survey Research Banch et chef d'équipe responsable de l'élaboration des bases de sondage aréolaires au sein de la Area Frame Section, Survey Sampling Branch, National Agricultural Statistics Service, 14e et Independence, S.W., bureau 4168-sud, Washington (DC) 20250, États-Unis.

2. BASES DE SONDAGE ARÉOLAIRES ÉTABLIES À PARTIR DE DOCUMENTS SUR SUPPORT PAPIER

2.1 Documents utilisés dans la procédure employée actuellement

Les bases de sondage aréolaires sont présentement élaborées État par État. Les documents utilisés dans le processus de stratification comprennent:

Une imagerie par satellite: l'imagerie par satellite est tirée des données numériques recueillies par des scanneurs à bord de satellites. Actuellement, nous utilisons l'imagerie sur support papier produite à partir des données recueillies par le satellite LANDSAT. Un scanneur à bord du satellite capte l'énergie reflétée et émise par le sol. Deux genres de scanneurs sont utilisés: un scanneur multibande (SMB) et un appareil de cartographie thématique (TM). Pour la stratification, nous préférons utiliser les produits obtenus à l'aide du TM. Les données du TM sont plus dispendieuses à cause de la meilleure résolution de l'appareil. L'échelle des produits sur support papier obtenus à partir des données recueillies par le TM est de 1:250000.

Le programme national de photographie aérienne (National Aerial Photography Program (NAPP)): Le NAPP est le produit d'un consortium d'organismes fédéraux qui ont besoin de photographies aériennes. Les épreuves par contact employées ont neuf pouces de côté et sont produites à l'échelle de 1:40000. Le NAPP est un outil de base pour la stratification. Presque toute la surface des États-Unis a été photographiée dans le cadre du programme NAPP.

Coupure topograhique (Quad): Ces cartes sont produites par la U.S. Geological Survey (USGS) et l'échelle préférée est celle de 1:24000 (série de 7.5 minutes - 2.6 pouces au mille) ce qui les rend utiles pour la stratification et l'échantillonnage dans des régions urbaines et dans des régions à la fois agricoles et urbaines.

Carte du Bureau of Land Management (BLM): Ces cartes, à l'échelle de 1:100000, montrent la répartition des terres fédérales et d'État. Elles sont utiles dans les États de l'Ouest pour délimiter les strates des parcours et pour situer les limites des réserves indiennes.

Carte de la U.S. Geological Survey au 1:100000: Ces cartes sont de haute qualité et fournissent au NASS un fond de carte précis qui peut être utilisé pour la stratification et la numérisation (deux termes qui seront définis plus loin).

Ces documents sont présentement tous utilisés sur support papier.

2.2 Stratification

Le processus de stratification selon l'utilisation des terres consiste à délimiter des surfaces en catégories d'utilisation des terres sur des photographies ainsi qu'à produire un fond de carte correspondant à l'aide de la couverture par satellite. Le tableau 1 montre l'ensemble de catégories d'utilisation des terres dont on s'est servi lors de l'élaboration de la base de sondage aréolaire du Missouri en 1987. La stratification a pour objet de réduire la variabilité d'échantillonnage en créant des groupes homogènes d'unités d'échantillonnage. Bien que certaines parties du processus soient de nature très subjective, les personnes qui s'occupent de la stratification du territoire (appelées stratificateurs) doivent effectuer un travail de précision afin qu'il n'y ait pas de chevauchement ou d'omission de superficie de terrain et que ce dernier soit stratifié de la façon appropriée.

Il se peut que le concept le plus important que l'on cherche à inculquer aux employés pendant leur formation initiale soit l'idée d'utiliser des limites de qualité. Une limite de qualité est un élément géographique permanent ou, à tout le moins, de longue durée, qu'un intervieweur peut trouver et identifier facilement. Si un intervieweur ne peut trouver avec précision un segment en temps opportun, il se peut que des erreurs non dues à l'échantillonnage soient introduites dans les données recueillies lors d'une enquête. Si, inconsciemment, l'intervieweur sur le terrain ne recueille pas les données associées à toute la superficie à l'intérieur de la région échantillonnée ou s'il recueille des données pour une superficie qui se trouve à l'extérieur de la région sélectionnée, les résultats de l'enquête seront biaisés.

Tableau 1. Codes et définitions des strates employés pour l'utilisation des terres

CODE DE STRATE	DÉFINITION	TAILLE-CIBLE	
		milles^2	kilomètres^2
11	Terre labourable générale avec 75% ou plus de la superficie cultivée.	6-8	15.5-20.7
12	Terre labourable générale avec de 50 à 74% de la superficie cultivée.	6-8	15.5-20.7
20	Terre labourable générale avec de 15 à 49% de la superficie cultivée.	6-8	15.5-20.7
31	Région à la fois agricole et urbaine avec moins de 15% de la superficie cultivée, on y trouve plus de 100 logements au mille carré, région résidentielle combinée à de l'agriculture.	1-2	2.6-5.2
32	Région résidentielle/commerciale, aucune culture, plus de 100 logements au mille carré.	.5-1	1.3-2.6
40	Parcours et pâturage, avec moins de 15% de la superficie cultivée.	12-16	31.1-41.4
50	Région non agricole, segments de taille variable.		
62	Eau.		

Il arrive souvent, qu'en pratique, l'objectif d'utiliser des limites permanentes et celui d'obtenir des unités d'échantillonnage homogènes dans une strate s'opposent lors de la stratification d'une base de sondage aréolaire. Il se peut qu'on doive faire des concessions dans des cas limites. Compte tenu du fait que la base de sondage aréolaire sera utilisée pendant 15 à 20 ans et qu'elle représente un investissement important, on doit utiliser les meilleures limites et celles qui sont les plus durables. Les routes ainsi que les fleuves et les rivières constituent de bonnes limites de strates, alors que ce n'est pas le cas pour les cours d'eau intermittents et les lisières des champs qu'on ne devrait donc utiliser que rarement. La liste ci-après renferme les éléments géographiques les plus utilisés pour représenter des limites de strates. Les éléments dans la liste sont classés en fonction de leur qualité, de la plus élevée à la plus faible:

- Routes pavées.
- Routes secondaires toute saison.
- Routes locales reliant la ferme au marché.
- Chemins de fer.
- Fleuves, rivières et cours d'eau permanents.

À des fins administratives, la stratification est effectuée comté par comté. On attribue un comté à chaque stratificateur et cette personne travaille sur ce comté jusqu'à ce que le traitement de ce dernier soit terminé. On commence généralement la stratification en déterminant les strates urbaines et agricoles-urbaines pour le comté. Les régions agricoles sont ensuite stratifiées à l'aide de l'imagerie par satellite obtenue par l'appareil de cartographie thématique (TM). L'imagerie est utilisée principalement pour vérifier où se trouvent les régions cultivées et les régions non cultivées dans un comté. L'imagerie est très utile lors de la stratification de par son actualité. Bien que les photographies aériennes puissent avoir été prises de une à cinq années auparavant, l'imagerie Landsat porte habituellement sur la dernière saison de végétation, fournissant une vue très récente de la région. Utilisant l'imagerie Landsat pour repérer les cultures ainsi que les pâturages et les photographies pour trouver les limites, le stratificateur doit prendre des décisions subjectives pour placer les régions dans leurs strates respectives.

L'assurance de la qualité est une préoccupation importante au cours de la phase de la stratification. Pendant tout le processus, on vérifie et revérifie le travail afin de s'assurer d'obtenir un produit de haute qualité et pour tirer profit d'une seconde opinion fournie par une personne plus expérimentée.

Une fois que la stratification sur les photographies a été examinée et approuvée, les limites des strates seront transférées sur la carte, recréant le fond de carte. Cette carte sera ensuite numérisée (mesurée

électroniquement) afin de déterminer les superficies des unités primaires d'échantillonnage (expression qui sera définie plus loin). Une fois ce transfert terminé, le vérificateur en fait l'examen et on commence la phase suivante de la stratification...la construction des unités primaires d'échantillonnage.

2.3 Construction des unités primaires d'échantillonnage

L'étape suivante de l'élaboration de la base de sondage aréolaire consiste à diviser les strates en unités primaires d'échantillonnage (UPÉ). La taille désirée pour les UPÉ varie selon la strate, mais elles renferment, en moyenne, de six à huit unités finales d'échantillonnage ou segments. La taille minimale d'une UPÉ est généralement d'un segment. L'utilisation des unités primaires d'échantillonnage permet de réaliser des économies quand on effectue un échantillonnage à l'aide d'une base de sondage aréolaire. Il n'est pas nécessaire de diviser une base de sondage entière en segments afin de choisir un échantillon. Les strates sont divisées en UPÉ et l'on choisit un échantillon aléatoire d'UPÉ. Seules les UPÉ choisies aléatoirement seront encore divisées pour obtenir des segments - ce qui permet une économie considérable au niveau des coûts de la main-d'oeuvre. Quand on délimite des UPÉ, on ne se concentre pas sur l'homogénéité de l'utilisation des terres - cela aura déjà été réalisé lors de la stratification en fonction de l'utilisation des terres. On vise surtout à obtenir la taille désirée avec de bonnes limites tout en essayant de faire en sorte que chaque UPÉ constitue une représentation réduite de la strate dans son ensemble.

Comme dernière vérification, les fonds de carte sont examinés par un statisticien afin de s'assurer qu'ils sont complets. Les polygones créés en traçant les UPÉ sont examinés afin de s'assurer qu'ils forment une figure fermée. Le système de numérotation est vérifié tant du point de vue de l'exactitude de l'identification des strates que pour s'assurer que la numérotation est suivie. Les fonds de carte sont encore vérifiés pour s'assurer qu'il n'existe pas d'omission ni de chevauchement. Une fois que ces vérifications ont été effectuées, les fonds de carte sont prêts pour l'étape suivante du processus...la mesure des UPÉ.

2.4 Numérisation

La conversion des points de la carte en coordonnées à deux dimensions (abscisses et ordonnées) s'appelle numérisation. Cette opération comprend la mesure électronique de la superficie des UPÉ sur les fonds de carte. Il faut mesurer la superficie des UPÉ afin de déterminer le nombre de segments par UPÉ à des fins d'échantillonnage. L'enregistrement électronique de la superficie des UPÉ permet:

- de mesurer avec précision la superficie des UPÉ,
- de s'assurer de la qualité,
- de conserver une copie de sauvegarde numérique du fond de carte au cas, peu vraisemblable, où un de ces fonds de carte serait perdu.

Le NASS utilise des tablettes de conversion analogique-numérique (des tables à numériser de 4 pi. x 5 pi.) pour établir un système de coordonnées qui recouvre le fond de carte. Un point de référence, connu sous le nom d'origine (0,0), est établi pour les coordonnées (abscisse, ordonnée) (ou (X,Y)) sur la carte. Ces coordonnées, identifiées de la façon appropriée, décrivent de façon unique les limites d'une UPÉ et créent, par conséquent, un polygone pour chaque UPÉ. Le logiciel de numérisation enregistre les coordonnées dans un fichier. À l'aide de l'échelle de la carte, la superficie de chaque polygone (UPÉ) du comté est calculée en milles carrés et enregistrée dans un fichier pour ce comté.

La superficie des UPÉ dans chaque comté est totalisée et comparée à la superficie officielle du comté. La même procédure est appliquée pour la superficie de l'État. On accepte que la superficie des comtés varie de jusqu'à 3.0 pour cent par rapport à leur superficie officielle. Le total des superficies pour l'État ne peut varier de plus de 0.5 pour cent par rapport à la superficie officielle de l'État. La superficie d'un comté peut varier plus que celle d'un État parce que l'on a affaire à de plus petites régions et que les unités primaires d'échantillonnage peuvent chevaucher les limites de comtés. Puisqu'on ne permet jamais à la stratification de chevaucher les frontières entre les États, on tolère seulement une petite erreur dans ces cas.

La superficie des UPÉ est alors totalisée pour chaque strate au niveau de l'État. La superficie de l'UPÉ divisée par la taille cible d'un segment pour la strate est égale au nombre total d'unités d'échantillonnage ultimes

(segments) dans cette UPÉ (arrondi à l'entier le plus rapproché). L'addition du nombre de segments donnera le nombre total de segments dans la strate. Ce renseignement sera utilisé pour déterminer le nombre d'échantillons à choisir pour tout l'État.

2.5 Sélection de l'échantillon

Une fois que l'on a déterminé le nombre total de segments à échantillonner dans un État, on fait passer un programme pour choisir les UPÉ qui seront subdivisées en segments à échantillonner. Chaque segment a une taille cible particulière selon la strate à laquelle il est associé de sorte que chaque segment ressemble étroitement à toute l'UPÉ (dans la mesure du possible) avec les meilleures limites physiques disponibles. On localise les UPÉ sélectionnées sur le fond de carte et on transfère leurs limites sur les photographies. Les UPÉ sélectionnées sont ensuite subdivisées selon des limites identifiables pour obtenir le nombre requis de segments. Ces derniers sont numérotés manuellement et on choisit un nombre aléatoire entre un et le nombre de segments dans l'UPÉ. Le segment correspondant au nombre aléatoire choisi est celui qui est sélectionné. Un agrandissement photographique de ce segment sera envoyé plus tard à l'intervieweur chargé du dénombrement.

3. BASES DE SONDAGE ARÉOLAIRES NUMÉRISÉES

3.1 Historique de la recherche

Dans le cadre d'un accord de collaboration, le NASS a travaillé avec la National Aeronautics and Space Administration (NASA), l'agence spatiale américaine, à l'élaboration de bases de sondage aréolaires à l'aide de données numériques. Le projet avec la NASA a commencé il y a trois ans. Bien que l'accord de recherche initial avec cet organisme prenne fin à l'automne de 1991, la NASA continuera à fournir un soutien logiciel dans le cadre d'un accord de collaboration avec la Ecosystem Science and Technology Branch (ECOSAT), un groupe dont les bureaux sont situés au Ames Research Center, Moffett Field, Californie.

Le NASS et ECOSAT ont collaboré depuis longtemps à la recherche en télédétection. Les deux organismes ont travaillé ensemble à un certain nombre de projets depuis la fin des années 70. EDITOR, un progiciel pour l'estimation de grandes superficies en culture qui utilise les données du scanneur multibande du LANDSAT, et des données de terrain connexes a été rédigé par des employés du NASS et de ECOSAT. Quatre-vingts pour cent de PEDITOR, une version portable de EDITOR, c'est-à-dire, un système pouvant être mis en application sur diverses machines, a été rédigé à ECOSAT. PEDITOR est le principal outil logiciel utilisé pour le travail opérationnel en télédétection effectué au NASS. ECOSAT a assemblé, pour le NASS, un prototype de poste de travail articulé autour d'un microprocesseur, appelé MIDAS, et a aidé à réaliser une expérience visant à déterminer les caractéristiques de rendement du système quand ce dernier produit des estimations de superficies dans l'environnement opérationnel du NASS. ECOSAT a créé, pour le NASS, un logiciel d'affichage compatible avec PEDITOR et un précurseur du système nécessaire pour réaliser l'élaboration de bases de sondage aréolaires.

L'élaboration du logiciel utilisé pour réaliser des bases de sondage aréolaires a été un processus itératif. Le logiciel fourni au NASS par ECOSAT a fait l'objet d'une évaluation constante afin de s'assurer qu'il était complet et facile à utiliser par le personnel de la Area Frame Section du NASS. Cette évaluation par les étudiants et les surveillants qui utilisent le système a entraîné l'établissement de relations de travail étroites avec les employés de la NASA. Les suggestions pour les améliorations et les modifications sont appliquées très rapidement.

Un nouveau système pour la production de bases de sondage aréolaires appelé le système CASS (Computer Assisted Stratification and Sampling), basé sur les concepts de PEDITOR et intégré à ce dernier renforcera les programmes de recherche en télédétection et d'application opérationnelle de cette technologie au NASS et à la NASA ainsi que l'élaboration de bases de sondage aréolaires. Un avantage particulier qu'offre cette méthode est la possibilité d'utiliser des renseignements numériques, sur l'utilisation des terres, tirés des enquêtes d'années antérieures pour aider à l'élaboration ou à la mise à jour de bases de sondage aréolaires. D'autres avantages seront mentionnés plus loin.

3.2 Le système CASS

Le système CASS incorpore des données numériques provenant de l'appareil de cartographie thématique (TM) des satellites LANDSAT et des données de graphiques linéaires numériques (Digital Line Graph (DLG)) de la USGS. Les données du TM (à l'échelle de 1:100000) servent de base pour délimiter l'utilisation des terres selon notre plan de stratification. Pour l'identification des limites, les données DLG à l'échelle de 1:100000 sont superposées à l'image du TM à l'aide d'un plan graphique.

Affichage des données de satellite - Trois des sept bandes des images Landsat sont chargées dans le système pour affichage avec une résolution de 30 mètres. Une fonction zoom peut faire varier l'échelle de l'image sur l'écran. Cette image prise par satellite fournit la vue la plus récente de la région à stratifier. Ce produit numérique est l'équivalent de l'utilisation de l'imagerie satellite sur support papier ainsi que de la photographie à haute altitude en noir et blanc.

Cartographie en couleurs - La fenêtre d'affichage utilise 8 bits pour chacune des trois bandes ce qui donne une image nette et colorée. Une carte en couleurs est une carte pour laquelle on attribue des degrés de luminosité à chacune des bandes. Une fonction de cartographie dynamique permet à l'utilisateur de régler la luminosité et le contraste de chacun des trois plans. Ces réglages peuvent alors être conservés dans un fichier pour usage ultérieur. Plusieurs autres fonctions permettent à l'utilisateur de voir une seule bande, de voir un histogramme de chaque bande et de faire une cartographie linéaire, par morceaux, ou équiprobable. La cartographie optimale est celle qui permet de distinguer le mieux les parties cultivées et les limites. Pour vérifier cette situation, on repère sur l'écran des champs utilisés lors d'enquêtes antérieures et on en relève les couleurs.

Affichage et repérage des données DLG - Les données DLG sont utilisées comme référence. Les stratificateurs doivent être certains qu'ils utilisent de bonnes limites pour les étapes de délimitation des UPÉ et des segments. Les données numériques utilisées actuellement comprennent le réseau de transport et le réseau hydrographique. Les limites politiques et administratives seront disponibles sous peu. Les fichiers DLG originaux sont produits pour une unité appelée un "panel" (dont la dimension est de 7.5 ou 15 minutes). Une fonction du système CASS lit ces fichiers à partir d'une bande et crée des fichiers sur disques. Ces derniers fichiers peuvent alors être combinés et placés dans un fichier pour couvrir une superficie plus grande, comme un comté. Si un comté chevauche des zones de la projection transversale universelle de Mercator (UTM), les données peuvent être converties afin qu'on puisse utiliser un seul fichier de comté. (On ne peut combiner des "panels" provenant de zones différentes.) Une autre fonction du système CASS affiche le fichier de données sur l'écran. Afin de recouvrir avec précision l'image obtenue par l'appareil de cartographie thématique avec les données DLG, on utilise un autre programme pour faire le repérage des données DLG avec la toile de fond des données de satellite en choisissant plusieurs points de coïncidence dans chaque ensemble de données et en faisant passer un programme de régression utilisant la méthode des moindres carrés pour ajuster le reste des données. Ces points ainsi que les résultats de la régression sont aussi conservés dans un fichier et utilisés chaque fois que ce fichier DLG particulier est affiché. Finalement, ce fichier de repérage permet à l'utilisateur de déterminer les coordonnées (latitude et longitude) d'un point donné.

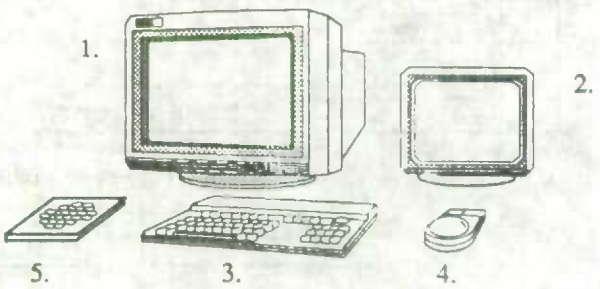
Délimitation des UPÉ - Dans chaque comté, on tracera des polygones et on leur attribuera le numéro d'UPÉ approprié, ce dernier numéro étant composé d'un numéro de strate et d'un numéro d'ordre. Pour ce faire, on détermine à quelle strate d'utilisation des terres une parcelle de terrain appartient en interprétant l'affichage couleur des données recueillies par l'appareil de cartographie thématique (TM). En même temps, on délimite une UPÉ à l'intérieur d'une certaine superficie, à l'aide de limites physiques déterminées au moyen des données DLG et (ou) TM. Dans le système CASS, on obtient ce résultat en introduisant le numéro de l'UPÉ, puis en utilisant la souris pour désigner des points le long des limites désirées. Quand le polygone qui forme l'UPÉ est fermé, sa superficie est immédiatement calculée et affichée. Cela permet au stratificateur de déterminer si la superficie de l'UPÉ se trouve dans les limites de superficie cibles pour cette strate. Si le polygone est trop petit ou trop grand, on peut combiner, séparer ou refaire des polygones. Quand la stratification d'un comté est terminée, les polygones sont conservés dans un fichier de stratification pour être examinés, par la suite, par un stratificateur expérimenté. L'utilisateur peut vérifier s'il y a chevauchement de polygones ou s'il existe des trous (ou des superficies omises). En tout temps, l'utilisateur peut obtenir la liste des UPÉ qui ont été créées jusqu'à ce moment afin de vérifier la numérotation et les superficies.

Décomposition des UPÉ en segments - Après que tout l'État a été stratifié et que l'on a calculé la superficie totale de chaque strate, on fait passer un programme distinct afin de tirer l'échantillon d'UPÉ (première phase de l'échantillonnage) qui seront décomposées par la suite en unités finales d'échantillonnage ou segments. Toutes les UPÉ ne sont pas décomposées - seulement celles qui ont été choisies à l'aide du programme de sélection de l'échantillon. L'utilisateur fait afficher le fichier de stratification (sauvegardé lors de l'étape précédente) et introduit le numéro de l'UPÉ dans laquelle on doit tirer un échantillon. Le logiciel efface ensuite toutes les UPÉ à l'écran sauf celle dans laquelle on doit tirer l'échantillon. Un bon nombre des fonctions qui ont été utilisées pour délimiter les UPÉ au cours de l'étape de la stratification sont utilisées pour diviser l'UPÉ en segments de même superficie. Des vérifications de contrôle de la qualité semblables sont effectuées. Quand l'UPÉ a été décomposée en segments, on peut choisir, de façon aléatoire, un segment à l'aide de la commande de sélection de segments. Cela constitue la deuxième phase de l'échantillonnage.

3.3 Le poste de travail du système CASS

Le poste de travail du système CASS (voir la figure 1) comprend plusieurs éléments. L'écran d'affichage (1) est utilisé pour afficher l'image Landsat en couleurs, alors que l'écran des menus (2) est employé pour afficher des éléments relatifs au logiciel. Le clavier (3) sert à introduire les commandes sur l'écran des menus, alors que la souris (4) est utilisée pour "dialoguer" avec l'écran d'affichage. Le bloc de touches (5) est aussi employé avec l'écran d'affichage pour effectuer le traitement des plans de recouvrement (pour modifier la couleur de l'affichage ou afin d'activer ou de désactiver l'affichage) et pour faire des zooms avec l'image.

Figure 1.
Poste de travail du système CASS



Actuellement, nous employons un poste de travail Hewlett-Packard (HP) qui utilise UNIX, pour répondre aux besoins de traitement et de stockage de la quantité considérable de données à utiliser. Les postes de travail HP possèdent les fonctions minimales pour l'élaboration de bases de sondage aréolaires, c'est-à-dire, trois plans d'image, quatre plans de recouvrement et un système d'affichage à coordonnées d'une résolution de 1 024 x 1 280 pixels. La configuration permet d'utiliser trois bandes de données de satellite dans les plans d'image tout en utilisant les plans graphiques pour afficher les DLG, les UPÉ ainsi que le menu des commandes. Une modification récente permet d'afficher les DLG dans plus d'un plan de recouvrement. Une des raisons qui justifient cette modification est qu'elle permet de colorer de façon distincte les données sur le réseau de transport et celles relatives à l'hydrographie.

3.4 L'essai initial

L'essai initial avait pour but de nous permettre d'acquérir une expérience de base dans l'utilisation du logiciel, de comparer le système CASS à la méthode utilisée actuellement et de déterminer la rapidité de la construction des bases de sondage. Nous avons utilisé des données numériques pour trois comtés du centre nord du Missouri (Linn, Livingston et Macon). Ces comtés ont été choisis en partie parce que la Area Frame Section avait élaboré une nouvelle base de sondage aréolaire pour le Missouri en 1987, base qui devait être utilisée en 1988 et parce que les données numériques étaient disponibles. Voici certains des résultats obtenus :

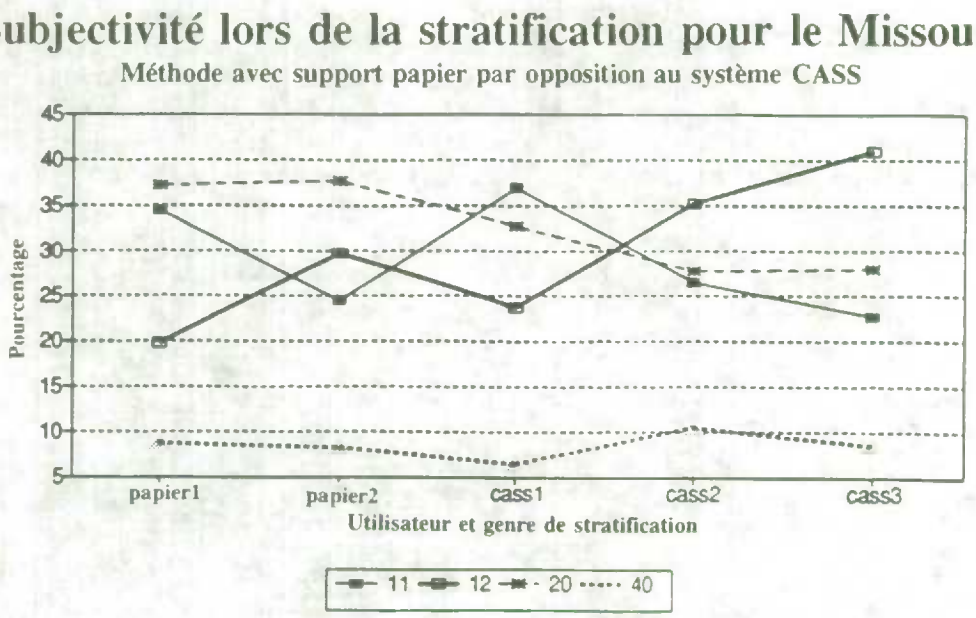
1) Cet essai a démontré qu'il était possible d'effectuer une stratification à l'aide du système CASS.

2) Le système CASS s'est révélé plus rapide puisqu'il faut entre 2.5 et 3 semaines pour effectuer une stratification avec la méthode utilisée actuellement et de 2 à 3 jours avec CASS.

3) Plusieurs améliorations ont été apportées au logiciel. C'est à peu près à ce moment que nous avons décidé d'ajouter des modules pour effectuer la sélection des échantillons.

4) Il faut remarquer un point important, soit la nature subjective du travail. Pour cet essai, la stratification des mêmes comtés a été effectuée par 3 personnes avec le système CASS et par 2 personnes avec les documents sur support papier. (Voir la figure 2.)

Figure 2.



5) C'est à peu près à ce moment que nous sommes passé d'un poste de travail SUN à un poste de travail Hewlett-Packard (HP). La taille de l'écran a augmenté d'environ 512 à 1 024 pixels. Le nombre de "superpositions" est passé de 2 à 4 et l'emploi du poste de travail HP a permis de charger toute une scène de l'appareil de cartographie thématique (TM) en mémoire plutôt que seulement un écran.

3.5 L'essai élargi

La phase suivante de la recherche consistait à simuler un environnement opérationnel. Une région plus considérable a été choisie au Michigan (21 comtés) et il a fallu tenir compte de la configuration exigée. Cette région a été choisie tout d'abord parce qu'on venait de produire une nouvelle base de sondage pour cet État en 1989 (elle est entrée en vigueur en 1990) et ensuite que la Remote Sensing Section du NASS avait, peu de temps auparavant, effectué des travaux relatifs à la classification sous surveillance dans la région du Michigan où l'on récolte des haricots secs (les données étaient donc disponibles). Dans cet essai, une seule personne a travaillé sur chaque comté. Comme la stratification à l'aide des documents sur support papier avait été effectuée très peu de temps auparavant, il se peut qu'un biais existe dans la stratification effectuée avec le système CASS.

3.6 Analyse

Cette analyse est basée sur les résultats de l'essai élargi qui portait sur les 21 comtés choisis au Michigan. Les évaluations sont de nature à la fois quantitative et qualitative.

Dans le tableau 2 on compare, pour l'étude sur le Michigan, la superficie totale stratifiée au moyen du système CASS à celle qui a été stratifiée à l'aide de la méthode manuelle utilisée actuellement. L'écart en pourcentage (exprimé sous forme du rapport essai opérationnel de CASS/mode opérationnel) est aussi indiqué. On trouve des différences pour la strate 12 ainsi que pour la strate 31 avec un écart plus considérable pour la strate 40. Les commentaires des stratificateurs favorisent généralement le système CASS pour l'écart relevé au niveau de la strate 40 car ils pensent avoir été mieux en mesure de trouver les régions boisées à l'aide de ce système.

Tableau 2. Comparaison de la superficie totale par strate pour les 21 comtés de l'étude sur le Michigan. La superficie est exprimée en milles carrés.

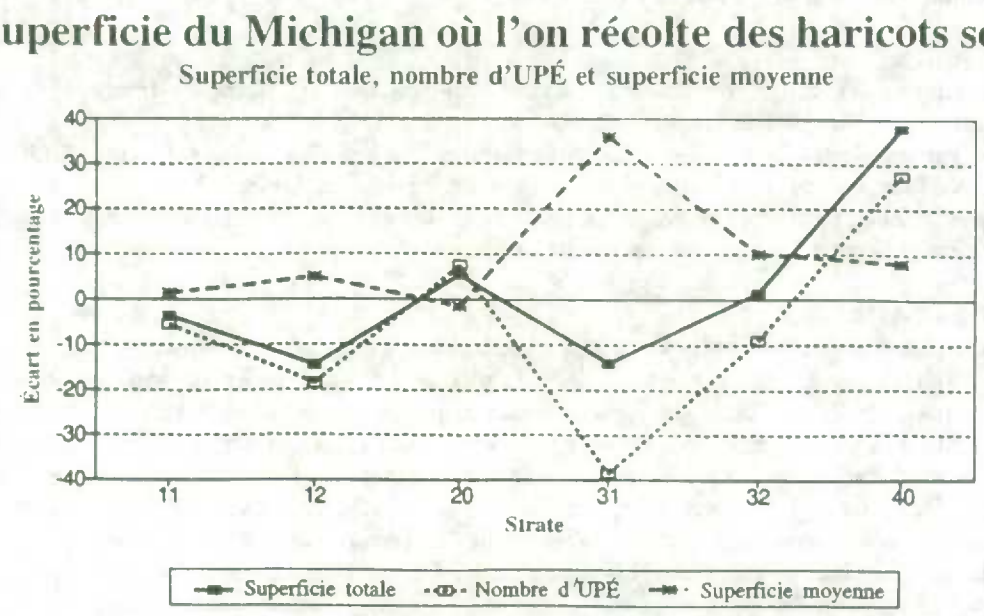
Strate	Mode opérationnel	CASS	Écart en pourcentage
11	5 427.4	5 225.8	-3.7
12	2 547.8	2 175.6	-14.6
20	3 264.0	3 448.0	5.6
31	648.0	555.3	-14.3
32	168.4	170.7	1.4
40	1 347.3	1 856.2	37.8

Dans le tableau 3 on décrit la taille moyenne des UPÉ par strate sous forme d'une comparaison entre les résultats obtenus en mode opérationnel et avec le système CASS.

Tableau 3. Taille moyenne des UPÉ selon la strate pour les 21 comtés de l'étude sur le Michigan. La superficie est exprimée en milles carrés.

Strate	Mode opérationnel	CASS	Écart en pourcentage
11	6.4	6.5	1.7
12	5.3	5.6	4.9
20	6.1	6.0	-1.5
31	1.2	1.6	35.8
32	0.6	0.7	10.0
40	9.7	10.5	8.1

Figure 3.



Il est intéressant d'examiner les données quantitatives (voir la figure 3), mais l'aspect qualitatif de l'étude porte sur des questions plus instructives. Quand l'essai pour le Michigan a été terminé, les utilisateurs ont comparé les bases de sondage aréolaires obtenues par la méthode qui emploie des supports papier et par le système CASS pour cinq comtés du Michigan. Les questions soulevées, portent sur l'affichage numérique, la stratification et la sélection d'échantillons.

Quelques améliorations à l'affichage numérique aideront l'utilisateur à effectuer un meilleur travail la prochaine fois.

a) APPAREIL DE CARTOGRAPHIE THÉMATIQUE (TM)

- On a déterminé que l'affichage de l'image prise par satellite pouvait être amélioré. Au début, les niveaux de grossissement ("zoom") étaient fixés à 1, 2, 4, 8 et 16. À la fin de l'essai sur le Michigan, on les a remplacés par les niveaux 1, 2, 3, 4 et 5.
- Au début, chaque stratificateur créait une carte en couleurs pour le comté sur lequel il travaillait. Cela a créé des difficultés lors de l'examen car chaque carte en couleurs produite par une personne différait de celles produites par les autres. On a donc décidé d'essayer, à l'avenir, une carte en couleurs par scène LANDSAT. De plus, les stratificateurs ne connaissaient pas certaines des techniques utilisées en cartographie, ce qui aurait pu aider dans l'interprétation des données LANDSAT.

b) DLG

- Les données graphiques linéaires numériques ont été introduites dans un seul plan de recouvrement. Ce n'est qu'à la fin de l'essai qu'une modification a été apportée pour afficher ces données dans des plans différents. De cette façon, on peut afficher les routes dans un plan (dans une couleur) et l'eau dans un autre plan (dans une autre couleur).
- De plus, l'affichage des lignes a été amélioré. À cause d'une erreur, les lignes ne conservaient pas les points et les tirets quand l'image était affichée de nouveau ou quand on en faisait un "zoom". Cette erreur a été corrigée.
- De plus, les points marqués par l'utilisateur paraîtront plus gros au niveau d'agrandissement 2 ou supérieur.
- Finalement, les lignes "semblent" toutes les mêmes dans un plan de recouvrement donné. Par conséquent, au début, les utilisateurs ont employé des limites interdites (comme des limites 3z3z pour les marécages) qui "semblaient" identiques à des limites permises. On a alors rédigé un programme pour supprimer toutes les limites interdites des fichiers DLG. Toutefois, on travaille actuellement à tenter de mieux afficher les données DLG afin que l'utilisateur puisse "reconnaître" le genre particulier de limite.

Voici quelques commentaires à propos de la stratification:

a) EXPÉRIENCE

- L'interprétation des couleurs était une nouvelle expérience et il faudra du temps avant que les stratificateurs deviennent très compétents dans ce travail.
- De plus, sur les support papier, on délimite des strates entières puis on les subdivise en UPÉ. Avec CASS, on détermine les strates mais on les délimite au niveau des UPÉ.
- Rétrospectivement, plusieurs personnes ont dit qu'elles auraient peut-être fait un meilleur travail si elles avaient eu alors l'expérience qu'elles possèdent maintenant.

b) STRATIFICATION

- En général, les stratificateurs pensaient que la stratification des superficies cultivées se faisait mieux avec CASS. Pour ce qui est des régions urbaines, la méthode qui utilise des supports papier donne de meilleurs résultats à cause de la résolution accrue des photographies à haute altitude. Sur papier, on peut quelque peu voir les maisons mais avec CASS le seul élément qui nous guide vers les villes est le rapprochement des routes. La section "Éléments qui devront être pris en considération" renferme quelques idées pour régler ce problème. Finalement, la superficie recouverte d'eau était plus détaillée sur les cartes sur support papier mais, en affichant une bande, l'utilisateur peut voir clairement les superficies recouvertes d'eau et les DLG montrent les fleuves, les rivières et les autres cours d'eau.

c) LIMITES/NUAGES

- Pour ce qui est des limites, les utilisateurs estimaient que la méthode utilisée actuellement était légèrement supérieure. Cela est dû en partie à la résolution des données.
- Dans deux cas, les nuages ont causé des difficultés. Dans un cas, un nuage a entièrement caché une parcelle de terrain, ce qui a entraîné sa mauvaise classification (elle aurait dû faire partie de la strate 11). Dans le deuxième cas, on n'a pas vu une île assez grosse à cause d'un nuage mince qui la cachait entièrement. Si le nuage est mince, il se peut qu'une combinaison différente de bandes nous permette de "voir" à travers lui.

On effectue présentement une analyse de la partie de CASS qui sert à la sélection des échantillons. Nous ferons une comparaison à l'aide de la superficie des segments, de l'homogénéité des UPÉ et du genre de limites utilisées. Le résultat important était la distinction (au niveau du logiciel) entre les processus de stratification et de sélection des échantillons. Les premières tentatives de sélection d'échantillons ont pris pas mal de temps à cause de difficultés au niveau de l'affichage. On peut améliorer cette situation en travaillant avec de plus petits fichiers de données de l'appareil de cartographie thématique (TM) (plutôt que d'utiliser toute une scène LANDSAT), en automatisant la détermination de l'emplacement des UPÉ et en travaillant à différents niveaux d'agrandissement. On peut aussi améliorer la vitesse du traitement, à long terme, à l'aide d'une interface utilisateur graphique (voir la section "Éléments qui devront être pris en considération") et, à court terme, en modifiant les commandes qui peuvent être utilisées avec la souris et celles qui doivent être introduites au terminal employé avec les menus. Finalement, le choix des bandes peut permettre d'améliorer la résolution de l'image produite par l'appareil de cartographie thématique (TM); les satellites qui seront lancés au cours des années à venir produiront aussi des images avec une meilleure résolution. Il faut remarquer un point important, soit le fait que les limites sont plus importantes pour ce processus puisque les polygones sont beaucoup plus petits.

3.7 Avantages qui découlent de l'utilisation du système CASS

Dans tout projet, comme CASS, on peut tenir compte de plusieurs composantes de la qualité. Ce sont: l'exactitude, les ressources, l'actualité et la pertinence (Statistical Policy Working Paper 20, 1991, traduction libre).

Au niveau de l'exactitude, nous pouvons penser à plusieurs avantages:

a) Les données de satellite sont plus récentes que celles obtenues à partir des photographies à haute altitude (ce qui est important quand une base de sondage est utilisée pendant 15 ans) et leur résolution est supérieure (1:100000 plutôt que 1:250000) à celle de l'imagerie satellite sur support papier. Cela permet d'effectuer une meilleure stratification pour l'agriculture.

b) L'opération ennuyeuse et sujette à erreur qui consiste à transférer des renseignements de l'impression d'une image prise par satellite à une photographie à haute altitude, puis sur une carte à l'échelle de 1:100000 de la USGS et finalement dans un fichier numérique peut être éliminée.

c) Il est plus facile de réviser les UPÉ avec CASS que lorsque les données sont sur support papier.

d) Le logiciel choisit les segments échantillonnés à l'aide d'un générateur de nombres aléatoires, on n'a donc pas à consulter une série de feuilles de nombres aléatoires.

e) Les limites, la stratification urbaine et la couverture nuageuse constituent des problèmes. Les données DLG ne valent que jusqu'à la date de leur création ou de leur révision. Il est plus difficile de repérer les limites sur les images LANDSAT que sur les photographies. Dans les cas où la photographie en noir et blanc est plus récente que les cartes au 1:100000, on pourrait déterminer les modifications dans les limites. De plus, on ne peut bien voir les maisons quand on utilise les données LANDSAT disponibles actuellement. On trouvera dans la section "Éléments qui devront être pris en considération" certaines façons d'améliorer ces situations.

Les ressources et l'actualité vont de pair. Avec CASS, nous pouvons traiter plus d'États avec le même nombre d'employés ou traiter le même nombre d'États en employant moins de personnes. Le NASS préfère la deuxième option, puisqu'on peut réaffecter les employés dans d'autres services qui ont besoin de personnel.

a) On peut procéder à une réduction des effectifs car, avec le système CASS, le processus de stratification qui emploie beaucoup de main-oeuvre est automatisé.

b) On peut procéder à une réduction des effectifs, car le processus de numérisation (afin de calculer les superficies et pour servir de copie de sauvegarde) n'est plus requis.

c) En général, l'emploi de CASS prend moins de temps que le processus utilisé actuellement. Il faut environ deux semaines pour traiter un comté avec des documents sur support papier, mais seulement de deux à trois jours pour effectuer le même travail à l'aide de CASS.

d) La possibilité de déterminer immédiatement la superficie d'une UPÉ devrait aider à réduire le travail additionnel relié au choix des échantillons.

e) L'argent est une ressource importante. Au début, le coût du matériel sera élevé, puis il diminuera de façon marquée. Le coût du matériel augmentera à cause de l'achat de données numériques produites par l'appareil de cartographie thématique (TM) plutôt que d'imagerie TM. Les coûts salariaux devraient diminuer considérablement à cause de la réduction de l'effectif et du fait qu'il faut moins de temps pour effectuer le travail. On ne dispose pas encore de détails à ce sujet.

La pertinence de la méthode s'évalue à l'emploi, par l'utilisateur final, de bases de sondage et de segments échantillonnés.

a) L'aspect numérique de la base de sondage permettra d'en effectuer la mise à jour plutôt que d'avoir à recommencer à zéro (ce qu'on doit faire actuellement avec les documents sur support papier parce que ces derniers ne sont plus à jour et qu'il est impossible d'effacer les traits en couleur).

b) On peut déterminer la position (latitude et longitude) des segments échantillonnés et introduire ces renseignements dans la base de données d'un système d'information géographique (SIG).

c) On peut étudier plus facilement des bases de sondage aréolaires spécialisées. La Remote Sensing Section du NASS peut fournir une imagerie satellite classée selon la culture pour aider à l'élaboration de bases de sondage aréolaires spécialisées.

3.8 Éléments qui devront être pris en considération

Nous espérons améliorer CASS de plusieurs façons. Ces améliorations porteront sur le processus de stratification, sur la détermination des limites et sur d'autres aspects.

Le processus de stratification peut être amélioré de plusieurs façons:

a) À ce jour, nous n'avons utilisé que les bandes 2, 3 et 4 à des fins d'affichage. L'accès à d'autres bandes pourrait nous permettre de voir plus de détail au niveau des couleurs et même d'être en mesure de voir à travers les nuages.

b) Nous étudions l'utilisation de classifications, effectuées ou non sous surveillance, de données de satellite et espérons utiliser de telles classifications comme moyens d'appuyer la stratification.

c) Nous étudions la possibilité d'utiliser des données TIGER du Census Bureau et du fichier "Public Law" correspondant pour localiser les limites d'un îlot de recensement dans le système CASS et pour trouver le nombre de ménages afin de mieux déterminer la classification des terrains urbains. Il se peut que nous puissions utiliser des coupures topographiques (Quad) pour déterminer la classification urbaine ainsi que les limites appropriées.

d) Un fonctionnaire du U.S. Department of Agriculture (USDA) étudie la possibilité de grouper les achats de données LANDSAT effectués par tous les organismes qui font partie du USDA. Le fait de partager les données devrait être avantageux pour tous ces organismes et il se peut que, pour le système CASS, on profite de la possibilité de disposer d'une scène LANDSAT prise à des dates différentes.

Le processus de sélection des limites revêt une importance capitale et on peut l'améliorer de plusieurs façons:

a) L'utilisation d'autres bandes (particulièrement la bande 5) devrait fournir des données plus détaillées.

b) Nous étudions l'utilisation de filtres afin d'être plus en mesure de détecter les "bords" ou les limites.

c) Quand les données de Landsat 6 deviendront disponibles, nous devrons utiliser la bande panchromatique (qui donne une résolution de 15 mètres) avec les bandes spectrales (qui auront toujours une résolution de 30 mètres).

d) La disponibilité de données DLG au 1:100000 pour les données sur les limites administratives et politiques devrait être utile. De plus, on n'utilise pas encore les enregistrements sur les "régions" dans le fichier DLG et ces renseignements pourraient se révéler utiles.

Voici quelques autres améliorations qui pourraient être apportées:

a) Des programmeurs de la NASA travaillent à la réalisation d'une interface utilisateur graphique pour CASS, ce qui permettrait d'afficher des menus dans des "fenêtres" et éliminerait ainsi le besoin d'utiliser le terminal des "menus".

b) Les fichiers de polygones créés dans CASS ont pu être mémorisés dans un SIG à l'aide de ARC/INFO. Il se pourrait qu'on puisse échanger des renseignements dans l'autre direction.

c) CASS permet de déterminer la latitude et la longitude de points. Cela aidera le NASS à commander des photographies plus facilement et avec plus de précision et permettra de combler d'autres besoins éventuels.

BIBLIOGRAPHIE

Ciancio, N.J., Rockwell, D.A., et Tortora, R.D. (1977). An Empirical Study of Area Frame Stratification. U.S. Department of Agriculture, Statistical Reporting Service Staff Report. Washington, D.C.

Cotter, J., et Nealon, J. (1987). Area Frame Design for Agricultural Surveys. U.S. Department of Agriculture, National Agricultural Statistics Service.

Fecso, R., et Johnson, V. (1981). The new California area frame: a statistical study. U.S. Department of Agriculture, *Statistical Reporting Service Publication*, SRS 22. Washington, D.C.

Fecso, R., Tortora, R.D., et Vogel, F.A. (1986). Sampling frames for agriculture in the United States, *Journal of Official Statistics*, 2, 3. Statistics Sweden.

Geuder, J. (1983). *An Evaluation of SRS Area Sampling Frame Designs: Ordering Count Units and Creating Paper Strata*. U.S. Department of Agriculture, Statistical Reporting Service Staff Report, AGES830715. Washington, D.C.

Geuder, J. (1984). *Paper Stratification in SRS Area Sampling Frames*. U.S. Department of Agriculture. Statistical Reporting Service, SF&SRB Staff Report, 79. Washington, D.C.

Hanuschak, G., et Morrissey, K. (1977). *Pilot Study of the Potential Contributions of Landsat Data in the Construction of Area Sampling Frames*. U.S. Department of Agriculture, Statistical Reporting Service Staff Report. Washington, D.C.

Houseman, E. (1975). *Area Frame Sampling in Agriculture*. U.S. Department of Agriculture, Statistical Reporting Service Publication SRS, 20. Washington, D.C.

Huddleston, H. (1976). *A Training Course in Sampling Concepts for Agricultural Surveys*. U.S. Department of Agriculture, Statistical Reporting Service Publication SRS, 21. Washington, D.C.

Jessen, R.J. (1942). *Statistical Investigation of a Sample Survey for Obtaining Farm Facts*, Iowa Agricultural Experiment Station Research Bulletin 304.

Pratt, W. (1974). *The Use of Interpenetrating Sampling in Area Frames*. U.S. Department of Agriculture, Statistical Reporting Service Staff Report. Washington, D.C.

Statistical Policy Working Paper 20 (1991). Seminar on Quality of Federal Data. *Federal Committee on Statistical Methodology Report*. Washington D.C., 32-33.

U.S. Department of Agriculture (1983). Scope and Methods of the Statistical Reporting Service. Publication 1308. Washington, D.C.

U.S. Department of Agriculture (1984). Area Frame Analysis Package. Statistical Reporting Service Staff Report. Washington, D.C.

U.S. Department of Agriculture (1987). International Training: Area Frame Development and Sampling. National Agricultural Statistics Service. Washington, D.C.

U.S. Department of Agriculture (1987). Supervising and Editing Manual, June Enumerative and Agricultural Surveys. National Agricultural Statistics Service. Washington, D.C.

Wigton, W., et Bormann, P. (1978). A Guide to Area Sampling Frame Construction Utilizing Satellite Imagery. U.S. Department of Agriculture. Statistical Reporting Service Staff Report. Washington, D.C.

SESSION 3

Analyse spatiale des données sur la santé et l'environnement

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

L'AUTOCORRÉLATION SPATIALE: UN PROBLÈME OU BIEN UN NOUVEAU PARADIGME?

P. Legendre[1]

L'autocorrélation spatiale peut être définie assez librement comme la propriété qu'ont les variables aléatoires observées de prendre, à deux endroits situés à une certaine distance l'un de l'autre, des valeurs plus comparables (autocorrélation positive) ou moins comparables (autocorrélation négative) que prévu pour des paires d'observations formées aléatoirement. L'autocorrélation est une propriété très générale des variables écologiques et, à vrai dire, de toutes les variables observées dans des séries chronologiques (autocorrélation temporelle) ou dans l'espace géographique (autocorrélation spatiale). La majorité des phénomènes écologiques naturels présentent une microrépartition géographique et on retrouve cette microrépartition à toutes les échelles spatiales -- du micromètre à l'échelle continentale et à l'échelle océanique. Lorsqu'on reproduit sur une carte la variation spatiale de la ou des variables étudiées, on observe soit une structure relativement uniforme ou bien une structure marquée par de fortes irrégularités.

Le problème statistique associé à la représentation spatiale des données écologiques peut être illustré par l'exemple typique des données autocorrélées spatialement. Les valeurs observées de la variable à l'étude -- par exemple, la composition spécifique -- sont le plus souvent influencées, peu importe la localité, par les assemblages présents dans les localités avoisinantes à cause de processus biotiques contagieux comme la croissance, la reproduction, la mortalité, la migration, etc. Ainsi, comme il est possible de prédire au moins en partie la valeur de la variable à une localité quelconque à l'aide des valeurs observées aux points avoisinants, ces valeurs ne sont pas stochastiquement indépendantes les unes des autres. Cela étonnera peut-être les écologistes qui ont été habitués à croire que la nature se comporte selon les hypothèses de la statistique classique, l'une d'elles étant l'indépendance des observations. Toutefois, les écologistes itinérants savent par expérience que les êtres vivants ne sont pas répartis uniformément ni aléatoirement dans la nature; il en va de même des variables physiques dont nous nous servons pour décrire le milieu. Dans les écosystèmes, l'hétérogénéité spatiale est une caractéristique active et non le résultat d'un processus aléatoire quelconque générateur de bruit; c'est pourquoi il devient important de l'étudier pour ce qu'elle est vraiment. Le message que cette communication vise à transmettre en premier lieu est qu'il faut revoir nos théories et nos modèles, faire en sorte qu'ils renferment des hypothèses réalistes sur la configuration spatiale et temporelle de communautés.

L'autocorrélation pour des variables soulève un problème statistique propre; elle diminue notre capacité d'exécuter des tests d'hypothèses courants. Heureusement, on voit surgir graduellement de nouvelles notions et de nouvelles méthodes statistiques qui nous permettent de traiter les données autocorrélées. Nous évoquerons ces notions et ces méthodes.

Enfin, nous décrirons des façons d'introduire des structures spatiales dans les modèles écologiques. Deux groupes de techniques ont été expérimentées la méthode des données brutes, fondée sur la régression multiple, et la méthode matricielle, basée sur le test de Mantel.

RÉFÉRENCE

Legendre, P. Spatial autocorrelation: Trouble or new paradigm? Ecology (article présenté dans le cadre d'un supplément spécial).

---

[1] P. Legendre, Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale A, Montréal (Québec), Canada H3C 3J7.

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

ANALYSE À PONDÉRATION LOCALE DE DONNÉES GÉOGRAPHIQUES SUR LES NAISSANCES: ESTIMATION ET PRÉSENTATION DU DEGRÉ D'INCERTITUDE

D.R. Brillinger[1]

RÉSUMÉ

L'objet de cette communication est d'analyser et de présenter des données agrégées par région géographique (par ex.: divisions de recensement) et qui décrivent des phénomènes variant graduellement dans l'espace. Dans Brillinger (1990b), des analyses à pondération locale de données géographiques ont été faites sur les naissances chez les femmes de 25 à 29 ans pour les années 1986 et 1987 pour la province de Saskatchewan au niveau des divisions de recensement. Divers résultats ont été présentés à l'aide de cartes en courbes de niveau. Le présent article développe cet argument et cherche plus particulièrement à calculer et à afficher des degrés d'incertitude pertinents pour les cartes en courbes de niveau. L'existence d'un effet des jours ouvrables est noté mais celui-ci ne varie pas de façon appréciable dans l'espace. La méthode exposée peut remplacer les méthodes empiriques de Bayes qui ont été proposées pour des problèmes semblables.

MOTS CLÉS: Données agrégées; distribution binomiale-logit normale; établissement de cartes en courbes de niveau; variation non représentée; interpolation; analyse à pondération locale; distribution logit normale; cartes; simulation; données géographiques; présentation de l'incertitude; covariables non mesurées.

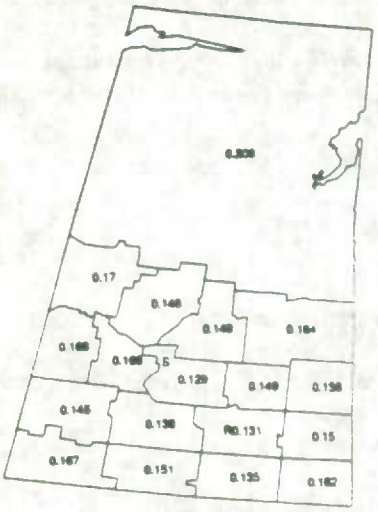
1. INTRODUCTION

Cette communication a pour but d'analyser des phénomènes qui varient graduellement dans l'espace et de trouver une façon d'indiquer le degré d'incertitude des résultats de l'analyse.

Les données dont il est question ici ont été agrégées et représentent des totaux par division de recensement, plus précisément les 18 divisions de recensement de la province de Saskatchewan. Les données du recensement portant sur le total quotidient des naissances par division de recensement chez les femmes de 25 à 29 ans pour la période 1986-1987 sont disponibles pour analyse. Les chiffres de population établis le jour du recensement de 1986 (3 juin) sont aussi disponibles. Nous cherchons à étudier et à représenter la structure géographique de ces données.

Figure 1

Saskatchewan: naissances chez les femmes de 25 à 29 ans pour la période 1986-1987



taux annuel de natalité

[1] D.R. Brillinger, Département de statistique, Université de Californie, Berkeley, CA., États-Unis.

La figure 1 montre les divisions de recensement de la Saskatchewan ainsi que le taux annuel de natalité observé dans chaque division. Les lettres R et S indiquent où sont situées respectivement les villes de Regina et de Saskatoon. Les taux ont été estimés directement à l'aide des chiffres fournis pour les deux années. Pour plus de détails sur ces données, prière de se référer à Brillinger (1990a,b).

2. PONDÉRATION ET ESTIMATION

L'objectif est d'analyser et de représenter des données géographiques au moyen de cartes en courbes de niveau. Lorsque des courbes de niveau sont établies à partir des valeurs d'une fonction en des points dispersés, on commence d'habitude par interpoler les valeurs dans une grille régulière; voir Pelto et coll. (1968). Étant donné (x_i,y_i,z_i), $i=1,...,n$ où z_i est la valeur d'une variable aléatoire mesurée au point (x_i,y_i), diverses méthodes ont été proposées pour cette interpolation. Une méthode courante est celle de Shepard (1968). Celui-ci calculez à la position (x,y) au moyen de la formule

$$z(x,y) = \sum_{i=1}^{n} w_i(x,y)\, z_i \,/\, \sum_{i=1}^{n} w_i(x,y) \tag{2.1}$$

où $w_i(x,y) = [(x-x_i)^2 + (y-y_i)^2]^{-\mu}$, $\mu > 0$. Le poids relatif à chaque point est dans une relation inverse avec la distance qui sépare ce point de (x,y). D'autres méthodes sont décrites dans Franke (1982) et Sabin (1985).

Dans le cas présent, une interpolation ne peut être effectuée car il existe des variations dues à l'échantillonnage. En outre, les données représentent des totaux, à partir desquels des proportions sont calculées, de sorte que les variations ne sont pas élémentaires. L'analyse de vraisemblance à pondération locale est une technique d'estimation appropriée pour les distributions non élémentaires qui varient dans l'espace; voir, par exemple, Gilchrist (1967), Brillinger (1977), Tibshirani et Hastie (1987), Cleveland et Devlin (1988), Staniswalis (1989), Brillinger (1990a,b). Supposons qu'une variable aléatoire Z a une distribution de probabilité $p(z \mid \theta)$ qui dépend d'un paramètre inconnu θ. Désignons par $\psi(z \mid \theta)$ la fonction de caractérisation, $\partial \log p / \partial \theta$, et choisissons $\hat{\theta}$, la valeur estimée de θ à la position (x,y), de manière à satisfaire l'équation

$$\sum_i w_i(x,y)\, \psi(z_i \mid \hat{\theta}) = 0 \tag{2.2}$$

pour une fonction de poids quelconque $w_i(x,y)$. Comme dans la méthode de Shepard, $w_i(x,y)$ dépend de la distance entre le point (x_i,y_i) et la position (x,y).

Pour un exemple simple d'une estimation à pondération locale, prenons le cas où B_i a une distribution binomiale avec π, N_i comme paramètres. On peut calculer directement l'estimation

$$\pi(x,y) = \frac{\sum_i w_i(x,y)\, B_i}{\sum_i w_i(x,y)\, N_i} \tag{2.3}$$

Cet estimateur est une extension naturelle de (2.1).

Dans le cas présent, où les données consistent en des agrégats pour des régions R_i, la fonction de poids appropriée est

$$w_i(x,y) = \frac{1}{|R_i|} \int_{R_i} \int W(x-u,y-v)\, du dv \tag{2.4}$$

$W(.)$ étant le bipoids,

$$W(x,y) = (1-u^2)^2 \quad \text{pour } |u| \leq 1 \tag{2.5}$$

et est égal à 0 dans le cas contraire, où $u = b\sqrt{x^2+y^2}$ pour une valeur $b > 0$. Dans l'équation (2.4), $|R_i|$ est la superficie de la division i. Par ailleurs, on peut imaginer que $w_i(x,y)$ représente l'influence de la division de recensement i sur une personne à la position (x,y), cette influence pouvant découler de facteurs comme les déplacements, la nutrition, le climat, l'ethnicité, l'instruction, la télévision, les lois. On évalue ces poids à l'aide d'une transformée de Fourier, tirant profit de la forme convolutée. Une fonction de poids élémentaire serait $w_i(x,y) = 1/|R_i|$, pour (x,y) dans R_i, et $= 0$ dans le cas contraire. Cela correspond à $W(.)$, une fonction delta.

La figure 2 montre l'effet de la variation de la valeur du paramètre b pour la division de recensement n° 18 qui couvre la partie septentrionale de la province. Les trois cas illustrés dans cette figure correspondent à trois valeurs de b différentes: (de haut en bas) - aucun lissage; - lissage modeste; - lissage moyen. Pour nos calculs, la valeur de b qui correspond au troisième cas a été retenue. La figure 3 donne la représentation graphique de l'équation (2.4) pour chacune des 18 divisions.

Figure 2

Division de recensement n° 18 -- effet du lissage

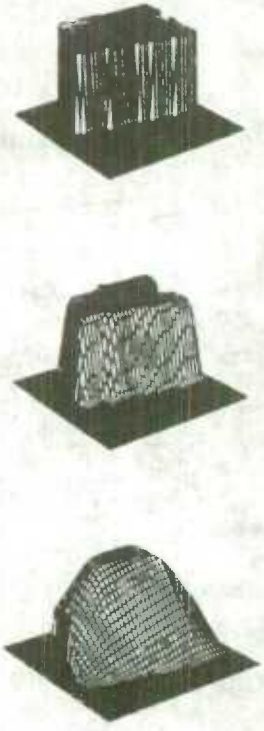
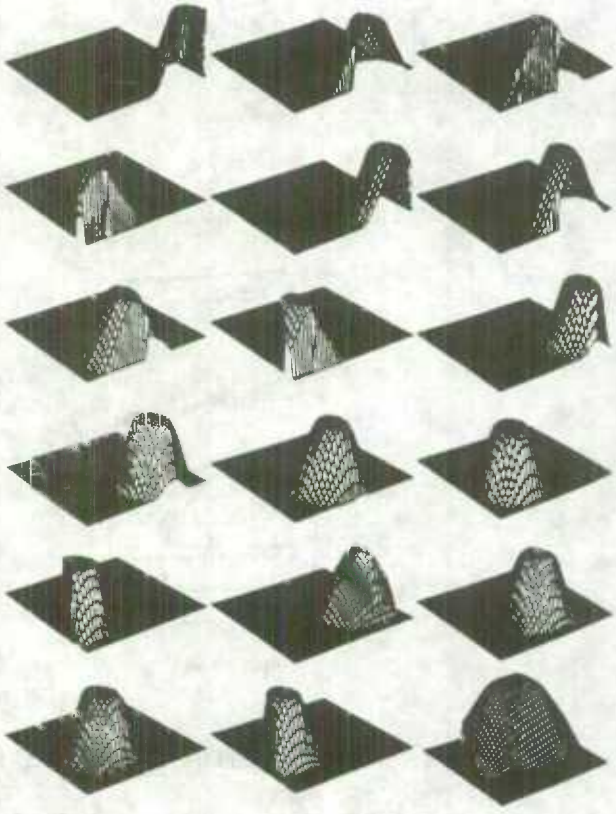


Figure 3



D'autres poids auraient pu être utilisés; voir à ce sujet Tobler (1979) et Dyn et Wahba (1982). L'avantage des poids utilisés dans la présente analyse est qu'ils ont un caractère local.

3. ESTIMATION AVEC BINOMIALE

Soit B_{ijk} le nombre de naissances enregistrées chez les femmes de 25 à 29 ans dans la division de recensement i pour l'année j, où k peut prendre la valeur 1 ou 2 selon qu'il s'agit de données pour un jour de semaine ou pour un jour non ouvrable. Désignons par N_i le nombre de femmes de 25 à 29 ans recensées dans la division i. On peut considérer la variable B_{ijk} comme le nombre de naissances enregistrées chez cette population. La distribution de cette variable peut être assimilée à une distribution binomiale. (Celle-ci semble mieux convenir

que la distribution de Poisson utilisée dans Brillinger (1990a,b) puisque la probabilité qu'une femme ait un bébé dans une année peut être aussi élevée que .2). Posons $x_1 = 2$ et $x_2 = -5$ pour faire en sorte que les estimations soient orthogonales. Nous allons supposer explicitement que B_{ijk} a une distribution binomiale avec, comme paramètres, π_k et N_i, où

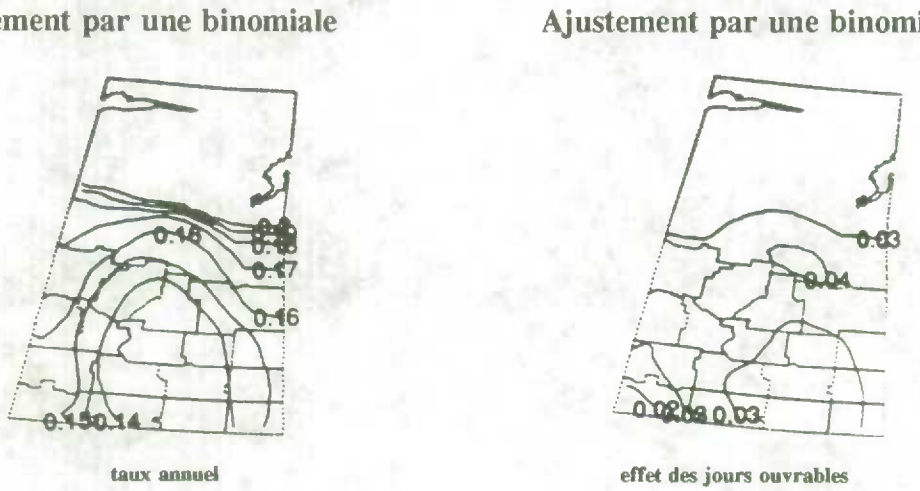
$$logit\ \pi_k = \log\ (\pi_k/(1-\pi_k)) = \alpha + \beta x_k \tag{3.1}$$

Le schéma de gauche de la figure 4 représente le tracé de contours du taux annuel de natalité estimé au moyen de l'expression

$$\frac{5}{7}\frac{\exp\ \hat{\eta}_1}{1+\exp\ \hat{\eta}_1} + \frac{2}{7}\frac{\exp\ \hat{\eta}_2}{1+\exp\ \hat{\eta}_2} \tag{3.2}$$

où $\hat{\eta}_k = \hat{\alpha} + \hat{\beta}x_k$, pour $k = 1, 2$, et il faut se rappeler que $\hat{\alpha} = \hat{\alpha}(x,y)$ et $\hat{\beta} = \hat{\beta}(x,y)$. Le premier terme de l'expression (3.2) correspond aux jours de semaine et le second, aux jours non ouvrables. On remarque dans le schéma de gauche des courbes de niveau qui s'étirent vers le haut, tendant à s'éloigner des divisions de recensement, y compris Regina et Saskatoon. Le schéma de droite est la représentation graphique de l'effet estimé des jours ouvrables $\hat{\beta}(x,y)$. On remarque que toutes les courbes de niveau sont positives, ce qui signifie qu'il y a plus de naissances les jours de semaine. Il convient de souligner que toutes ces valeurs sont exposées à des variations d'échantillonnage. La section 5 montre comment présenter ces variations dues à l'échantillonnage.

Figure 4



4. ESTIMATION AVEC BINOMIALE-LOGIT NORMALE

Dans Brillinger (1990a,b), on affirme qu'un bon nombre de variables explicatives utiles -- comme le régime alimentaire, le mode de vie, les conditions atmosphériques, l'environnement, les jours de congé, la structure par âge, l'urbanicité -- ne sont pas mesurées. À cause de cela, il y a des variations du nombre des naissances qui ne sont pas représentées par la loi binomiale. Une façon de corriger la situation est d'introduire un terme d'effet aléatoire, σz, et de remplacer (3.1) par l'équation

$$logit\ \pi_k = \alpha + \beta x_k + \sigma z \tag{4.1}$$

où z est une variable normale centrée réduite. Le modèle devient alors binomial-logit normal. On suppose que les variables z pour les différentes divisions de recensement sont indépendantes. L'ajustement de ce nouveau modèle peut se faire par l'intégration numérique; voir, par exemple, Bock et Lieberman (1970), Pierce et Sands (1975), Sanathanan et Blumenthal (1978), et Brillinger (1990a,b). Dans le cas présent, les N_i sont grands et un

modèle logit normal peut être substitué au nouveau modèle, en supposant que les logits de B_{ijk}/N_i suivent une distribution normale. Le modèle logit normal peut être exprimé comme suit:

$$\log B_{ijk}/(N_i - B_{ijk}) = \alpha + \beta x_k + \epsilon_{ijk} \tag{4.2}$$

où les ϵ_{ijk} sont indépendants et identiquement distribués selon une loi normale de moyenne 0 et de variance σ^2. L'hypothèse ci-dessus peut être vérifiée en partie en ajustant le modèle logit normal et en examinant les courbes de probabilité des résidus. Cela a été fait pour les données concernant les totaux annuels pour chaque jour de la semaine et chaque division de recensement. Aucun écart appréciable n'a été noté, exception faite d'une curieuse valeur aberrante.

Le taux annuel de natalité est maintenant estimé au moyen de l'expression

$$\frac{5}{7}\int \frac{\exp\,\hat{\eta}_1}{1+\exp\,\hat{\eta}_1}\,\phi(z)dz + \frac{2}{7}\int \frac{\exp\,\hat{\eta}_2}{1+\exp\,\hat{\eta}_2}\,\phi(z)dz \tag{4.3}$$

où $\eta_k = \hat{\alpha} + \hat{\beta}x_k + \hat{\sigma}z$, pour $k = 1, 2$, et où $\phi(.)$ est la densité normale. Le premier terme de cette expression correspond aux jours de semaine et le second, aux jours non ouvrables, comme dans l'expression (3.2). Crouch et Spiegelman (1990) étudient l'évaluation numérique d'intégrales comme celles de l'expression (4.3). Dans le cas présent, l'intégration gaussienne avec 21 noeuds est utilisée.

La partie supérieure gauche de la figure 5 illustre les résultats de la fonction d'estimation des taux (4.3). Si on compare ce schéma à l'ajustement binomial de la figure 4, les courbes de niveau semblent plus "aplaties". La partie supérieure droite de la figure 5 illustre l'effet des jours ouvrables, $\hat{\beta}(x,y)$. Dans ce cas-ci, les courbes semblent moins "aplaties" que pour le modèle binomial. La partie inférieure de la figure 5 contient les courbes de niveau pour la valeur estimée $\hat{\sigma}(x,y)$. Cette valeur semble moins élevée dans la région de Saskatoon, mais n'oublions pas qu'elle est soumise à des variations d'échantillonnage.

Figure 5

5. CALCUL ET PRÉSENTATION DE L'INCERTITUDE

L'utilisation de cartes simples soulève une foule de problèmes au point de vue de la présentation et de l'interprétation des résultats; voir, par exemple, Monmonier (1991). Les difficultés semblent encore plus nombreuses lorsqu'il s'agit d'indiquer le degré d'incertitude des résultats. Cette section présente quelques méthodes s'appliquant aux courbes de niveau.

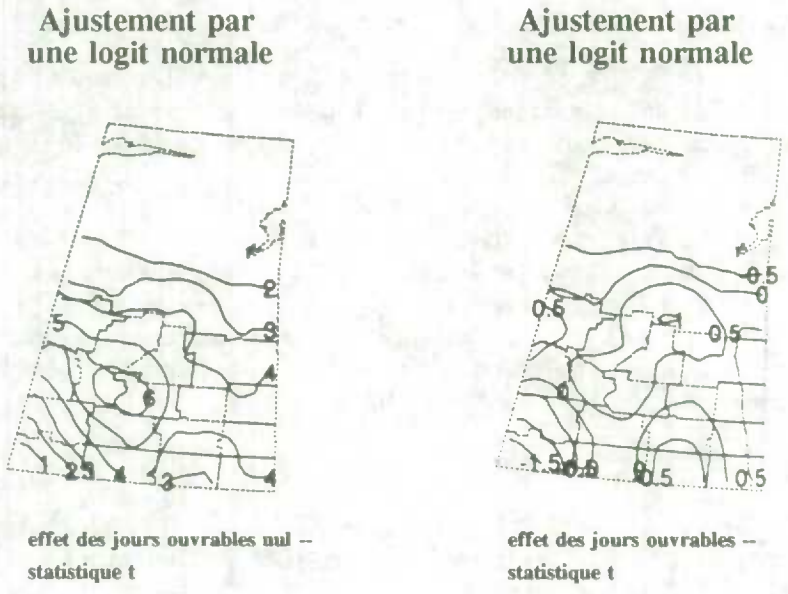
Le calcul même du degré d'incertitude à chaque position (x,y) ne semble pas poser problème. Les estimations obtenues au moyen de l'ajustement par une logit normale sont des estimations par les moindres carrés pondérés, de sorte que les variances de $\hat{\alpha}$ et de $\hat{\beta}$ peuvent être estimées par la formule (selon la notation classique)

$$\hat{\sigma}^2 (X'WX)^{-1} X'W^2X(X'WX)^{-1} \tag{5.1}$$

et l'écart-type de $\hat{\sigma}$ par

$$\frac{\hat{\sigma}}{\sqrt{2}} \frac{\sqrt{\sum w_i^2}}{\sum w_i} \tag{5.2}$$

Figure 6



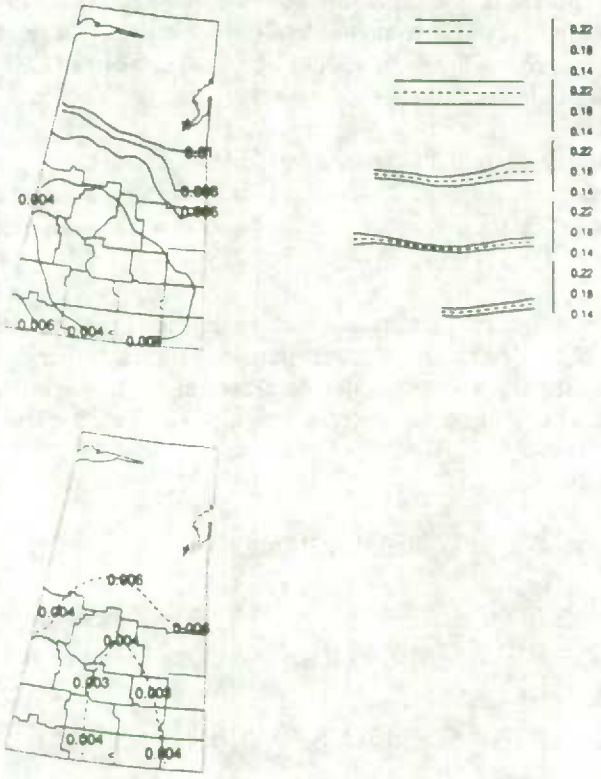
effet des jours ouvrables nul -- statistique t

effet des jours ouvrables -- statistique t

Les estimations du degré d'incertitude peuvent servir à analyser des hypothèses. La figure 6 a pour objet de vérifier s'il existe un effet des jours ouvrables et si cet effet varie selon les régions. Le schéma de gauche de la figure 6 représente le tracé de contours du quotient de l'estimation $\hat{\beta}(x,y)$ par l'erreur-type estimée correspondante. Les valeurs obtenues varient de 1 à 6, ce qui prouve l'existence d'un effet des jours ouvrables. Pour ce qui est de vérifier si cet effet varie dans l'espace, la statistique t est recalculée en prenant soin de soustraire du numérateur la valeur estimée pour l'ensemble de la province. On obtient alors des valeurs t qui varient de -1.5 à 1 et il n'est pas possible d'affirmer que l'effet varie dans l'espace.

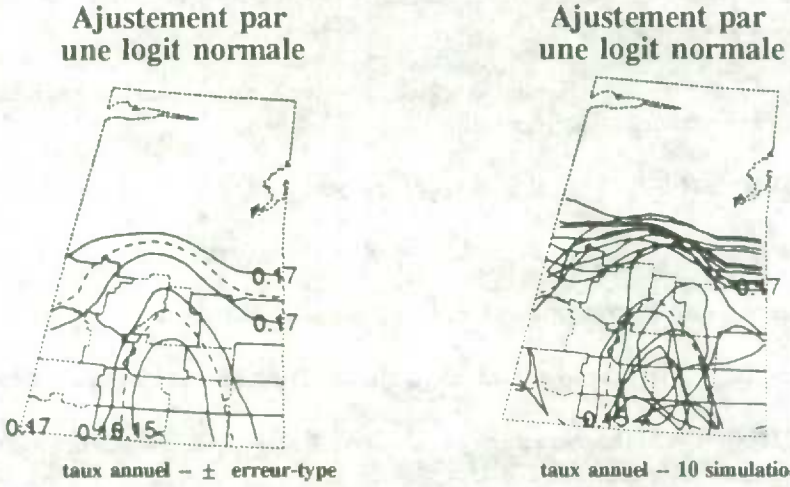
Les deux figures suivantes ont pour objet de représenter le degré d'incertitude de l'estimation du taux annuel de natalité calculée au moyen de l'expression (4.3) et représentée graphiquement dans la figure 5. L'erreur-type de cette estimation est estimée à l'aide de la méthode delta. Le schéma supérieur gauche de la figure 7 représente le tracé de contours de l'erreur-type estimée. Les cotes de courbe varient de .004 à .010 et laissent supposer l'existence d'une "concavité" centrée sur Regina et Saskatoon. Le schéma supérieur droit a été construit en fonction de cinq bandes "est-ouest" du territoire de la province. Il contient des intervalles équivalant à 2 erreurs-types de part et d'autre du taux de natalité estimé pour chaque bande. C'est une façon classique d'indiquer le degré d'incertitude pour les fonctions à une seule variable. Dans le schéma du bas, on représente par une ligne pointillée les contours à .15 et à .17 du taux de natalité et on inscrit à des endroits choisis le long de ces contours les erreurs-types estimées. Il semble plus facile d'illustrer le degré d'incertitude de cette manière sauf que les estimations de l'erreur-type ne peuvent être connues que pour une minorité de positions.

Figure 7



Le schéma de gauche de la figure 8 montre comment se comporte une courbe de niveau lorsqu'une erreur-type est ajoutée ou soustraite de la fonction d'estimation du taux de natalité. Une bande lisse se dessine autour de la courbe de niveau du taux estimé. Bien que l'objet de ces bandes soit clair, il est nécessaire d'en faire une interprétation attentive. Le schéma de droite de la figure 8 illustre le résultat de dix simulations du modèle logit normal. Les estimations $\hat{\alpha}(x,y)$, $\hat{\beta}(x,y)$, $\hat{\sigma}(x,y)$ sont d'abord agrégées pour chaque division de recensement. Des estimations du processus sont ensuite exécutées de façon indépendante suivant l'équation (4.2). L'ajustement du modèle est refait à la manière "bootstrap" pour chaque simulation et les contours à .15 et à .17 sont déterminés. Diaconnis et Efron (1983) proposent l'établissement de contours par des méthodes "bootstrap". Les différents contours à .17 forment une espèce de bande, tandis que les contours à .15 décrivent une figure plutôt irrégulière. Les lignes pointillées représentent les contours à .15 et à .17 originaux.

Figure 8



6. ANALYSE ET RÉSUMÉ

La combinaison de la fonction de poids, $w(.)$, avec l'effet aléatoire σz, permet de faire en sorte que la valeur estimée à la position (x,y) "emprunte de l'information" aux valeurs de toutes les divisions de recensement; voir, par exemple, Mallows et Tukey (1982). Cette méthode d'estimation peut donc être substituée aux méthodes empiriques de Bayes qu'ont élaborées Clayton et Kaldor (1987), Tsutakawa (1988), Cressie et Read (1989) et Manton et coll. (1989) pour les données de ce genre.

Le calcul du degré d'incertitude a permis d'analyser l'hypothèse de l'effet des jours ouvrables nul et celle de l'effet des jours ouvrables constant sur tout le territoire de la province. L'avantage de la représentation graphique est que s'il y a variation dans l'espace, les schémas peuvent nous donner une idée du genre de variation.

Beaucoup de recherches restent à faire. Notons, par exemple, des sujets comme le choix de la valeur du paramètre b dans l'équation (2.5), l'utilisation d'autres fonctions de poids, l'analyse formelle et informelle de la validité de l'ajustement, l'emploi d'autres méthodes de présentation de l'incertitude, y compris la mesure de variables explicatives, et enfin, la définition des critères d'asymptoticité qu'il convient d'utiliser dans l'étude de la méthode.

REMERCIEMENTS

Cette étude a été rendue possible en partie grâce à une subvention de la National Science Foundation (n° DMS-8900613).

BIBLIOGRAPHIE

Bock, R.D., et Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, 35, 179-197.

Brillinger, D.R. (1977). Discussion of Stone (1977). *Annals of Statistics*, 5, 622-623.

Brillinger, D.R. (1990a). Représentation cartographique de données agrégées sur les naissances. *Recueil du Symposium de 1989 sur l'analyse des données dans le temps* (Éds. A.C. Singh et P. Whitridge), Statistique Canada, Ottawa, Canada, 77-83.

Brillinger, D.R. (1990b). Modélisation spatiale et temporelle de données agrégées sur les naissances. *Techniques d'enquête*, 16, 267-282.

Clayton, D. et Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrika*, 43, 671-681.

Cleveland, W.S. et Devlin, S.J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83, 596-610.

Cressie, N. et Read, T.R.C. (1989). Spatial analysis of regional counts. *Biometrika*, 31, 699-719.

Crouch, E.A.C. et Spiegelman, D. (1990). The evaluation of integrals of the form: $\int_{-\infty}^{\infty} f(t)\exp(-t^2)dt$ application to logistic-normal models. *Journal of the American Statistical Association*, 85, 464-469.

Diaconnis, P. et Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 248, 116-130.

Dyn, N. et Wahba, G. (1982). On the estimation of functions of several variables from aggregated data. *SIAM Journal of Mathematical Analyses*, 13, 134-152.

Franke, R. (1982). Scattered data interpolation: tests of some methods. *Math. Comp.*, 38, 181-200.

Gilchrist, W.G. (1967). Methods of estimation involving discounting. *Journal of the Royal Statistical Society*, 29, 355-369.

Mallows, C.L. et Tukey, J.W. (1982). An overview of techniques of data analysis emphasizing its exploratory aspects, *Some Recent Advances in Statistics* (Éds. J. Tiago de Oliveira et coll.). Academic, London, 111-172.

Manton, K.G., Woodbury, M.A., Stallard, E., Riggan, W.B., Creason, J.P. et Pelom, A.C. (1989). Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. *Journal of the American Statistical Association*, 84, 637-650.

Monmonier, M. (1991). How to Lie with Maps. Chicago Press, Chicago.

Pelto, C.R., Elkins, T.A. et Boyd, H.A. (1968). Automatic contouring of irregularly spaced data. *Geophysics*, 33, 424-430.

Pierce, D.A. et Sands, B.R. (1975). Extra-binomial variation in binary data. Technical Report 46, Statistics Department, Oregon State University.

Sabin, M.A. (1985). Contouring - the state of the art. Fundamental Algorithms for Computer Graphics (Éd. R.A. Earnshaw). *NATO ASI Series*, F17. New York: Springer-Verlag.

Sanathanan, L. et Blumenthal, S. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association*, 73, 794-799.

Shepard, D. (1968). A two-dimensional interpolation function for irregularly spaced data. *Proceedings of the 23rd National Conference ACM*, 517-523.

Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, 84, 276-283.

Stone, C.J. (1977). Consistent nonparametric regression. *Annals of Statistics*, 5, 595-620.

Tibshirani, R. et Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, 82, 559-567.

Tobler, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association*, 74, 519-536.

Tsutakawa, R.K. (1988). Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*, 83, 37-42.

LÉGENDES DES FIGURES

Figure 1. Taux annuel de natalité chez les femmes de 25 à 29 ans pour les années 1986 et 1987, par division de recensement. Les lettres "R" et "S" indiquent où sont situées les villes de Regina et de Saskatoon respectivement.

Figure 2. Représentation graphique de l'équation (2.4) pour la division de recensement n° 18 dans trois cas particuliers: absence de lissage; lissage modeste; et lissage moyen.

Figure 3. Représentation graphique de l'équation (2.4) pour chacune des 18 divisions de recensement.

Figure 4. Analyse binomiale. Les estimations de taux sont calculées au moyen de l'expression (3.2). Le schéma de droite est la représentation graphique de $\hat{\beta}(x,y)$.

Figure 5. Analyse logit normale. Les estimations de taux sont calculées au moyen de l'expression (4.3). Les autres schémas sont la représentation graphique de $\hat{\beta}(x,y)$ et de $\hat{\sigma}(x,y)$.

Figure 6. Valeurs de la statistique t visant à vérifier l'hypothèse d'un effet des jours ouvrables nul et celle d'un effet des jours ouvrables constant dans l'espace.

Figure 7. Le premier schéma est un tracé de contours de l'erreur-type estimée de l'estimation du taux de natalité. Le schéma de droite contient des intervalles équivalant à 2 erreurs-types de part et d'autre du taux de natalité estimé pour 5 bandes "est-ouest" du territoire de la province. Le troisième schéma indique l'erreur-type estimée à certains points des contours à .15 et à .17 du taux de natalité estimé.

Figure 8. Le premier schéma illustre par une ligne pointillée les contours à .15 et à .17 et par une ligne continue, les contours correspondants lorsqu'on ajoute ou qu'on soustrait une erreur-type. Le second schéma illustre les résultats des simulations du modèle (4.2) selon le mode de présentation décrit dans cette communication. Il s'agit de dix simulations.

SESSION 4

Développements spatiaux au niveau du traitement des données

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

CODAGE AUTOMATISÉ DES DONNÉES SUR LA MOBILITÉ ET LES NOMS DE LOCALITÉ POUR LE RECENSEMENT DE 1991

M.J. Norris et S. Coyne[1]

RÉSUMÉ

Des données spatiales sur le "lieu de résidence il y a cinq ans" sont recueillies lors du recensement de la population du Canada. La question du recensement portant sur la mobilité demande aux migrants d'écrire le nom de la municipalité, de la division de recensement (DR) ou du comté et de la province ou du territoire dans lesquels ils habitaient cinq ans auparavant. Dans le cadre du traitement des données, ces réponses en toutes lettres sont converties en codes à 7 chiffres de la Classification géographique type (CGT) qui comprennent un code de province à 2 chiffres, un code de division de recensement (DR) à 2 chiffres et un code de subdivision de recensement (SDR) à 3 chiffres. Jusqu'au dernier recensement (1991), les réponses en toutes lettres (renseignements sur la localité) aux questions sur la mobilité étaient converties manuellement en codes de la CGT. Maintenant, pour la première fois dans l'histoire du recensement au Canada, un codage automatisé a été mis en application afin de convertir les réponses en toutes lettres en codes numériques pour un certain nombre de variables, y compris la mobilité. Dans la présente communication, on décrit la stratégie et la structure de la base de données utilisées pour effectuer le codage automatisé des noms de localité en toutes lettres et pour résoudre un certain nombre de problèmes relatifs aux réponses tels que: l'utilisation de noms communs plutôt que du nom officiel des localités; les renseignements incomplets (réponses partielles) et les noms de localité répétés. On discute des implications du codage automatisé des données spatiales, particulièrement de l'incidence de l'amélioration de la précision du codage sur les données sur la migration au niveau de la SDR.

MOTS CLÉS: Codage automatisé; recensement; migration; mobilité; noms de localité.

1. INTRODUCTION

1.1 Renseignements généraux

Depuis 1961, le recensement du Canada comprend une question sur le "lieu de résidence il y a 5 ans". Dans cette question on demande aux répondants s'ils ont déménagé ou non, c'est-à-dire, s'ils vivaient à une adresse différente cinq ans auparavant et, dans l'affirmative, s'ils vivaient dans une autre ville ou un autre village. Dans les cas où ces personnes vivaient dans une ville ou un village, etc. différent, on demandait aux répondants d'écrire le nom de la ville, du comté et de la province. On trouvera à la figure 1, la version de la question sur la mobilité 'depuis 5 ans' posée dans le cadre du recensement de 1991 ainsi qu'un exemple de réponse.

Pendant les opérations reliées au recensement, on attribue à ces noms de localité en toutes lettres le code de la Classification géographique type (CGT) qui correspond à la subdivision de recensement (SDR), c'est-à-dire à la ville, au village, etc., à la division de recensement (DR) (comme un comté) et à la province. Le code de la CGT est un code à 7 chiffres composé d'un code de province à 2 chiffres, d'un code de DR à 2 chiffres et d'un code de SDR à 3 chiffres.

---

[1] M.J. Norris et S. Coyne, Division de la démographie, Statistique Canada, 6-A6, édifice Jean-Talon, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.

Le codage de ces noms de localité en toutes lettres nous permet d'obtenir, pour une SDR particulière, le nombre de personnes qui y avaient habité 5 ans auparavant. En d'autres mots, nous pouvons obtenir, pour chaque SDR, le nombre de sortants (émigrants internes), ainsi que d'entrants (immigrants internes) (les personnes qui vivaient à l'extérieur de cette SDR 5 ans auparavant). Ces données sur la migration au niveau de la SDR sont utilisées par les planificateurs et par les chercheurs. La mesure dans laquelle le code attribué à ces réponses en toutes lettres est identique aux codes géographiques correspondants est un facteur critique pour la qualité des données sur la migration dans le cas des SDR, particulièrement pour la sortie (émigration interne).

FIGURE 1. Questions sur la mobilité, demandant le lieu de résidence il y a 5 ans, posées lors du recensement de 1991

21. Cette personne habitait-elle à l'adresse actuelle il y a 5 ans, c'est-à-dire le 4 juin 1986?

22. Où cette personne habitait-elle il y a 5 ans, c'est-à-dire le 4 juin 1986?

Certaines grandes villes sont formées de petites villes appelées municipalités. S'il y a lieu, faites la distinction entre la municipalité et la grande ville, par exemple, Anjou et Montréal, Scarborough et Toronto, Burnaby et Vancouver, Saanich et Victoria.

Cochez un seul cercle.

25.

01 ○ Oui, habitait à l'adresse actuelle
Passez à la question 23

02 ⊗ Non, habitait à une autre adresse

03 ○ Habitait le ou la même ville, village, canton, municipalité ou réserve indienne

OU

04 ⊗ Habitait un ou une autre ville, village, canton, municipalité ou réserve indienne du Canada
Inscrivez en lettres moulées ci-dessous.

Ville, village, canton, municipalité ou réserve indienne

05 WINDSOR

Comté (si vous le connaissez)
ESSEX

Province ou territoire
ONT

OU

06 ○ Habitait en dehors du Canada
Inscrivez en lettres moulées le nom du pays.

07

Jusqu'au dernier recensement (1991), les réponses en toutes lettres aux questions sur la mobilité étaient converties manuellement en codes de la CGT dans le cadre des opérations régionales. Cette opération de codage était basée sur le Cahier des codes des noms de localité (CCNL) un ouvrage de référence renfermant environ 40 000 noms de localité au Canada (pour 1986), y compris des noms de localité officiels ou non, comme des noms de quartier et de localité non constituée. Ces noms de localité correspondent à environ 6 000 codes de la classification géographique type. Un exemple du genre de renseignements que renferme le CCNL est donné à la figure 2. Maintenant, pour la première fois dans l'histoire du recensement au Canada, un codage automatisé a été mis en application afin de convertir les réponses en toutes lettres en codes numériques pour un certain nombre de variables, comme le principal domaine d'études et le lieu de naissance, ainsi que la mobilité.

FIGURE 2. Exemple du contenu du Cahier des codes des noms de localité utilisé lors du recensement de 1986

Ontario

CMA/CA RMR/AR	PR PR	CD DR	CSD SDR	NAME NOM	TYPE GENRE	CMA/CA RMR/AR	PR PR	CD CD	CSD CSD	NAME NOM	TYPE GENRE
539	35	26	028	PELHAM CORNERS			35	09	021	PERTH	T
539	35	26	053	PELHAM ROAD			35	09	008	PERTH AIRPORT	
539	35	26	028	PELHAM TOWNSHIP			35	10	018	PERTH ROAD	
539	35	26	028	PELHAM UNION			35	43	068	PETAGUISHENE BEACH	
	35	60	090	PELICAN			35	47	078	PETAWAWA	TP
	35	60	090	PELLATT TOWNSHIP			35	47	079	PETAWAWA	VL
559	35	37	039	PELTON			35	47	078	PETAWAWA POINT	
	35	07	061	PELTONS CORNERS			35	57	095	PETERBELL	
	35	47	062	PEMBROKE	TP	529	35	15	014	PETERBOROUGH	C
	35	47	064*	PEMBROKE	C		35	15	006	PETERBOROUGH AIRPORT	
	35	47	074	PEMBROKE AIRPORT		537	35	25	030	PETERS CORNERS	

1.2 Structure de la communication

Dans la présente communication on fait un examen des développements dans le domaine du codage automatisé des données sur la mobilité spatiale, en commençant avec certaines données de base sur le besoin d'effectuer un codage automatisé des noms de localité en toutes lettres inscrits en réponse aux questions sur la mobilité; on présente aussi les objectifs du codage automatisé des noms de localité. Cette discussion est suivie d'une description des méthodes et stratégies appliquées pour coder automatiquement les noms de localité en toutes lettres à l'aide du logiciel de codage automatisé par reconnaissance de texte (CART) mis au point par Statistique Canada. On examine aussi, en plus de la stratégie et de la solution des problèmes de réponse pendant le codage automatisé, les développements liés à la production de données codées sur la mobilité, y compris le contrôle de la qualité ainsi que le contrôle et l'imputation. Les résultats des essais et de la production jusqu'à ce jour sont analysés, ainsi qu'une évaluation du codage automatisé de ces données spatiales jusqu'à ce jour. Finalement, on examine les implications du codage automatisé des données sur la mobilité spatiale, particulièrement en rapport avec la qualité améliorée des données sur la migration au niveau de la SDR ainsi que de nouveaux renseignements sur l'utilisation des noms de localité.

2. LE BESOIN D'EFFECTUER LE CODAGE AUTOMATISÉ DES DONNÉES SUR LA MOBILITÉ

Trois raisons principales expliquent pourquoi la variable sur la mobilité était considérée comme variable pouvant faire l'objet d'un codage automatisé: la qualité des données, le volume et le coût. Le besoin d'effectuer le codage automatisé des noms de localité fournis en réponse aux questions sur la mobilité a été reconnu à la suite d'une étude d'évaluation des données sur la mobilité et sur la migration, provenant du recensement de 1986, au niveau des petites régions. Cette étude (Norland 1988) a montré qu'il existait des problèmes au niveau de la qualité des données dans le cas des données sur la migration pour les régions, particulièrement pour les SDR et, jusqu'à un certain point, pour les DR. On a trouvé que ces problèmes de qualité des données étaient imputables à deux sources d'erreur: les erreurs de codage et les erreurs de réponse. En plus des considérations d'ordre qualitatif, du simple point de vue volume, le codage automatisé était justifié, compte tenu du fait qu'il y a près d'un million de noms de localité en toutes lettres et de la possibilité de réduire les coûts de production.

2.1 Qualité des données sur la migration au niveau des petites régions

La combinaison des erreurs de codage et des erreurs de réponse entraînait un certain nombre de problèmes relatifs à la qualité des données au niveau de la SDR, tels que: les taux de migration pour les "petites SDR" (les SDR avec une population de moins de 250 personnes) sont douteux; on trouve un nombre important de SDR plus grosses avec des taux de sortie (émigration interne) excessifs; il y a des problèmes spéciaux dans le cas des données pour des "noms de localité répétés" (p. ex., Barrie, nom pour lequel il existe le township de Barrie dans le comté de Frontenac et la ville de Barrie dans le comté de Simcoe) et des problèmes spéciaux portant sur certaines SDR dans des RMR, comme Saanich et Victoria. Cette dernière situation est plus due à une erreur de réponse, en ce sens que le répondant a tendance à confondre le nom de la région métropolitaine plus grande (p. ex., Victoria) avec le nom de la municipalité de banlieue (p. ex., Saanich). Cependant, pour les autres genres de problèmes, l'erreur de codage a été une source importante d'erreurs, particulièrement dans les cas où il fallait régler les problèmes liés aux codes de la CGT pour les noms de localité répétés. Les erreurs de réponse et celles des codeurs ont un effet sur l'affectation appropriée des codes de la CGT pour "la SDR du lieu de résidence il y a cinq ans". Et par conséquent, sur la mesure du nombre de sortants (émigrants internes) des SDR ainsi que sur les taux correspondants de sortie (émigration interne). Comme on l'a fait remarquer dans l'étude, c'est pour cette raison que les taux de sortie (émigration interne) pour les SDR ont tendance à être moins bons que les taux d'entrée (immigration interne), pour lesquels les erreurs de ce genre n'ont pas le même effet, puisque les taux d'entrée (immigration interne) sont basés sur les codes de la CGT du lieu de résidence au moment du recensement.

Certains exemples des genres de problèmes qui ont été relevés lors de l'étude des données sur la migration recueillies dans le cadre du recensement de 1986 sont présentés dans le tableau 1. Les taux de sortie (émigration interne) pour certaines SDR sont donnés afin de montrer l'importance du problème dont on soupçonne l'existence. La SDR de Tungsten dans les Territoires du Nord-Ouest, qui compte environ 200 personnes, montre le problème lié à une faible population, avec un taux de sortie (émigration interne) calculé pour 1981-1986 de 149 migrants pour 100 résidents. Des taux de sortie (émigration interne) tellement élevés qu'on s'en méfie sont observés même pour des SDR plus grandes comme Dawson Creek en Colombie-Britannique, avec une population de plus de 9 000 habitants et un taux de sortie (émigration interne) de 50 pour cent. Le problème spécial des noms de localité répétés est illustré par le cas de la ville de Barrie et du township de Barrie, tous les deux en Ontario. La SDR plus petite du township de Barrie a un taux de sortie (émigration interne) de 133%, comparativement à 19% pour la ville de Barrie, probablement parce que les codeurs ont attribué, par erreur, le code de la CGT du township de Barrie plutôt que celui de la ville de Barrie quand seulement "Barrie" était écrit pour le nom de la localité. Le problème spécial dû au fait que les répondants confondent des SDR particulières avec la RMR plus grande est illustré par le cas de Saanich et de Victoria. Ce problème peut aussi devenir un problème de codage dans le cas où les deux noms de localité sont fournis.

À cause de ces divers problèmes reliés à la qualité des données, on a fait, dans l'étude, un certain nombre de recommandations relativement à l'utilisation de données sur la migration au niveau de la SDR, comme le fait que les utilisateurs devraient se reporter aux régions avec de grosses populations de base et qu'ils devraient être informés des "situations spéciales" comme les "noms de localité répétés". Aussi, les plans originaux pour la publication de données sur l'entrée (immigration interne), la sortie (émigration interne) et la migration nette pour les SDR ont été modifiés à cause du nombre important de SDR avec des taux de sortie (émigration interne) excessifs. À la fin, les études schématiques sur les SDR n'ont renfermé que des données sur la mobilité. De plus, dans l'étude, on a recommandé que des améliorations soient apportées aux procédures de codage et aux manuels pour le recensement de 1991 (à ce moment, les recommandations étaient faites en fonction du codage manuel). D'autres renseignements sur les problèmes liés à la qualité des données sont fournis dans le 'Guide à l'intention des utilisateurs - Données du recensement de 1986 sur la mobilité' (Norris 1990). On trouve aussi, dans le rapport sur la mobilité du Test du recensement national (Norris 1989), les changements apportés à la question sur le 'lieu de résidence il y a 5 ans' afin de réduire, pour 1991, certains des problèmes dus aux répondants.

On a reconnu que le codage automatisé pouvait permettre d'améliorer la qualité des données sur la migration recueillies au cours du recensement de 1991 en fournissant un moyen plus précis et plus cohérent de coder les noms de localité en toutes lettres. Le codage manuel avait posé des problèmes en matière de qualité des données à cause des erreurs des codeurs, du codage incohérent et subjectif, ainsi que des instructions et des

manuels complexes qu'il fallait suivre. Contrairement au codage manuel, le codage automatisé est cohérent et offre la possibilité d'isoler de façon systématique les genres de réponses qui posent un "problème spécial".

Tableau 1: Exemples de SDR pour lesquelles on soupçonne qu'il pourrait exister des problèmes de qualité des données, 1986

Genre de problème	SDR	Population en 1986 (personnes âgées de 5 ans et plus)	Taux de sortie (émigration interne)[1] (pour 100 habitants) (1981-1986)
Faible population	Tungsten (T.N.-O.)	205	149
Plus grande SDR avec des taux de sortie (émigration interne) trop élevés	Dawson Creek (C.-B.)	9 470	50
Noms de localité répétés	Ville de Barrie (Ontario)	44 440	19
	Township de Barrie (Ontario)	690	133
Certaines SDR dans des RMR	Saanish (C.-B.)	77 045	2
	Victoria (C.-B.)	60 540	52

[1] Le nombre de sortants (émigrants internes) est obtenu à partir des réponses à la question du recensement de 1986 sur le 'lieu de résidence il y a 5 ans', les réponses à cette question permettent d'obtenir le nombre de sortants (émigrants internes) âgés de 5 ans et plus, pour la période allant de 1981 à 1986.

Source: Recensement de 1986, données non publiées sur la mobilité.

3. OBJECTIFS DU CODAGE AUTOMATISÉ DES DONNÉES SUR LA MOBILITÉ

L'objectif principal du codage automatisé des données sur les noms de localité fournis en réponses aux questions sur la mobilité est d'améliorer la qualité du codage par rapport au codage manuel. De plus, d'autres objectifs importants du codage automatisé en général comprennent la réduction du temps de traitement et des frais.

Pour être précis, l'objectif du codage automatisé des données sur la mobilité est d'attribuer avec précision à un code géographique à sept chiffres à un nom de localité en toutes lettres correspondant. En d'autres mots, le système de codage automatisé doit attribuer automatiquement, avec un minimum d'erreur, un code au plus grand nombre possible d'inscriptions parmi le million d'inscriptions qui doivent être traitées. À cause de la nature des problèmes attribuables aux répondants en matière de mobilité, on a réalisé, au cours de l'élaboration du codage automatisé, que le système ne pouvait attribuer avec précision des codes pour toutes les réponses en toutes lettres. À des fins d'efficience et de précision, le système a été élaboré avec deux objectifs: réaliser le codage automatisé d'au moins 70% des réponses en toutes lettres avec un niveau de précision acceptable et acheminer les problèmes de codage spéciaux, comme les noms de localités répétés, vers un service de traitement manuel. Ainsi, la phase de traitement manuel du codage automatisé ne porterait que sur les réponses en toutes lettres pour lesquelles le système ne pourrait attribuer des codes avec précision.

3.1 Problèmes de codage imputables à la réponse

Au début, on pourrait penser que l'attribution d'un code de la CGT à un nom de localité en toutes lettres est un problème simple. Après tout, on demande au répondant de fournir le nom de la ville, du village, de la réserve, etc., du comté et de la province, ce qui correspond clairement à la SDR, au DR et à la province du code de la CGT. Toutefois, divers problèmes dus aux répondants ou reliés aux noms de localité peuvent rendre le codage des données sur le nom de localité assez compliqué, particulièrement quand on utilise un système automatisé. Le tableau 2 renferme certains genres de problèmes, avec des exemples, qui peuvent rendre problématique le couplage du nom de localité en toutes lettres aux codes de la CGT. Par exemple, une réponse incomplète peut parfois, mais pas toujours, mener aux problèmes des noms de localité répétés - la réponse

incomplète 'Kingston' peut se rapporter à six SDR possibles au Canada avec 'Kingston' comme nom de localité. Dans le cas d'une faute d'orthographe ou d'une erreur de frappe simple, mais non prévue, le système automatisé ne peut attribuer de code, comme dans le cas de 'Totonto'.

Tableau 2: Problèmes reliés à l'attribution de codes de la CGT pour des données relatives aux noms de localité fournis en réponses aux questions sur la mobilité

Genre de réponse	Exemples
• Réponse incomplète	Kingston
• Usage courant, comme des noms de quartier et de localité non constituée	Glebe (quartier), Lake Park
• Fautes d'orthographe	Totonto
• Abréviations	Mtl.
• Combinaisons erronées de ville, de comté et (ou) de province	Kingston, Manitoba
• Noms de comté erronés	Circonscriptions électorales provinciales déclarées plutôt que les comtés, particulièrement au Québec
• Plus d'un nom de localité	Saanich/Victoria Wawa/Hawk Junction
• Ordre des mots	Poplar Point/Point Poplar
• Adresse	235 Chemin Montréal, Ottawa
• Réponse non géographique	Sur la ferme
• Noms et limites périmés	Galt

4. SYSTÈME DE CODAGE AUTOMATISÉ

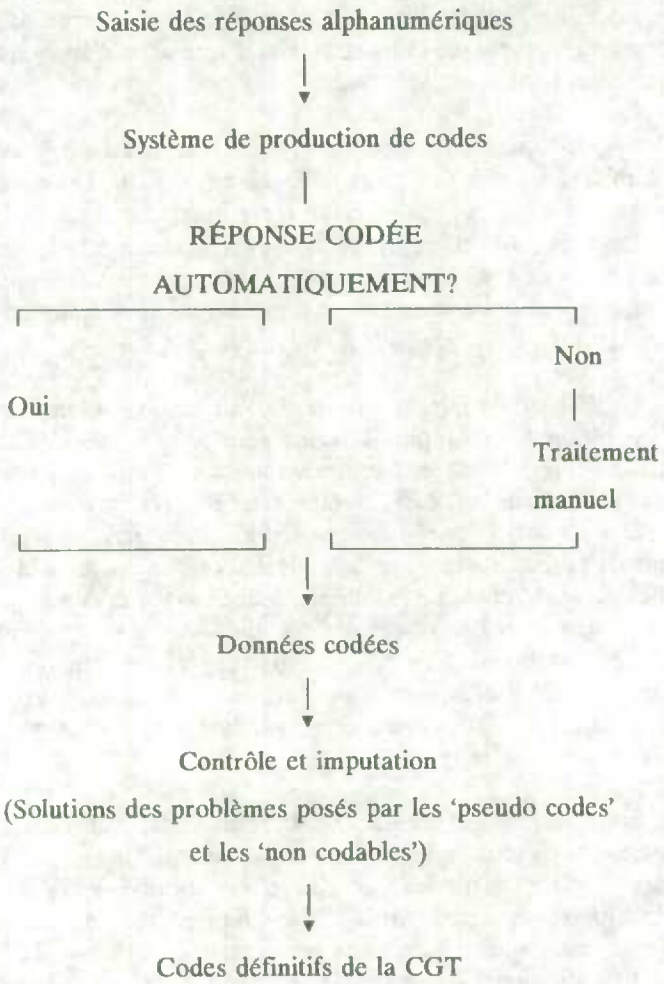
4.1 Aperçu du codage automatisé

La figure 3 donne un aperçu de la méthode de codage automatisé utilisée pour le recensement de 1991. Ce processus commence avec les noms dont la saisie a été effectuée, c'est-à-dire, dans le cas des questions sur la mobilité, avec des noms de localité en toutes lettres. Ces réponses sont introduites dans le système de codage automatisé qui interprète les réponses saisies afin d'attribuer un code numérique. Il y a deux résultats possibles: un code est attribué automatiquement à la réponse en toutes lettres ou, à défaut, la réponse en toutes lettres est transmise pour subir un traitement manuel. Dans ce dernier cas, des codeurs spécialement formés essaient d'attribuer des codes aux réponses en toutes lettres auxquelles le système automatisé n'a pu en attribuer. Certains des codes attribués pendant le traitement manuel sont des 'pseudo' codes à 7 chiffres, on attribue aussi des codes réels de la CGT. Les codeurs doivent aussi déterminer s'il est possible ou non d'attribuer un code à une réponse en toutes lettres. Ainsi, les résultats du système de production de codes seront composés de données codées, dont la majorité sont des codes de la CGT; ainsi que de 'pseudo' codes et de réponses 'non codables', pour ces deux dernières catégories, une solution est trouvée plus tard au cours de l'étape du contrôle et de l'imputation à la suite de laquelle des codes définitifs de la CGT sont obtenus.

4.2 Codage automatisé par reconnaissance de texte (CART)

Le système automatisé de production de codes est basé sur le 'CART', ce qui signifie 'Codage automatisé par reconnaissance de texte', un logiciel élaboré par la Section de la recherche et des systèmes généraux de Statistique Canada. Les résultats de la recherche sur l'utilisation du CART comme système de codage automatisé pour le recensement sont présentés dans le rapport de Fyffe et coll. (1988). Le système CART interprète des réponses en toutes lettres introduites manuellement afin de leur attribuer un code numérique correspondant. Par exemple, dans l'application du CART à la mobilité, le système attribue au nom de localité en toutes lettres Ottawa, Ontario le code de la CGT 3506014.

Figure 3: Aperçu du codage automatisé



Afin d'apparier les réponses en toutes lettres à un code numérique, le système CART contient un fichier de référence composé d'expressions (p. ex., noms de localité) ainsi que des codes numériques correspondants (p. ex., des codes de la CGT). On utilise aussi une stratégie d'analyse pour normaliser, le plus possible, les nombreuses variations des réponses en toutes lettres, comme les abréviations, les fautes d'orthographe courantes et les mots omis ou superflus. Le processus de codage automatisé est composé d'une série d'étapes successives au cours desquelles on apparie la réponse en toutes lettres du répondant à une expression dans le fichier de référence.

Chaque service spécialisé utilise le CART pour élaborer son propre fichier de référence et sa propre stratégie d'analyse particuliers aux variables du recensement, comme le principal domaine d'études. Dans le cas de la mobilité, le codage des données sur les noms de localité représente l'application géographique d'un système de codage général.

4.3 Fichier de référence géographique

L'adaptation d'un logiciel de codage général comme CART à une application géographique a comporté la nécessité de travailler avec certaines restrictions et contraintes. Par exemple, la version de CART qui est exploitable pour le recensement de 1991 ne fait pas la distinction entre l'ordre des mots de sorte que, deux noms de localité distincts comme Spring Hill et Hill Spring, sont interprétés comme la même expression. En plus des contraintes du système CART, la conception des questions sur la mobilité elles-mêmes imposait certaines restrictions sur l'utilisation de fonctions de CART comme l'option "filtre". Si l'on avait demandé la province tout d'abord puis le comté et ensuite la ville, la base de données ou le fichier de référence aurait pu être structuré de façon à permettre l'appariement des expressions sur le nom de localité province par province. Toutefois,

puisque l'on demande tout d'abord la ville et à la fin la province, il est moins probable que les répondants indiquent la province et, par conséquent, il serait difficile de faire un premier appariement selon la province et ensuite par élément géographique plus détaillé. Il a donc fallu tenir compte de ces restrictions quand on a élaboré le fichier de référence géographique et les stratégies d'analyse et d'appariement pour le nom de localité fourni en réponse aux questions sur la mobilité.

Le fichier de référence géographique utilisé pour le codage automatisé de la réponse aux questions sur la mobilité est basé sur le Cahier des codes des noms de localité (CCNL). Les données du fichier de référence ont été tirées du CCNL sous forme lisible par machine renfermant: les noms de localité pour les subdivisions de recensement (SDR), les comtés (DR) et les provinces; les abréviations pour les genres de SDR (p. ex., C pour cité, R pour réserve); et les codes de la CGT correspondants. De plus, le CCNL renferme des noms de localité courants mais qui ne correspondent pas à une SDR, comme des noms de quartier et des noms de localité non constituée ainsi que les codes de la CGT auxquels ils correspondent.

Le fichier de référence a été élaboré de façon à maximiser l'attribution automatique de codes. La stratégie a entraîné l'élaboration d'une hiérarchie de réponses sous forme de noms de localité classés selon qu'ils étaient plus ou moins complets. Des enregistrements renfermant des combinaisons 'partielles' de renseignements sur les noms de localité ont été inclus dans le fichier de référence. D'après les réponses en toutes lettres recueillies lors du recensement de 1986, on sait que les répondants ne fournissent pas des réponses complètes. La réponse habituelle ou la plus fréquente est composée du nom de localité correspondant à la SDR ainsi que de la province (p. ex., Hull, Québec). Parfois seul le nom de la ville est donné sans la province (p. ex., Toronto). Dans le cas de noms de localité 'répétés' dans le même comté, on doit disposer du genre de SDR pour faire la différence entre les deux localités (p. ex., le township de Kingston par opposition à la ville de Kingston) et parfois des répondants indiqueront le genre de SDR quand il est pertinent. Ainsi, bien qu'on demande aux répondants d'indiquer le nom correspondant à la SDR, le nom correspondant au DR et la province, il arrive souvent qu'on ne nous fournisse que des renseignements partiels.

Dans le cas de réponses partielles, l'affectation automatisée de codes dépend du genre de réponse. Si, par exemple, une réponse partielle correspond à un nom de localité unique, le système peut alors attribuer un code. Si, toutefois, une réponse partielle correspond à plus d'une localité, ou s'il n'y a pas suffisamment de renseignements comme lorsque seulement la province est indiquée, alors le système ne peut attribuer un code et la réponse en toutes lettres est soumise à un traitement manuel. Au cours de ce traitement, il se peut que des pseudo codes soient attribués quand on ne peut résoudre la difficulté que posent des réponses en toutes lettres partielles pour l'attribution de codes. En plus des codes de la CGT, le fichier de référence renferme ces pseudo codes et signale les codes de la CGT qui sont préférés et que les codeurs attribuent au cours du traitement manuel.

Afin de pouvoir traiter la grande variété de réponses et pour obtenir le plus grand nombre possible d'appariements uniques, nous avons conçu une hiérarchie géographique d'expressions de réponse dans le fichier de référence. Cette structure de réponses et les conditions qu'elles doivent satisfaire pour qu'un code soit attribué automatiquement sont présentées sommairement dans la figure 4. La structure de ce fichier de référence a entraîné la création d'environ 170 000 enregistrements correspondants à environ 40 000 noms de localité et à 6 000 SDR.

4.4 Exemples de fichier de référence géographique

L'exemple de Kingston présenté à la figure 5 sert à illustrer à la fois la structure hiérarchique des expressions contenues dans le fichier de référence géographique ainsi que le problème posé par les réponses partielles. Ce tableau montre les quatre composantes des renseignements géographiques qu'une réponse pourrait fournir: le nom de la SDR, le genre de SDR, le DR et la province. Nous avons huit variantes possibles des expressions dans le fichier de référence pour la cité de Kingston. Dans le premier cas, nous avons une réponse complète 'cité de Kingston, comté de Frontenac, Ontario' qui correspond à une localité unique, pour laquelle un code de la CGT peut être attribué automatiquement. Les trois expressions suivantes renfermant des réponses partielles qui comprennent l'expression 'cité de Kingston' correspondent aussi aux codes de la CGT pour la cité de Kingston. Toutefois, les trois expressions de réponses partielles qui suivent comprenant simplement le mot 'Kingston' combiné avec comté de Frontenac et (ou) Ontario, laissent supposer deux possibilités - soit la cité de

Kingston, soit le township de Kingston, les deux se trouvant dans le comté de Frontenac. Ces trois expressions ne permettraient pas d'attribuer automatiquement un code, mais les cas seraient soumis à un traitement manuel. Finalement, dans le scénario qui envisage le pire, nous avons tout simplement 'Kingston' comme réponse partielle, ce qui correspond à six noms de localités possibles au Canada, un tel cas serait donc soumis à un traitement manuel.

Figure 4: Hiérarchie des réponses dans le fichier de référence géographique

Expression de réponse					Attribution automatique d'un code
	NOM DE LOCALITÉ/ SDR	DR	GENRE DE SDR	PROV./ TERR.	L'expression de réponse entraînera l'attribution automatique d'un code:
1.	X	X	X	X	Toujours, l'expression de réponse est unique au Canada
2.	X	X		X	Seulement si le nom de localité est unique dans le DR
3.	X			X	Seulement si le nom de localité est unique dans la province
4.	X	X			Seulement si le nom de localité est unique dans le DR et que l'expression de réponse est unique au Canada
5.	X				Seulement si le nom de localité est unique au Canada
6.	X		X	X	Seulement si le nom de localité et le genre de SDR sont uniques dans la province
7.	X	X	X		Seulement si l'expression de réponse contenant le nom de localité, le DR et le genre de SDR est unique au Canada
8.	X		X		Seulement si l'expression de réponse contenant le nom de localité et le genre de SDR est unique au Canada

4.5 Stratégie d'analyse pour normaliser les expressions

Dans le cas de la mobilité, l'analyse était minimale comparativement au cas d'autres variables codées comme le "principal domaine d'études". L'analyse est minimale parce que, pour les noms de localité, les petites différences sont importantes, p. ex., Saint John, St. John's. Les exemples d'analyse de noms de localité comprennent: les lettres doubles, p. ex., BB devient B; les noms de province normalisés, p. ex., toute version prévue du mot Ontario devient ON, alors que les noms de provinces composés de deux mots sont combinés en un seul, p. ex., toutes les versions de Colombie-Britannique deviennent BC. L'analyse est aussi utilisée pour tenir compte des fautes d'orthographe et des abréviations courantes relatives aux noms de localité.

Figure 5: Exemple de fichier de référence géographique basé sur 'Kingston'

Expressions de réponse				Nombre de localités au Canada auxquelles l'expression peut correspondre
SDR	Genre de SDR	DR	Province	
1. Kingston	Cité	Frontenac	Ontario	1
2. Kingston	Cité	Frontenac		1
3. Kingston	Cité			1
4. Kingston	Cité			1
5. Kingston	Cité	Frontenac		2
6. Kingston		Frontenac	Ontario	2
7. Kingston			Ontario	2
8. Kingston				6

Figure 6: Exemple de traitement manuel assisté par ordinateur pour 'Kingston, Ontario'

```
CDMANMOB     RECENSEMENT DE LA POPULATION DE 1991/CODAGE AUTOMATISÉ          04/06/91
MMANUELI  CODAGE MANUEL - MOBILITÉ DEPUIS 5 ANS AU CANADA                   12:00:00.0
Réponse en toutes lettres à coder                          Genre   Code

KINGSTON ON                                                        M   __

Expressions produites par CART                                     Codes   (S)élection
KINGSTON ON (RÉP.)                                 __              3510009   - -
KINGSTON ON (RÉP.)                                 __              3510011   -
KINGSTON ON (RÉP. - ON)                            __              9935172   - -

N° d'identification : 35009020 100 2 22
================================================================================
Donnés pour la même question fournies par chaque
membre du ménage                                   N° de la personne :  1
Cases à cocher

_____ _____ _____ _____ _____ _____ _____ _____ _____ _____

Répondre en toutes lettres
KINGSTON ONTARIO

Introduire-PF1---PF2-----PF3------PF4-------PF5-------PF6-------PF7-----PF8---PF9--------PF10------PF11----PF12--
        AIDE HAUT BAS  <<<<  >>>>  PLUS  +HAUT  +BAS  RENVOI  VALIDER  RENDRE  SORTIE
                                                                        PERMANENT
```

```
NDISMOB          RECENSEMENT DE LA POPULATION/CODAGE AUTOMATISÉ          04/06/91
MMOBILITÉ1                          Mobilité                            12:00:00.0

                 IL Y A 1 AN
                 Au Canada          : MEMPRTE
                                    : _____
                 À l'extérieur du Canada : _____
                                    : _____
                 IL Y A 5 ANS
                 Au Canada          : AUTRVIL
                                    : KINGSTON ON
                                    _____
                 À l'extérieur du Canada : _____
                                    : _____
                                    : _____
          Lien avec la personne 1   : ÉPOUX/ÉPOUSE DE LA PERSONNE 1
               Date de naissance    : 23-06-1962

N° d'identification : 35009020 100 2 22

Introduire –  PF1 — PF2 — PF3 — PF4 — PF5 — PF6 — PF7 — PF8 — PF9 — PF10 — PF11 — PF12 –
              AIDE                                                          SORTIE
```

4.6 Traitement manuel assisté par ordinateur

La phase de traitement manuel qui fait partie du codage automatisé est assistée par ordinateur. Pour la majorité des réponses en toutes lettres qui sont transmises afin de subir un traitement manuel, CART propose des combinaisons de codes/d'expressions. Dans certains cas, le système ne peut proposer de codes pour les réponses en toutes lettres qui ne sont pas des réponses géographiques, (p. ex., 'sur la ferme') ou qui renferment des fautes d'orthographe. Dans l'exemple de la figure 6, le système CART propose trois codes possibles pour la réponse en toutes lettres 'Kingston, Ontario' pour la personne 1 du ménage: le code de la CGT pour la cité de Kingston (3510009); le code de la CGT pour le township de Kingston (3510011) et le pseudo code pour Kingston, Ontario (9935172). Si les codeurs peuvent obtenir des renseignements additionnels sur le ménage, comme une réponse plus détaillée fournie par d'autres membres, alors ils choisissent le code approprié. Le système fournit en direct des renseignements additionnels sur le ménage du répondant auxquels le codeur peut accéder facilement. De plus, des documents de référence géographique additionnels sont fournis aux codeurs. Si, pendant le traitement manuel, le codeur ne peut déterminer la réponse appropriée il attribue alors soit un pseudo code, soit un code réel de la CGT préféré dans le cas des problèmes posés par des noms de localité répétés. Dans les autres cas, on attribue la réponse 'impossible de coder'. Dans l'exemple de la figure 6, l'autre personne dans le ménage, l'époux/l'épouse de la personne 1, n'a pas fourni de renseignements additionnels, elle n'a donné que la même

réponse en toutes lettres 'Kingston, Ontario'. Ainsi, dans ce cas, on attribue le pseudo code pour 'Kingston, Ontario'.

4.7 Solution des problèmes présentés par les pseudo codes

Le problème présenté par les pseudo codes attribués par les codeurs au cours de la phase de traitement manuel du processus de codage automatisé sont réglés pendant le contrôle et l'imputation des données du recensement. Ces pseudo codes sont répartis selon N voies d'après le nombre réel de codes de la CGT auxquels ils correspondent, généralement entre 2 et 8. Nous utilisons à nouveau l'exemple de Kingston, Ontario. Pendant le traitement manuel, le codeur ne peut trouver de renseignements additionnels sur le ménage afin de résoudre le problème posé par la réponse 'Kingston, Ontario' qui pourrait correspondre soit à la cité de Kingston, soit au township de Kingston. Ainsi, un pseudo code, 9935172 est attribué à l'étape du traitement manuel.

Au cours du contrôle et de l'imputation, les réponses auxquelles on a attribué le pseudo code 9935172 pour Kingston, Ontario, sont réparties proportionnellement à la taille de la population entre les deux codes de la CGT pour la cité de Kingston et pour le township de Kingston (3510009 et 3510011, respectivement). D'autres problèmes, en plus de ceux posés par les pseudo codes, sont réglés à l'aide d'imputations à partir de données sur un membre de la famille ou à partir d'enregistrements donneurs, pour des cas tels que: les réponses pour lesquelles on n'a pu attribuer un code; les réponses manquantes dans des cas où une réponse en toutes lettres aurait dû être fournie et les réponses partielles où la seule réponse fournie est le 'comté' et (ou) la 'province'. L'attribution des codes de la CGT aux 'SDR' du 'lieu de résidence il y a 5 ans' est parachevée au cours des opérations de contrôle et d'imputation.

5. RÉSULTATS ET AVANTAGES DU CODAGE AUTOMATISÉ DES DONNÉES SUR LA MOBILITÉ

5.1 Résultats du codage

Compte tenu du fait qu'à l'heure actuelle le codage automatisé des données recueillies lors du recensement de 1991 bat son plein, seuls des résultats préliminaires sur le codage automatisé des données sur la mobilité sont disponibles. Pour fournir certaines données de base, les résultats de la phase de recherche et d'essai du codage automatisé des données sur la mobilité ainsi que du codage effectué jusqu'à ce jour, sont présentés au tableau 3.

Tableau 3: Résultats du codage automatisé des noms de localité pour la question sur la mobilité

Certains indicateurs	Recherche et essai (basé sur un échantillon de réponses en toutes lettres fournies lors du recensement de 1986)	Production jusqu'à ce jour (novembre 1991)
• Pourcentage de réponses en toutes lettres pour lesquelles on peut attribuer automatiquement un code:		
Intervalle	60 - 80%	70 - 80%
Moyenne	70%	75%
• Taux d'erreur lors du codage		
a) Basé sur un échantillon de réponses en toutes lettres lors du recensement de 1986 codées par le système:		
- automatisé	1%	Sans objet
- manuel (résultats du codage pour le recensement de 1986)	4%	
b) Avec une mesure de l'erreur incorporée pour la solution des problèmes présentés par les noms de localité répétés:		
- solutions automatisée et manuelle combinées (prévue)	Sans objet	1 - 2%
- solution manuelle (estimation basée sur les résultats du codage manuel de 1986)	8%	Sans objet

- **Recherche et essai**

Le codage automatisé des données sur la mobilité a été élaboré à l'aide d'un échantillon de réponses en toutes lettres à la question sur la mobilité posée lors du recensement de 1986. Pendant la recherche et l'essai, on a pu attribuer automatiquement un code pour près de 80% des noms de localité en toutes lettres. Les taux d'erreur de codage, basés sur les réponses en toutes lettres auxquelles le système automatisé a attribué un code, était de 1% pour le codage automatisé comparativement à 4% pour l'attribution manuelle de codes effectuée, pour ces mêmes enregistrements, par les codeurs en 1986. Ces taux d'erreur ne comprennent pas l'attribution de codes pour les noms de localité répétés, qui comportent l'utilisation de pseudo codes ou de codes préférés. (On n'a pas utilisé de pseudo codes lors du codage des données sur la mobilité en 1986.) Des détails sur la recherche et sur l'essai effectués au commencement des travaux sur le codage automatisé des données sur la mobilité sont présentés dans le rapport de Norris et Kirk (1989).

- **Production jusqu'à ce jour**

Les résultats préliminaires basés sur la production jusqu'à ce jour, qui sont présentés au tableau 3, montrent qu'en moyenne, le système automatisé peut attribuer un code aux noms de localités en toutes lettres dans 75% des cas. Pour les divers passages de production réalisés jusqu'ici, l'affectation automatisée de codes a été réalisée dans entre 70 et 80% des cas. Bien qu'il soit difficile de mesurer avec précision les taux d'erreur de codage, puisque les travaux ont commencé depuis peu, on prévoit que l'attribution automatisée et manuelle de codes en 1991 donnera des taux d'erreur compris entre 1 et 2%. Cela représente une amélioration considérable par rapport au taux d'erreur estimé d'environ 8% pour le codage manuel effectué en 1986, taux qui comprend une mesure de l'erreur pour la solution des problèmes présentés par les noms de localités répétés.

5.2 Avantages additionnels offerts par le codage automatisé

En plus d'une précision accrue du codage, le codage automatisé présente des avantages additionnels sur le plan de la production ainsi que de l'évaluation et il permet d'autres réalisations.

- **Traitement manuel plus efficient**

Tout d'abord, la phase de traitement manuel du codage automatisé est beaucoup plus efficiente que les opérations de codage manuel de 1986. Comme le traitement manuel est assisté par ordinateur, cela permet d'éliminer le fardeau excessif qu'imposait l'utilisation de documents sur support papier lors du recensement précédent - on ne perd plus de temps à feuilleter des dossiers et des livres de codes. Les réponses dans les unités de travail sont groupées en ordre alphabétique. Les codeurs prennent beaucoup moins de temps pour traiter une réponse en toutes lettres en 1991 qu'en 1986.

- **Meilleur contrôle et meilleure surveillance**

Le codage automatisé a permis d'assurer un meilleur contrôle et une meilleure surveillance des opérations de codage pour 1991. Lors du recensement précédent, le codage des données sur la mobilité a été effectué dans les bureaux régionaux à travers le Canada. Maintenant, les codeurs peuvent facilement consulter des spécialistes parce que les opérations sont centralisées. De plus, les résultats peuvent être plus facilement surveillés car on peut obtenir des rapports informatisés sur le codage pour les travaux de production.

- **Coûts de production inférieurs**

Les estimations préliminaires basées sur la production jusqu'à ce jour montrent que les coûts de production du codage des données sur la mobilité pour 1991 seront considérablement inférieurs aux coûts correspondants en 1986. Des estimations grossières pour 1991 laissent prévoir un coût d'environ $260 000 et de quatre années-personnes comparativement à $466 000 et près de 17 années-personnes en 1986. Une partie importante du coût pour 1991 est attribuable à la saisie des réponses alphanumériques pour les noms de localité.

• **Contrôle de la qualité (CQ) automatisé**

Le contrôle de la qualité (CQ) automatisé permet de suivre et de surveiller facilement les attributions tant manuelles qu'automatisées de codes à des réponses en toutes lettres non analysées. Un journal des réponses en toutes lettres non analysées et des codes correspondants est tenu à jour avec les données sur le CQ du système et, de plus, un système spécial de consultation nous permet d'examiner les affectations manuelles de codes. Élément très important, avec la méthode du CQ, il est possible de relever et de corriger les erreurs systématiques (au moyen de corrections effectuées après la production). Les stratégies d'assurance de la qualité appliquées au codage automatisé pour le recensement de 1991 sont décrites dans l'article de Ciok (1991).

• **Meilleure compréhension de la qualité des données pour les petites régions**

Les résultats du codage automatisé des véritables noms de localité en toutes lettres ainsi que les codes correspondants attribués à ces réponses, se révéleront utiles pour l'analyse de la qualité des données régionales. Pour la première fois, il sera possible de connaître l'importance de divers problèmes reliés à la réponse, tels que: la déclaration de noms de quartier ou de noms communs plutôt que de noms officiels de localité, les fautes d'orthographe, les réponses incomplètes, les noms de comté erronés, l'utilisation de plus d'un nom de localité et de plus d'une adresse de voirie. L'évaluation des données sur les noms de localité précisés pour les questions sur la mobilité, obtenues par codage automatisé permettra de mieux comprendre et de mieux évaluer la qualité des données sur la migration régionale pour 1991.

• **Information pour le prochain recensement**

L'expérience acquise avec le codage automatisé lors du recensement de 1991 fournira des informations pour l'élaboration des éléments relatifs à la mobilité au cours du prochain recensement. Par exemple, en matière de production pour le prochain recensement, on pourrait augmenter le pourcentage des noms de localité en toutes lettres pour lesquels un code peut être attribué automatiquement, rationaliser davantage le traitement manuel et améliorer, lors de la phase du contrôle et de l'imputation, la solution des problèmes relatifs aux codes de la CGT. L'évaluation et l'analyse des données sur les noms de localité recueillies lors du recensement de 1991 permettra d'acquérir de nouvelles connaissances sur la façon dont les répondants utilisent les noms de localité et dont ils comprennent les questions sur la mobilité ainsi que de se faire une meilleure idée à ce sujet. Ce genre de nouveaux renseignements, qu'il est maintenant possible d'acquérir parce qu'on utilise un codage automatisé, se révélera inestimable au cours de l'élaboration de nouvelles questions sur la mobilité pour le prochain recensement.

6. CONCLUSION

En conclusion, les objectifs originaux du codage automatisé des données sur les noms de localité fournis en réponse aux questions sur la mobilité sont atteints et, en même temps, le codage automatisé fournit des avantages additionnels pour la production, l'évaluation et pour d'autres réalisations. Non seulement la précision du codage et, par conséquent, la qualité des données sur la migration régionale sera améliorée, mais nous acquerrons une meilleure compréhension de la qualité des données sur les noms de localité. Par exemple, il sera maintenant possible de savoir à partir de quel genre de renseignements des codes ont été attribués, car on pourra examiner les réponses en toutes lettres fournies ainsi que les codes correspondants; une opération qu'il était impossible d'effectuer systématiquement en 1986 à moins de nous reporter aux questionnaires eux-mêmes. De plus, le système automatisé donne aussi de nouveaux renseignements sur l'utilisation que font les répondants du nom de localité - par exemple, la mesure dans laquelle les répondants déclarent des noms de quartier plutôt que des noms de municipalité. Finalement, l'évaluation de ces données recueillies en 1991 fournira des données pour produire un fichier de référence géographique amélioré et pour élaborer des questions sur la mobilité pour le prochain recensement.

REMERCIEMENTS

Les auteurs désirent remercier les personnes mentionnées ci-après, travaillant à Statistique Canada, pour leurs contributions: George Mori et Art Gardner pour leurs suggestions et conseils inestimables pendant l'étape de recherche sur le codage automatisé en matière de mobilité; Rick Ciok pour ses travaux qui ont permis d'incorporer les besoins spéciaux en matière de mobilité dans le système de CQ; Anna Rigakis pour son apport lors de l'élaboration des stratégies de traitement manuel en matière de mobilité et Judy Kirk pour ses rapports et observations sur les questions relatives à la production. Il faut aussi remercier Audrey Miles qui a effectué le traitement de texte.

BIBLIOGRAPHIE

Ciok, R. (1991). The use of automated coding in the 1991 Canadian census of population, papier présenté au 1991 American Statistical Association, Atlanta, Georgia.

Fyffe, S., Gardner, A., Ladouceur, D., Miller, D., Rakhra, M., et Swain, S. (1988). Research and testing of an automated coding system for census using the ACTR system (Automated Coding by Text Recognition): Analysis Report, Statistique Canada, 1991 Census of Canada, Ottawa. (Rapport interne, octobre 1988.)

Norland, J.A. (1989). Evaluation of mobility data form the 1986 Census, Statistique Canada, Division de démographie (Rapport interne, février 1989).

Norris, M.J. (1990). Guide à l'intention des utilisateurs - Données du recensement de 1986 sur la mobilité, Statistique Canada, novembre 1990.

Norris, M.J. (1989). National Census Test, Report No. 16, Questions 19, 20, 21: Mobility, Statistique Canada, août 1989.

Norris, M.J., et Kirk, J. (1989). Research and testing of an automatic coding system for the mobility status variable using the ACTR System: Analysis report. Statistique Canada, 1991 Census of Canada, Ottawa Rapport interne, avril 1989).

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

UN ASSISTANT EXPERT EN ANALYSE STATISTIQUE ET EN DÉCOUVERTE DE LA CONNAISSANCE

J. Muzard[1], E. Falardeau[1] et M.G. Strobel[2]

RÉSUMÉ

Ce papier décrit le système STATEX, un assistant expert en analyse statistique et en découverte de la connaissance. L'architecture du système est basée sur l'approche tableau-noir afin de permettre la communication entre l'usager, les modules à base de connaissances et les modules statistiques. Un environnement intégré a été construit, un micro-monde avec icônes et graphiques, dans le but de combiner la consultation, l'exécution et l'interprétation dans un système. L'usager sera assisté par le système dans toutes ces phases de l'analyse statistique. Un modèle pour l'analyse des données, basé sur une théorie de la perception humaine, a été développé. C'est à dire que pour distinguer un objet, pour le "voir", il faut le contraster, et que les relations entre les objets sont identifiées par leur association. Ce modèle a guidé la fabrication d'un système flexible et graphique pour l'analyse statistique. Le logiciel CSPAL est utilisé pour les calculs statistiques. STATEX assiste l'usager à travers les étapes de l'analyse quantitative et crée un environnement qui change de façon dynamique selon les besoins de l'usager et les exigences reconnues de la pratique statistique. Le système propose alors une interprétation des résultats dans le langage de l'usager. Une méthode d'analyse statistique est proposée et le système peut être utilisé pour l'acquisition de connaissances à partir de données numériques.

MOTS CLÉS: Assistant expert; système statistique; système à base de connaissances; modélisation de la connaissance experte et de l'usager; interface graphique.

1. INTRODUCTION

Dans sa lutte pour la survie, un organisme doit réaliser un bon accouplement structurel avec son milieu environnant (Maturana et Varela 1984). Pour cela, il apprend à faire des distinctions. L'environnement est très complexe et changeant. L'homme s'est donné un outil, parmi d'autres, pour réaliser sa connaissance de l'environnement. Il s'agit de la cueillette et l'analyse de données numériques, qui fait appel à des méthodes et techniques mathématiques. Des systèmes puissants ont été réalisés sur ordinateur pour faire le travail de calcul statistique. Mais l'analyse de données fait aussi appel à des méthodes plus ou moins formalisées, à des stratégies d'analyse et des règles d'interprétation de résultats, afin de révéler des objets, faits ou des événements significatifs selon le domaine du chercheur. Afin de modéliser la connaissance sur ces méthodes, stratégies et règles, nous avons utilisé les méthodes et techniques de l'intelligence artificielle et des sciences cognitives pour développer un systèmes à base de connaissances qui facilite l'analyse des données. L'intelligence artificielle nous permet de développer des systèmes dans lesquels il est possible de représenter la connaissance d'un domaine, celui de l'analyse des données. Les sciences cognitives nous apportent les fondements théoriques et pratiques qui permettent de faire en sorte que l'outil développé tienne compte des caractéristiques de l'usager en fonction de la tâche à réaliser. Un environnement artificiel a été développé sur ordinateur, un micro-monde, dans le but de faciliter la découverte de connaissances.

---

[1] J. Muzard et E. Falardeau, Centre Canadien de recherche sur l'informatisation du travail. 1575, boulevard Chomedey, Laval (Québec), Canada H7V 2X2.

[2] G. Strobel, Université de Montréal, Case Postale 6128, Succ. A. Montréal (Québec), Canada H3C 3J7.

Les logiciels statistiques actuels très puissants ne sont pas très utiles pour les utilisateurs qui ne connaissent pas suffisamment la stratégie, les méthodes et les règles de l'art de l'analyse de données (Gale 1986). La grande accumulation de résultats est aussi difficile pour les usagers qui n'ont pas la connaissance pour interpréter les résultats de l'analyse. Le projet STATEX a comme but: a) de fournir aux décideurs un moyen de faire des analyses afin de dégager des tendances, comparer, prédire; b) de fournir aux chercheurs et aux cogniticiens un outil de découverte de la connaissance; c) de faciliter l'analyse des données contenues dans les bases de données; d) de modéliser la connaissance dans le domaine de l'analyse des données; e) de fabriquer un système intelligent destiné à fournir un support aux usagers dans leur tâche; f) d'augmenter la connaissance dans le domaine de l'informatisation du travail avec les apports de l'intelligence artificielle et les sciences cognitives dans un domaine complexe; g) de contribuer à une meilleure compréhension de la pensée humaine.

2. MODÈLES, MÉTAPHORES ET ANALOGIES HEURISTIQUES

Pour comprendre l'environnement nous faisons un acte de distinction, c'est-à-dire que nous séparons un objet du contexte (Maturana et Varela 1984). Un objet, une entité ou une unité contient les distinctions implicites. Les distinctions que nous faisons sont adaptatives et n'impliquent pas une existence réelle de l'objet. Avec ces catégories, nous produisons des schémas qui sont des liens entre des catégories, des unités de connaissances ou shunks. Pour comprendre, il semble qu'il faut être capable de faire les distinctions nécessaires, avoir une connaissance initiale du domaine et être capable de passer d'une conception erronée à une nouvelle conception (Carey 1990). La connaissance est 'restructurée' pendant le processus d'acquisition. Il y a un besoin de décrire la restructuration qui a lieu au cours d'une acquisition de la connaissance complexe. Dans l'environnement nous pouvons donc distinguer des objets, des situations et des événements. Les organismes biologiques ont besoin de s'orienter vis-à-vis des objets, d'identifier les situations, de prédire les événements pour augmenter leur accouplement structurel avec l'environnement en vue de leur survie. Ils possèdent des mécanismes qui leur permettent de percevoir des objets, de faire des distinctions, de faire des catégories et de les ordonner, d'associer des événements et de prédire des changements. De façon analogique, la statistique invite à une description des objets, des situations et des événements, à établir leurs différences et à faire des associations ou corrélations entre ceux-ci. Il y a une progression dans le temps et dans la complexité. Cette analogie invite d'abord à décrire les objets, les unités, pour ensuite établir leurs différences et leurs associations.

3. L'ANALYSE DE LA TÂCHE DE COOPÉRATION ENTRE DEUX EXPERTS

3.1 Analyse de la tâche

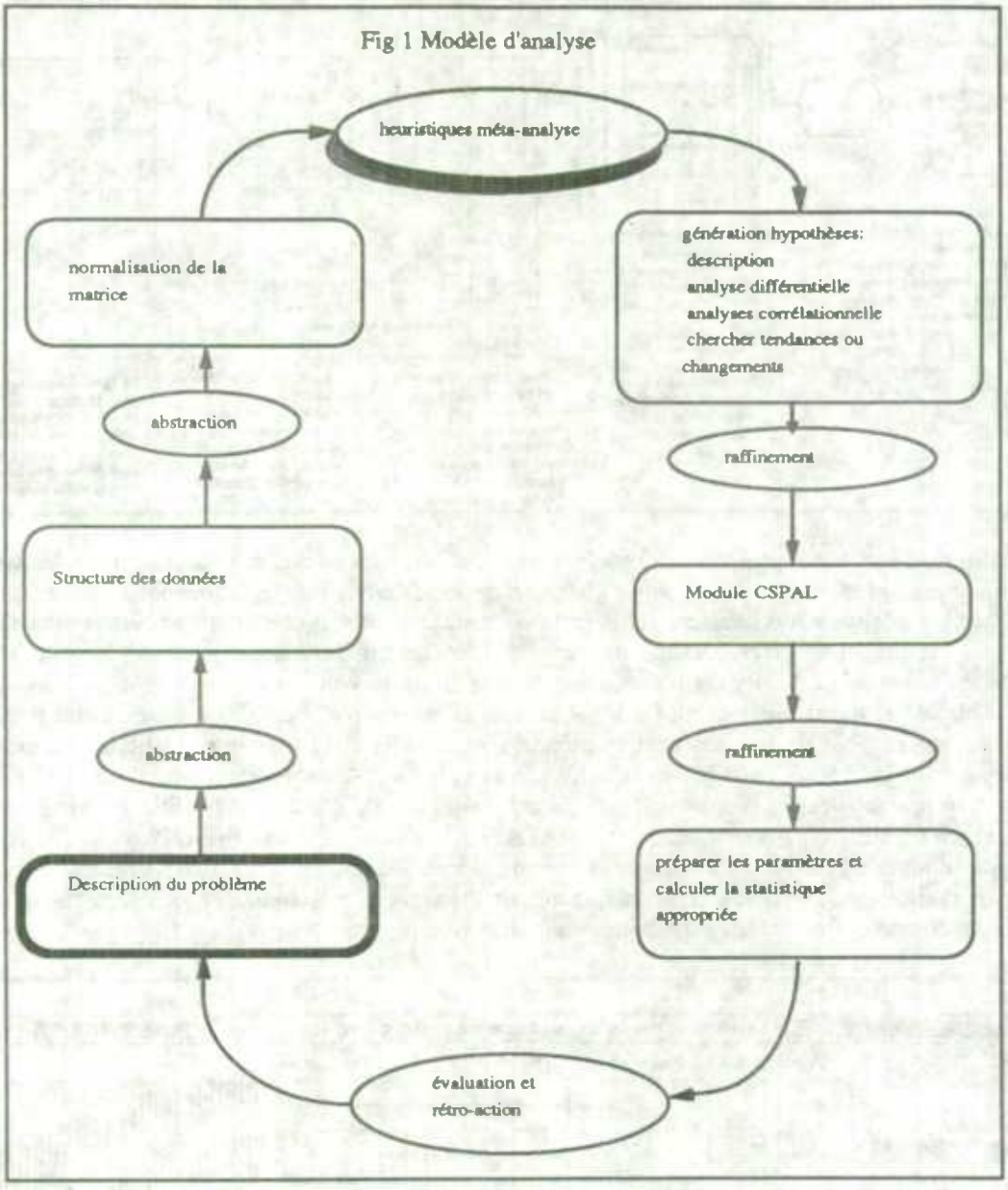
Une consultation entre un usager et un expert statisticien est caractérisée par la communication et la coopération entre deux experts en vue d'augmenter la connaissance dans le domaine de l'usager. Le client consulte parce qu'il cherche de l'assistance. Il communique les buts de sa recherche et apporte le fichier de données. Le statisticien pose des questions pour avoir des informations sur les données, sur les attributs du problème et la nature de la recherche. Par la suite, le statisticien recommandera et participera à l'observation et au nettoyage des données, à la structuration de la matrice de données, au choix de la stratégie et des méthodes, à faire l'analyse et à interpréter les résultats dans un langage compréhensible à l'usager. Le processus de résolution de problème est essentiellement une conversation pendant laquelle il y a un apprentissage réciproque. Le client est amené à augmenter ses connaissances en analyse des données et le statisticien augmente sa connaissance du domaine du client. Chacun des participants coopère en vue de l'objectif qui est de découvrir des connaissances nouvelles dans le domaine du client.

3.2 Le modèle de STATEX

À partir de ces observations, nous avons établi un modèle d'analyse (Fig. 1). Dans une première étape, il faut compléter la description du problème. Par un processus d'abstraction, les structures de données sont établies, telles que les variables, les groupes, les facteurs, le rang et le type des variables, etc. En fait, STATEX construit une représentation interne des données en exploitant les informations sur les données ou méta-données (Hand 1991; Oldford 1990). On arrive par un autre processus d'abstraction et de classification à la structure de la matrice. En utilisant les analogies heuristiques décrites précédemment, on génère des hypothèses de stratégie

comme par exemple la stratégie corrélationnelle. Par un processus de raffinement, on établi la méthode à utiliser et on prépare les paramètres pour calculer les statistiques appropriées. Les résultats sont alors présentés et interprétés dans le langage fourni par l'usager lors de la description du problème. À partir de ces résultats, une nouvelle boucle peut être réalisée. Ce modèle représente la façon de travailler de STATEX. L'architecture utilisée, le tableau-noir, facilite le processus de tenir compte des données, des connaissances du système, des résultats et des choix de l'usager (Muzard, Falardeau et Strobel 1991).

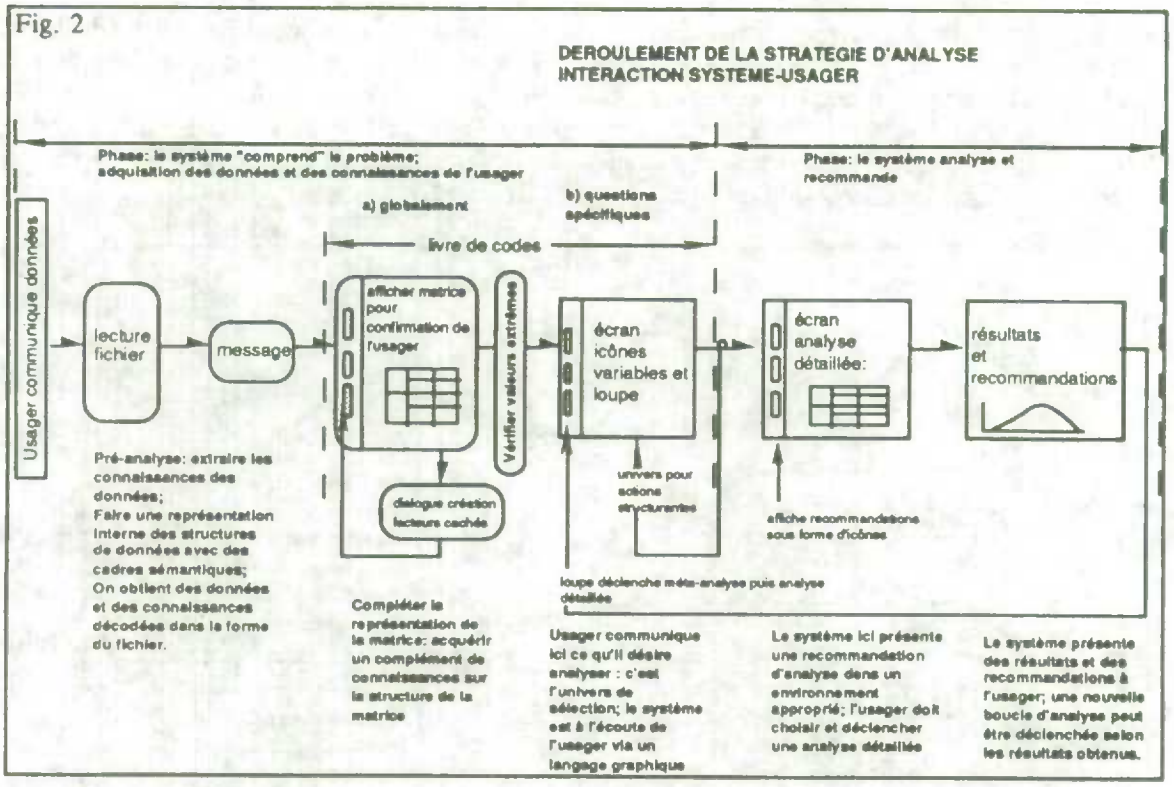
Fig 1 Modèle d'analyse



4. INTERACTION SYSTÈME-USAGER

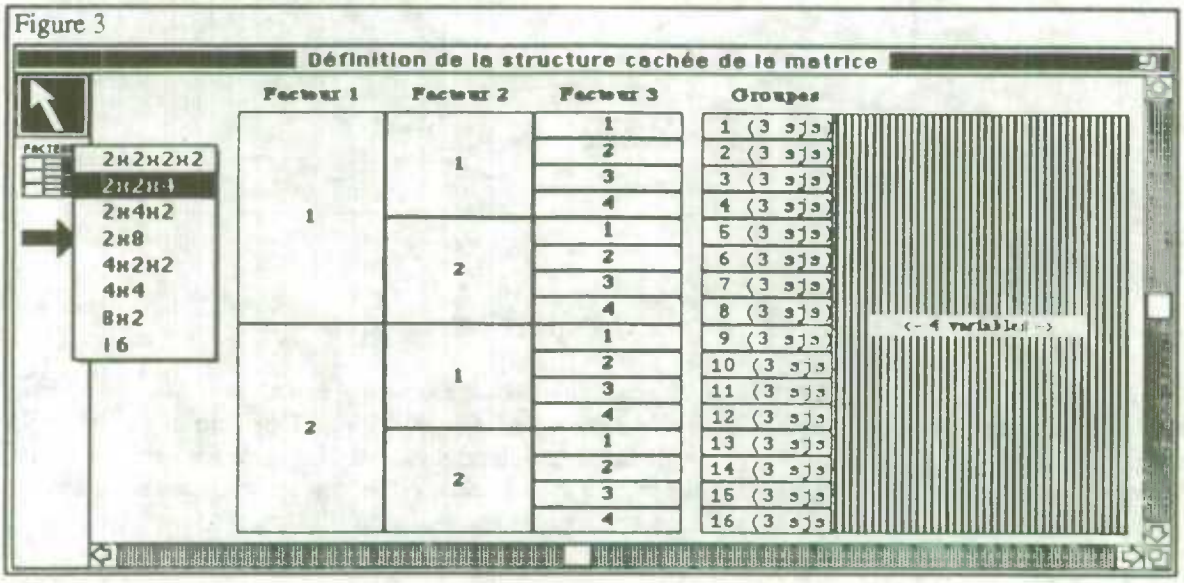
Le déroulement de la séance d'analyse est en tout temps sous le contrôle de l'usager. Afin de décrire l'interaction entre le système et une personne qui analyse des données, nous distinguerons deux phases générales: dans un premier temps le système acquiert les données et informations sur les données afin de se faire une représentation interne du problème. Cette représentation interne est la compréhension que le système a du problème. Dans une deuxième phase, le système analyse et recommande (Fig. 2).

Fig. 2



Au début, l'usager est invité à donner des informations sur lui-même, son nom et ses connaissances. Le système construira une représentation de l'usager et se chargera de la mettre à jour en fonction de l'utilisation du système. Le système adaptera son interface selon cette connaissance en utilisant comme principe d'offrir plus de possibilités à mesure que l'usager devient plus expert. L'utilisateur du système sera par la suite invité à communiquer ses données. Le fichier de données doit être de façon suivante: une ligne par sujet, les variables en colonnes séparées par des tabulations ou des espaces et les groupes séparés par des caractères alphabétiques. STATEX sera aussi capable de lire des fichiers préparés pour SPSS si l'on dispose du fichier de contrôle. STATEX suppose que les données proviennent d'échantillons au hasard. Avant la lecture des données, STATEX ouvrira une fenêtre qui montre un dessin d'une matrice et l'usager est invité à donner le schème de sa recherche à l'aide d'un menu déroulant, par exemple, 2 x 2 x 4. STATEX redessinera la matrice en fonction de ce qui a été indiqué pour donner du feed-back à l'usager et obtenir son accord. De cette façon, STATEX sera informé de la structure cachée de la matrice et en même temps l'usager pourra constater que le système a une représentation adéquate de ses données. L'usager, peut ainsi visualiser sa matrice (Fig. 3).

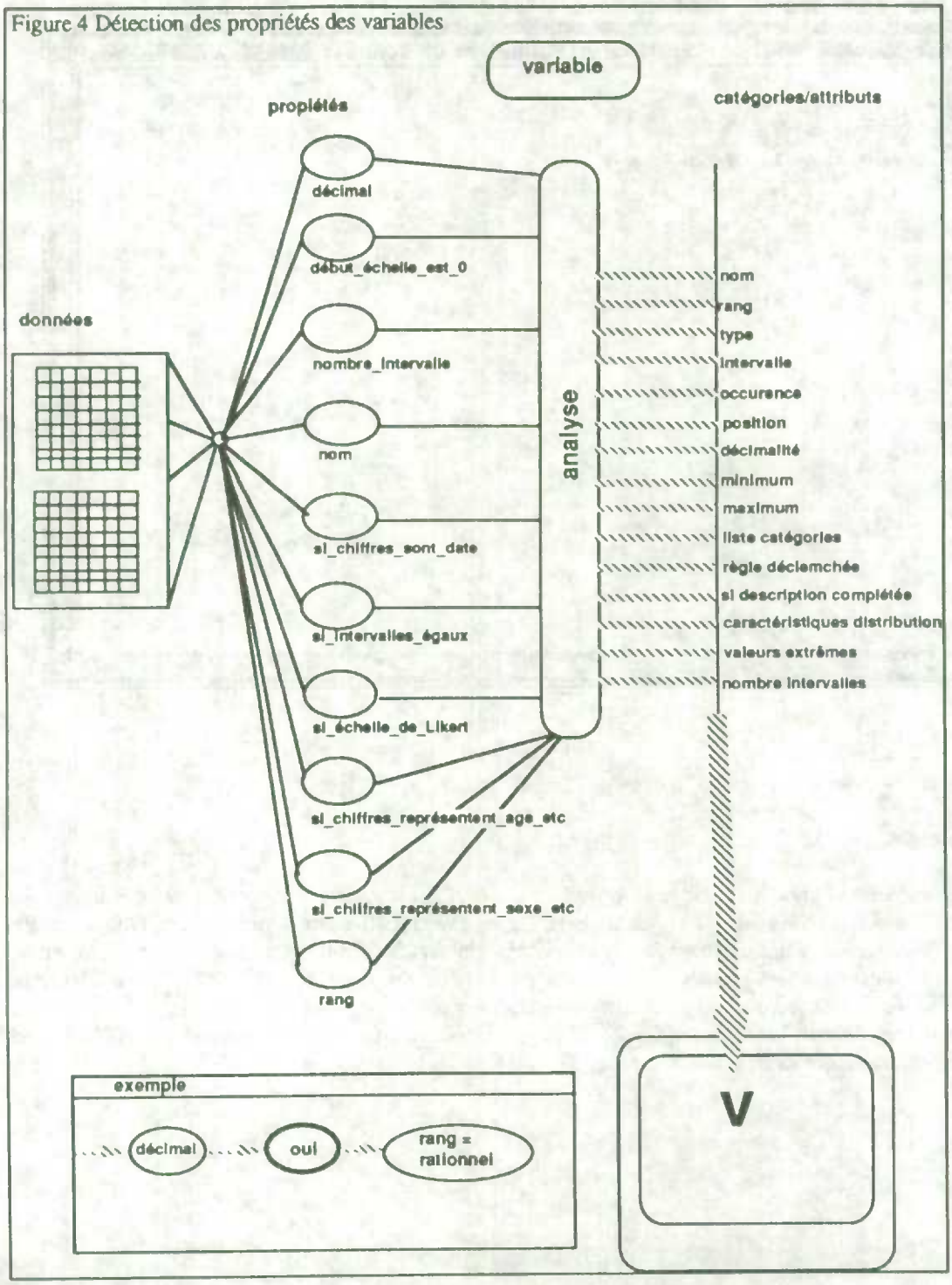
Figure 3



Afin de stimuler l'hémisphère droit du cerveau, celui qui est responsable de la reconnaissance des patterns et de voir l'ensemble, STATEX dessine chaque fois que possible une représentation graphique (Good 1983). Nous croyons que de cette façon le système peut avoir un couplage plus étroit avec l'usager et travailler à l'harmonisation des représentations mentales afin de faciliter la communication.

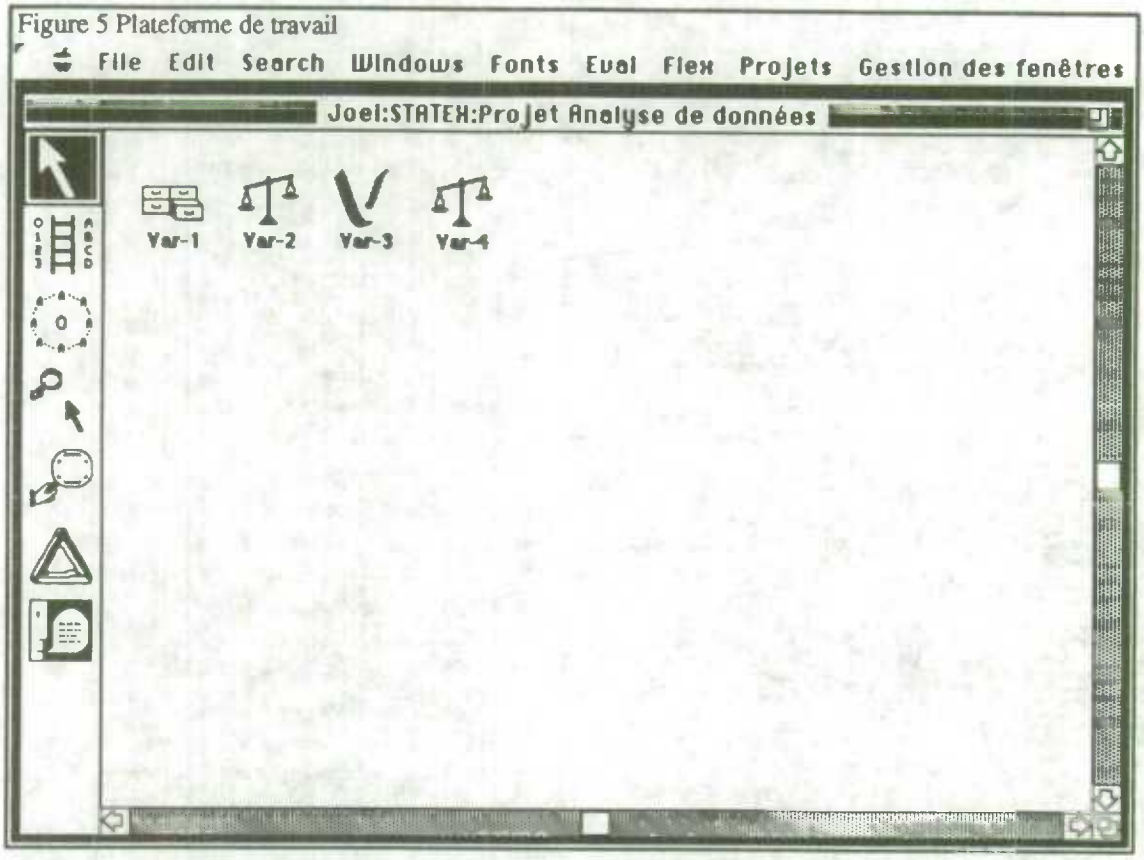
Lors de la pré-analyse, STATEX examinera variable par variable pour se construire une représentation interne des données. Cette représentation interne sera affichée via des icônes sur la plate-forme de travail. Elle sera aussi accessible aux modules à base de connaissances. Lors de la présentation des résultats, STATEX utilisera aussi cette information. Selon les propriétés, STATEX essayera de diagnostiquer le rang et le type de chaque variable (Fig. 4).

Figure 4 Détection des propriétés des variables



L'affichage à l'écran sera fonction de cette pré-analyse. Par exemple, si la variable possède un chiffre décimal, une règle sera déclenchée : si décimal = oui alors rang = rationnel. Et si le rang est rationnel, le type sera classé quantitatif ou 'capable-de-moyenne'. Les variables de type qualitatif seront montrées comme des icônes de tiroirs, l'analogie avec classification, et les variables de type quantitatif seront représentées par une balance, par analogie à leur capacité de 'capable-de-moyenne'. Les informations sur les données ou méta-données seront utilisées par STATEX pour guider la stratégie d'analyse. Lorsque la pré-analyse sera terminée, STATEX demandera à l'usager de compléter le livre de codes et affichera la plate-forme de travail (Fig. 5).

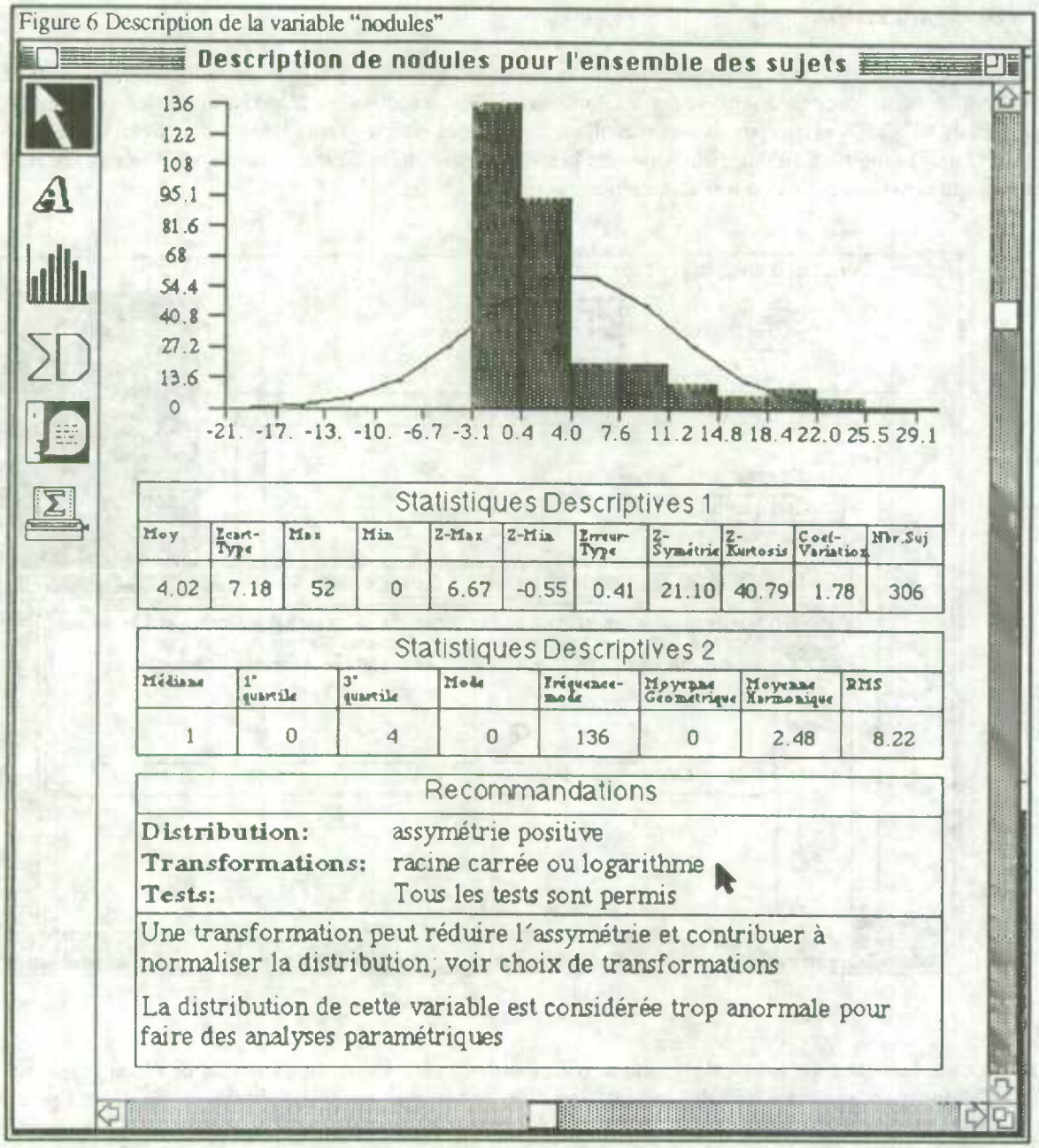
Figure 5 Plateforme de travail



4.1 Le livre de codes

Nous avons regroupé sous l'appellation 'livre de codes' les activités par lesquelles l'usager est invité à donner des informations sur les variables, les catégories, les facteurs, les mesures répétées, les données manquantes, etc. Par exemple, l'usager est invité à donner un nom aux variables qui est significatif pour lui. Lors de cette étape, l'usager sera invité à examiner les données et à voir s'il y a des données douteuses. STATEX lui montrera les données et la position de chaque donnée douteuse qui se trouve à plus ou moins trois écarts types. Le chercheur pourra faire la description de la variable (Fig. 6). Il pourra aussi modifier le rang proposé par STATEX. Si l'usager désire une assistance pour trouver le rang, STATEX la lui fournira.

Figure 6 Description de la variable "nodules"



Statistiques Descriptives 1

Moy	Ecart-Type	Max	Min	Z-Max	Z-Min	Erreur-Type	Z-Symétrie	Z-Kurtosis	Coef-Variation	Nbr.Suj
4.02	7.18	52	0	6.67	-0.55	0.41	21.10	40.79	1.78	306

Statistiques Descriptives 2

Médiane	1° quartile	3° quartile	Mode	Fréquence-mode	Moyenne Géométrique	Moyenne Harmonique	RMS
1	0	4	0	136	0	2.48	8.22

Recommandations

Distribution: assymétrie positive
Transformations: racine carrée ou logarithme
Tests: Tous les tests sont permis

Une transformation peut réduire l'assymétrie et contribuer à normaliser la distribution; voir choix de transformations

La distribution de cette variable est considérée trop anormale pour faire des analyses paramétriques

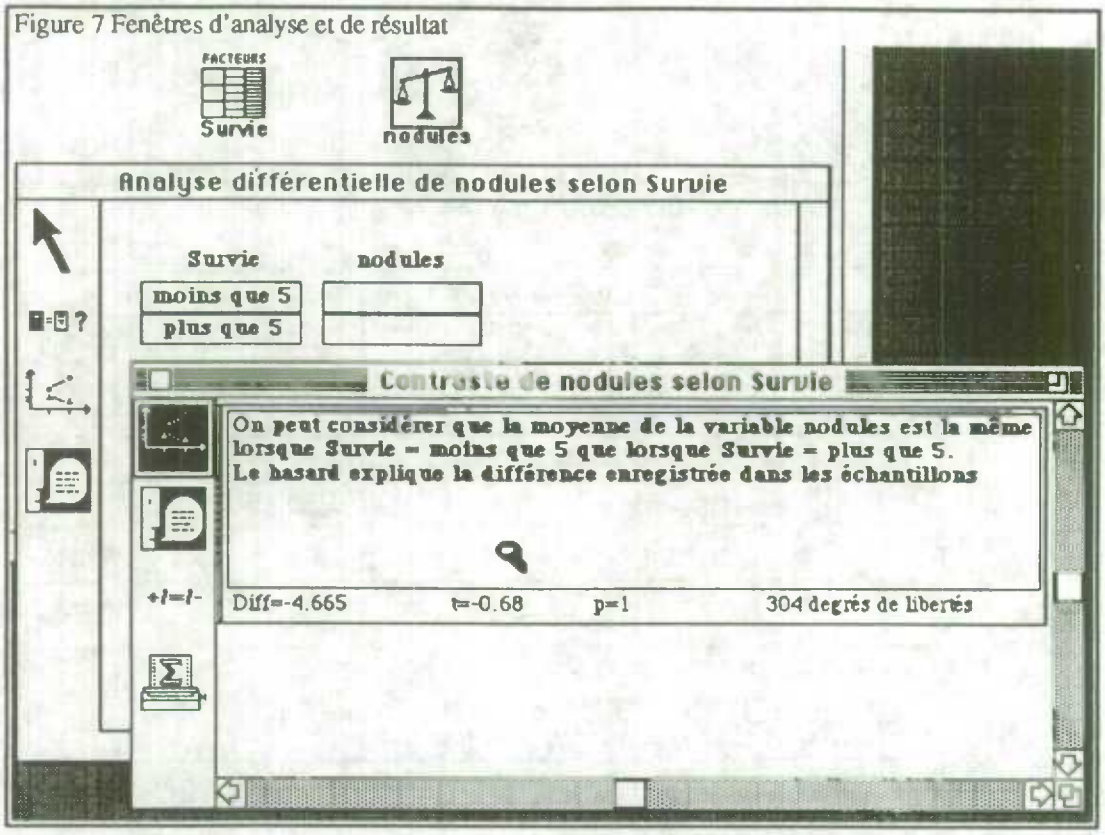
4.2 La méta-analyse

Une fois que la préparation des données est complétée, l'usager peut, s'il le désire sauvegarder son projet. STATEX permet à l'usager la gestion complète de ses sessions de travail par la gestion des projets. Un menu déroulant facilite cette opération. Il pourra par la suite sélectionner les variables et groupes qu'il désire analyser. Pour cela il cliquera sur les icônes correspondant avec l'outil 'Flèche qui pointe une loupe'. Ensuite il déclenchera la méta-analyse en cliquant sur la Loupe. Lors de la méta-analyse, STATEX procédera à classifier les données sélectionnées selon le dessin de la matrice, par exemple 'plusieurs-groupes-une-variable-capable-de-moyenne'. En fonction de cette classification, de l'étape en cours et du niveau de l'usager, deux types de recommandations seront proposées: des actions dites structurantes ou des stratégies d'analyse selon les analogies heuristiques décrites précédemment. L'usager pourra demander des explications qui le guideront sur les alternatives proposées. STATEX se chargera d'exécuter le choix. Les actions structurantes sont par exemple 'dégrader une variable quantitative' en vue d'une analyse corrélationnelle avec une autre variable de type classificatoire.

4.3 La stratégie d'analyse

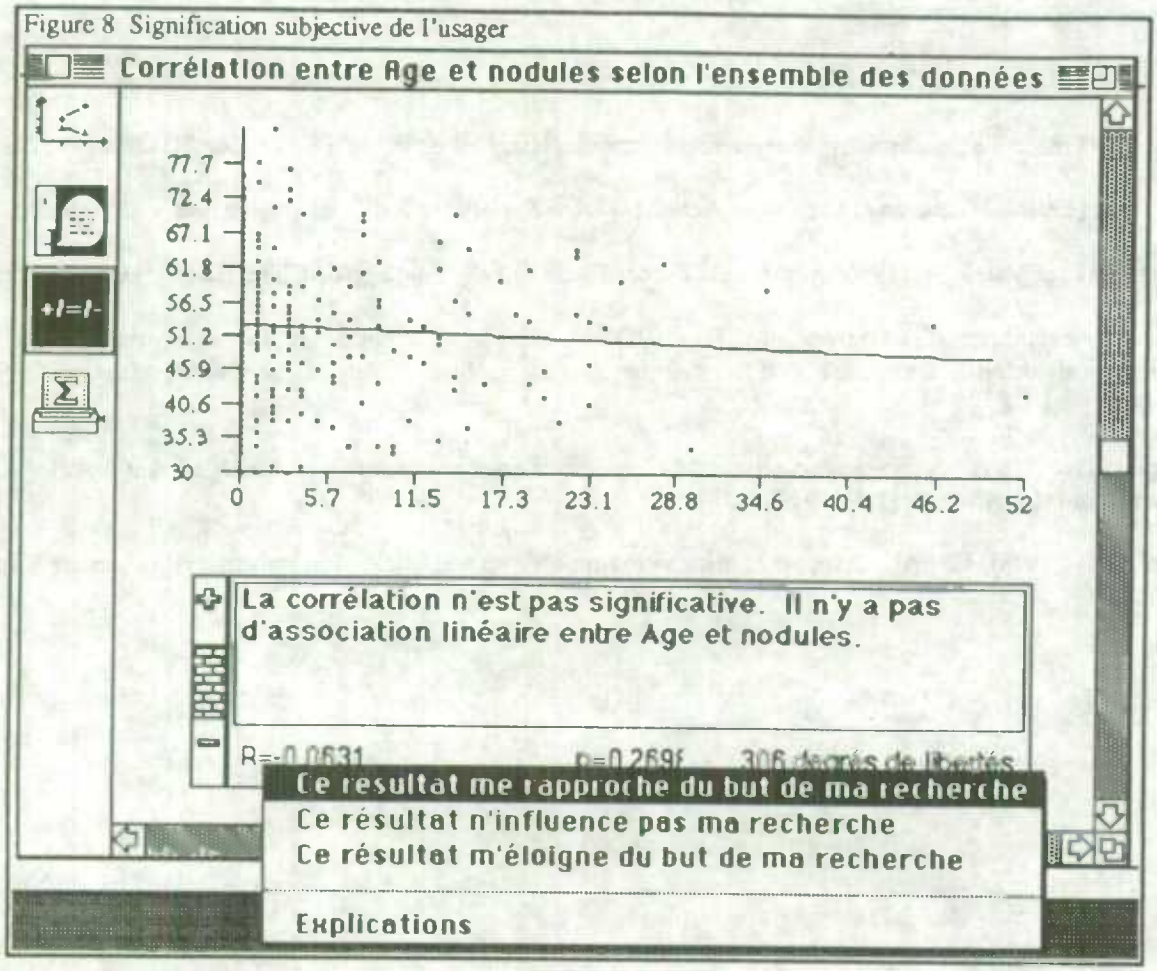
Si une stratégie d'analyse est proposée lors de la méta-analyse et si l'usager en fait le choix, STATEX déclenchera alors un processus qui, selon les caractéristiques des données sélectionnées, des connaissances statistiques de STATEX et des possibilités d'analyses disponibles, affichera une fenêtre d'analyse. Cette fenêtre d'analyse (Fig. 7) montrera à l'usager la forme de la matrice qui résulte de son choix de variables et facteurs et des icônes qui représentent des outils statistiques appropriés.

Figure 7 Fenêtres d'analyse et de résultat



Si l'usager en sélectionne un, le calcul statistique correspondant sera déclenché et affiché de façon graphique et textuelle. L'usager est alors invité à faire une introspection, et à dire si le résultat obtenu est significatif pour lui. De cette manière, STATEX pourra organiser les résultats et en tenir compte pour le rapport final, pour modifier le statut de l'usager et pour la suite de l'analyse (Fig. 8).

Figure 8 Signification subjective de l'usager



5. CONCLUSION

Un outil capable de contribuer efficacement à résoudre un problème statistique a été réalisé à un stade de prototype. Le prototype a été développé en Prolog et Flex de Quintus Mac Prolog et MPW Fortran pour CSPAL sur un Macintosh. CSPAL (Strobel 1978) a été modifié et modularisé pour être intégré à STATEX. Les routines statistiques implantées dans STATEX sont les statistiques de base, telles que le test t de Student , les corrélations, l'analyse de variance, etc. Cet outil de nature très interactive facilite le choix de stratégies d'analyse, le choix et le calcul de la statistique appropriée, la présentation des résultats sous forme graphique et l'interprétation des résultats dans le langage de l'usager. Cet un outil capable de satisfaire les besoins d'un usager qui désire analyser des données par une interface ergonomique et capable d'un support adéquat. Le système facilite la gestion des projets d'analyse. L'expérience accumulée dans ces analyses pourra être utilisée pour améliorer le système. Il est possible de visualiser la stratégie d'analyse d'un usager et de suivre son cheminement ainsi que de recueillir des statistiques sur l'utilisation du système. Le système facilite la production de rapports en permettant l'impression des fenêtres et le transport des affichages vers un logiciel de traitement de texte. Il est très utile pour l'exploration interactive de données et pour faciliter la découverte de la connaissance contenue dans les données. Le travail avec cet outil permet à l'usager d'augmenter ses connaissances dans le domaine de l'analyse de données.

BIBLIOGRAPHIE

Carey, S. (1990). Cognitive Development. (Éds. D. Osherson et E. E. Smith), *Thinking*, 147-172. Cambridge, Massachusetts: MIT Press.

Gale, W.A. (1986). Artificial Intelligence and Statistics. Reading, Massachusetts: Addison-Wesley.

Good, I.J. (1983). The philosophy of exploratory data analysis. *Philosophy of Science*, 50, 283-295.

Hand, D.J. (1991). Measurement scales as metadata. Dans Society for Artificial Intelligence and Statistics.

Maturana, H., et Varela, F. (1984). *El árbol del conocimiento* (4 Éd.). Santiago de Chile: Editorial Universitaria.

Muzard, J., Falardeau, E., et Strobel, M. G. (1991). A blackboard architecture for an expert assistant in statistical analysis. Dans EC2 (Ed.). *Avignon 91 Les systèmes experts et leurs applications*, 3, 79-93. Avignon: EC2.

Oldford, R.W. (1990). Sofware abstractions of elements of statistical strategy. *Annals of mathematics and artificial intelligence*, 2, (1-4), 291-308.

Strobel, M.G. (1978). CSPAL: Compact Statistics Programs for an Analytical Library. Suisse: Pirkhausser Basel.

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

UNE APPROCHE MULTIDIMENSIONNELLE DE LA LOCALISATION DES ENQUÊTÉS

L. Li, G. Deecker et P. Daoust[1]

RÉSUMÉ

La présente étude porte sur l'élaboration d'une approche multidimensionnelle des problèmes relatifs à la localisation des enquêtés et au couplage des données. On y précise comment les adresses, codes postaux, numéros de téléphone, noms de lieu et désignations officielles du territoire peuvent servir à localiser des enquêtés et, de ce fait, à relier les données pertinentes à la région géostatistique appropriée. Les auteurs décrivent une stratégie de regroupement de ces éléments spatiaux visant à améliorer la robustesse et la spécificité localisatrice, et ils brossent un tableau des règles décisionnelles servant à accroître l'insensibilité aux défaillances et à départager plusieurs solutions possibles. Enfin, ils se penchent sur l'utilité éventuelle de l'approche multidimensionnelle pour diverses opérations d'enquête et sur des occasions de recherche futures.

MOTS CLÉS: Géocodage; référence géographique; couplage de données; codage automatisé.

1. INTRODUCTION

Lorsqu'on mène une enquête ou un recensement, il faut définir où se trouve un enquêté donné si l'on veut bien coupler les données et en évaluer la qualité, rapprocher les résultats de plusieurs recensements ou enquêtes ou procéder à plusieurs autres tâches. La même règle s'applique lorsqu'il s'agit d'adapter des données administratives à des unités de déclaration géostatistiques dans le but de préparer un rapport statistique.

Traditionnellement, la vérification géographique et le couplage des données se sont avérés des tâches manuelles fastidieuses. Puis, vers la fin des années 1960, des outils tels que le fichier de conversion des codes postaux et le fichier principal de région de Statistique Canada, ainsi que les fichiers DIME et (aujourd'hui) TIGER du U.S. Bureau of the Census, ont permis d'automatiser une partie de ce travail. Jusqu'ici, on a privilégié les approches unidimensionnelles, le code postal servant le plus souvent de critère localisateur (Wilkin 1988a; Wilkin 1988b; Nadwodney 1989). Dernièrement, on retient plutôt des approches multidimensionnelles (Norris et Kirk 1989; Schneider 1987; Yergen 1987), car elles offrent la possibilité d'améliorer la résolution spatiale et la robustesse (c'est-à-dire l'insensibilité aux données manquantes ou erronées).

Le présent document fait le point sur la recherche effectuée à Statistique Canada en vue d'élaborer une approche multidimensionnelle des problèmes relatifs à la localisation des enquêtés et au couplage des données. Les auteurs cherchent à savoir comment divers éléments géographiques peuvent servir à localiser des enquêtés et à relier les données correspondantes à une unité géostatistique appropriée, puis comment on peut réunir ces éléments afin d'améliorer le processus de localisation des enquêtés. Ils passent en revue plusieurs approches et précisent les règles décisionnelles susceptibles de résoudre les difficultés attribuables à des réponses incorrectes. De plus, ils présentent des chiffres afin d'illustrer la résolution spatiale de certains procédés unidimensionnels et multidimensionnels. Enfin, ils se penchent sur l'utilité éventuelle de l'approche multidimensionnelle pour diverses opérations d'enquête et sur des occasions de recherche futures.

---

[1] L. Li et G. Deecker, Division de la géographie, Statistique Canada, édifice Jean-Talon, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.
P. Daoust, Division des méthodes d'enquêtes - entreprises, Statistique Canada, édifice R.H. Coats, Parc Tunney, Ottawa (Ontario), Canada K1A 0T6.

2. LES ÉLÉMENTS GÉOGRAPHIQUES DE LA LOCALISATION DES ENQUÊTÉS

2.1 Adresses

L'adresse de l'enquêté figure parmi les éléments géographiques les plus répandus dans les réponses aux enquêtes et les fichiers administratifs. En règle générale, les réponses correspondent à l'adresse postale de l'enquêté. Les caractéristiques des adresses postales sont analysées par Deguire (1988). Dans les zones urbaines, l'adresse postale dénote habituellement un logement, quoi qu'il existe des superboîtes, des cases postales, la poste restante et d'autres formes d'adresse qui ne sont pas spécifiquement localisatrices. Dans les zones rurales, par contre, les adresses postales sont moins informatives. Une adresse qui précise une route rurale («route rurale 3, Almonte (Ontario)», par exemple) donne une résolution spatiale assez faible. En outre, les habitants d'une région rurale sont plus susceptibles que les citadins de donner une adresse fondée sur une case postale ou la poste restante, qui ne fait qu'établir un lien entre l'enquêté et le bureau de poste.

En région urbaine, on peut conjuguer les adresses à un fichier de rues, comme le fichier principal de région de Statistique Canada ou les fichiers DIME et TIGER du U.S. Bureau of the Census, pour tâcher de localiser un enquêté (Yergen 1987, 221-230). Il faut d'abord faire correspondre l'adresse de l'enquêté à la fourchette d'adresses et au nom de rue rattachés à chaque segment de rue dans le fichier. Aujourd'hui, de nombreux progiciels commerciaux d'information géographique, par exemple MAPINFO (Mapping Information Systems Corporation 1989, 4-63) et ARC/INFO (ESRI 1989), prévoient des programmes spécifiques pour cette tâche.

En région rurale, il s'avère beaucoup plus difficile d'utiliser l'adresse postale pour localiser l'enquêté. Cela s'explique par la piètre spécificité/résolution spatiale des adresses rurales et par l'absence de fichiers des rues pour les zones rurales au Canada.

On peut aussi localiser un enquêté au moyen de son adresse postale en convertissant celle-ci en un code postal, puis en faisant appel à un outil tel que le fichier de conversion des codes postaux de Statistique Canada (Division de la géographie 1989) pour relier l'enquêté à un ou plusieurs secteurs de dénombrement. La conversion de l'adresse postale en code postal se fait habituellement en deux étapes. Premièrement, on analyse l'adresse brute afin d'en tirer une clé de recherche des adresses (Deguire 1988; Yergen 1987, 221-231). Deuxièmement, on utilise la clé pour extraire un code postal du fichier maître des adresses et des codes postaux de Postes Canada. C'est un peu comme si on se rendait au bureau de poste pour consulter les répertoires de codes postaux dans le but de trouver le code qui correspond à une adresse donnée.

2.2 Codes postaux

On retrouve le code postal dans la grande majorité des fichiers administratifs et des réponses aux enquêtes. Ce code alphanumérique à six caractères sert couramment de critère localisateur pour le traitement des enquêtes (Nadwodney 1989), l'analyse des données (Wilkin 1988a) et de nombreuses autres applications (Maloney 1988; Nadwodney 1989).

Comme on l'a vu plus haut, on peut conjuguer le code postal au fichier de conversion des codes postaux de Statistique Canada afin de rattacher un enquêté à un ou plusieurs secteurs de dénombrement. Le fichier de conversion sert ainsi de table à consulter qui facilite la conversion.

La résolution spatiale des codes postaux varie beaucoup, selon qu'on se trouve en zone urbaine ou rurale. Dans les villes, un code postal de six caractères peut correspondre à un îlot, à un grand immeuble d'appartements et même à un étage d'un immeuble de bureaux (Postes Canada 1983, 7). À la campagne, le code postal représente un secteur de service, qui peut englober des parties de plusieurs secteurs géostatistiques. Le tableau 1 montre la résolution comparative des codes postaux urbains et ruraux par province.

TABLEAU 1: NOMBRE MOYEN DE SD ET DE PARTIES DE SD PAR CODE POSTAL PAR PROVINCE

	Can.	C-B	Alb.	Sask.	Man.	Ont.	Qué.	N-B	N-É	Î-P-É	T-N
Urbains	1.04	1.04	1.03	1.03	1.04	1.04	1.03	1.04	1.03	1.02	1.05
Ruraux	4.31	5.13	6.16	4.92	3.90	4.31	2.99	4.29	3.92	5.18	2.14

2.3 Numéros de téléphone

Un numéro de téléphone se compose de trois éléments: un indicatif régional de trois chiffres, un indicatif de central de trois chiffres et un indicatif local de quatre chiffres. L'indicatif régional délimite une grande région. À titre d'exemple, le Manitoba tout entier est circonscrit dans l'indicatif régional 204. D'autre part, l'indicatif de central définit une région relativement restreinte, habituellement desservie par un seul centre de commutation téléphonique et correspondant souvent à la taille d'une municipalité (même si elle ne coïncide pas forcément avec le territoire municipal). Enfin, les quatre chiffres de l'indicatif local sont spécifiques à chaque abonné, exception faite de ceux qui sont desservis par une ligne partagée.

À l'heure actuelle, l'indicatif régional et l'indicatif de central sont les plus utiles pour localiser un enquêté, étant donné qu'ils délimitent un secteur de service particulier et que chaque indicatif de central se distingue au sein de son indicatif régional. L'indicatif local promet de faciliter un jour le couplage de l'enquêté avec un ménage, mais l'insuffisance des données et le coût prohibitif en restreignent pour l'instant l'utilité à l'échelle nationale.

Contrairement aux codes postaux, les indicatifs de central donnent une résolution spatiale meilleure à la campagne qu'en ville. Le tableau 2 montre le nombre moyen de SD par indicatif de central dans chaque province.

TABLEAU 2: NOMBRE MOYEN DE SD PAR INDICATIF DE CENTRAL ET PAR PROVINCE[2]

	Can.	C-B	Alb.	Sask.	Man.	Ont.	Qué.	N-B	N-É	Î-P-É	T-N.
Urbains	---------- Données non disponibles ----------										
Ruraux	5.84	4.18	7.96	7.88	---	7.13	5.29	5.22	5.13*		2.30

* Nouvelle-Écosse et Île-du-Prince-Édouard réunies.

2.4 Noms de lieu

On obtient souvent un nom de lieu désignant le lieu de résidence de l'enquêté dans les fichiers administratifs ou parmi les réponses aux enquêtes. Lorsqu'il est impossible de retracer un tel lieu dans le cadre d'une recherche fondée sur l'adresse ou le code postal, on peut tenter d'utiliser directement le nom de lieu pour localiser l'enquêté. Cette stratégie passe par l'élaboration d'un fichier de référence des noms de lieu susceptible de fournir les coordonnées XY ou l'unité géostatistique appropriée pour chaque lieu, et par une stratégie appropriée d'appariement des textes et d'analyse syntaxique.

---

[2] Il faudrait considérer ces moyennes comme des estimations approximatives, car certains secteurs de dénombrement englobaient plus d'un indicatif de central et ont de ce fait été comptés plus d'une fois. Aussi, les renseignements relatifs au téléphone ont été puisés à même le recensement de l'agriculture de 1986, qui ne tient pas compte de nombreux enquêtés dans les secteurs de dénombrement urbains. Ainsi ce recensement ne propose-t-il pas une couverture exhaustive de tous les indicatifs de central en zone urbaine.

Statistique Canada constitue actuellement un fichier de référence des noms de lieu, comportant à peu près 160,000 entrées, à partir des données des recensements antérieurs, des noms associés à des régions géostatistiques standard, de localités non constituées et d'autres sources pertinentes (Norris et Kirk 1989, 12-13). Il existe une source additionnelle qu'il faudrait approfondir: le fichier toponymique du Service de géographie (Énergie, Mines et Ressources Canada), qui incorpore les noms de lieux officiels des gazettes provinciales.

Les problèmes reliés à l'appariement des noms de lieu ressemblent à ceux de l'extraction automatisée de textes, du codage automatisé et du couplage des enregistrements. La riche documentation sur l'extraction automatisée de textes (Ashford et Willet 1988; Saffady 1989; Stanfill et Kahle 1986; Bouchard 1979) fait ressortir des approches et méthodes utiles: compromis entre le rappel et la précision, modalités d'incorporation de l'insensibilité aux défaillances dans les recherches, etc. Par ailleurs, Sellis (1988), Wu et Burkhard (1987), Ramamohanarao et Lloyd (1983) et Bouchard (1979) se penchent sur les moyens à employer pour optimiser la recherche.

Pourtant, l'appariement des noms de lieu se distingue nettement de l'extraction d'un texte à partir de documents autonomes. En effet, l'unicité d'un nom de lieu dépend de son contexte (il arrive qu'un nom soit unique sur son propre territoire seulement, par exemple lorsque deux villes du même nom sont situées à proximité l'une de l'autre) et les noms d'emprunt sont communs. Au Canada, la situation se complique davantage du fait qu'il existe des noms français, anglais et autochtones.

Les travaux de Norris et Kirk (1989) et du U.S. Bureau of the Census (Schneider 1987) donnent une bonne idée du codage automatisé des noms de lieu. Norris et Kirk, faisant appel au système de codage automatisé par reconnaissance de texte (Division du développement 1989), qui incorpore un algorithme à base d'entropie pour déterminer l'appariement, ont obtenu des taux de réussite d'environ 80% à partir d'échantillons nationaux. Pour sa part, Schneider (1987, B-3 à B-6), dans le cadre de la recherche préparatoire au recensement américain de 1990, a signalé des taux de réussite variables, allant d'à peu près 87% pour Los Angeles à 44% pour le Mississippi.

Les recherches en cours s'inspirent - pour les compléter - des approches générales retenues par Norris et Kirk (1989), Schneider (1987) et Yergen (1987). La démarche se résume comme suit: on soumet d'abord les noms à une analyse syntaxique, puis on cherche un appariement à partir des mots significatifs au sein d'une même chaîne de noms. On recherche ensuite un appariement direct pour chacune des variables d'entrée. Lorsqu'on n'obtient pas d'appariement direct, on retient un appariement approximatif. On définit alors une série d'intersections au moyen d'une comparaison croisée des candidats retenus pour chaque variable. S'il reste plusieurs candidats dans le fichier d'intersections, on peut appliquer une série de règles pour éliminer ceux qui ne conviennent pas en se basant sur les dissemblances entre les caractéristiques des lieux appariés. On fait également appel, le cas échéant, à des critères d'élimination selon la distance pour déterminer dans quelle mesure il est vraisemblable qu'un lieu donné soit celui de l'enquêté.

Lorsqu'il s'agit de localiser un enquêté au moyen d'un nom de lieu, la résolution spatiale varie considérablement selon la nature de la question à laquelle répond l'enquêté. S'agissant des applications nationales (une étude sur les migrations, par exemple), le rattachement à une municipalité semble une résolution souhaitable (Norris et Kirk 1989). Pour les fins des études sur la qualité des données du recensement et de la vérification des données, on peut préférer des lieux plus précis, comme un secteur de dénombrement particulier.

En 1986, le Canada comptait 6,009 municipalités dûment constituées. De ce nombre, environ 41% se trouvent dans un seul secteur de dénombrement, tandis que 17% englobent deux SD, 9%, trois SD et 33%, quatre SD ou plus. Ces chiffres expliquent bien pourquoi les noms des municipalités, qui correspondent souvent à ceux qui identifient les bureaux de poste, représentent un critère utile pour le couplage des données géographiques.

Les localités non constituées, qui sont habituellement des entités inframunicipales, sont relevés dans le cadre des opérations sur le terrain du recensement. Ils constituent une source de renseignements additionnelle servant à localiser un enquêté aux fins du couplage des données.

Dans certains de nos tests, les noms des municipalités et des localités non constituées ont permis d'intégrer des enquêtés à un même SD par voie de mappage déterministique, avec des taux de réussite allant jusqu'à 70% dans

les régions rurales, lorsque les données d'entrée étaient raisonnablement libres d'erreur. En région urbaine, par contre, où chaque municipalité tend à englober plusieurs SD, il est rare de rencontrer un appariement déterministique au niveau des SD.

À la lumière de nos recherches et des expériences de Norris et Kirk (1989) et de Schneider (1987, B-7a - B-12), les facteurs suivants expliquent une bonne partie des appariements ratés et des cas de non-appariement: les abréviations, les erreurs d'orthographe, les noms d'emprunt, les entrées dans un champ incorrect, les noms incomplets, les noms historiques qui ont été modifiés et les noms ambigus ou non exclusifs.

2.5 Désignation officielle d'un territoire

La nature des désignations officielles des unités territoriales varie considérablement d'une province à une autre, au sein d'une même province et même d'une région urbaine à une zone rurale. Cette hétérogénéité traduit la diversité historique du peuplement canadien: le jeu des influences française et anglaise à l'Est, le peuplement plus ordonné des Prairies et la primauté d'un environnement implacable en Colombie-Britannique. Aussi est-il très difficile de formuler une question généralisée pour obtenir une désignation officielle du territoire et de mettre au point une structure standard pour saisir les réponses des enquêtés dans le cadre d'une enquête nationale ou d'un recensement.

Par exemple, les données sur les cantons, rangs et sections qui proviennent des Prairies et du district de Peace River, en Colombie-Britannique, constituent des descripteurs numériques comportant parfois des modificateurs alphabétiques. Dans les autres provinces, les descripteurs sont alphanumériques. D'autre part, les réponses provenant du Québec, des Maritimes et de la Colombie-Britannique sont extrêmement variables. On peut d'ailleurs s'attendre à tout: réponses qui ne respectent pas la suite et la structure prévues des variables, inclusion d'adresses ordinaires, numéros d'identification de propriété, noms de ville ou de seigneurie, numéros de lot tirés d'un plan de lotissement, etc.

Malgré l'hétérogénéité des données, les désignations officielles du territoire constituent un puissant moyen de localiser les enquêtés à la campagne, car il s'agit de descripteurs extrêmement spécifiques qui sont bien connus dans bon nombre de milieux ruraux. Au niveau le plus finement détaillé, on peut identifier une parcelle appartenant à un particulier; le fichier de référence qu'il faudrait alors constituer pour l'ensemble du Canada nécessiterait un énorme volume de données et il serait très onéreux à compiler. Si l'on opte pour un niveau plus élevé d'agrégation spatiale, on constate que le lot de concession d'un canton (canton-rang-section) représente une échelle plus pratique sur le plan national. Un lot de concession et son équivalent dans l'Est canadien mesurent habituellement entre 100 et 200 acres. Dans les Prairies, une section s'étend sur à peu près 640 acres ou un mille carré.

Afin d'adapter les désignations officielles du territoire à la localisation des enquêtés, on a mis deux stratégies à l'essai. Pour les Prairies et le district de Peace River en Colombie-Britannique, où les réponses sont habituellement conformes au modèle de données prévu, l'appariement direct en fonction du canton-rang-section intégral et du méridien a souvent porté fruit. Lorsqu'un appariement direct s'avérait impossible, on a tenté un appariement partiel, sans tenir compte de la section ou du rang.

Dans certaines autres provinces, notamment au Québec et en Colombie-Britannique, dont les données sont beaucoup plus variables, la position de la réponse ne peut pas être considérée comme un facteur significatif quand il s'agit d'évaluer l'appariement. Par exemple, lorsque l'enquêté indique «Oxford» à la variable «Comté», on ne privilégierait pas l'appariement «comté d'Oxford» par rapport à celui de «canton d'Oxford». En effet, ces deux candidats «concurrents» seraient considérés à égalité et il faudrait appliquer d'autres règles pour les départager, incorporant des critères d'élimination selon la distance et des caractéristiques des lieux en cause, afin d'exclure de la course les candidats superflus.

3. STRATÉGIE MULTIDIMENSIONNELLE

La discussion ci-dessus expose cinq approches unidimensionnelles du problème de la localisation de l'enquêté. Une stratégie réunissant la totalité ou une partie de ces approches promet d'ajouter de la robustesse au processus et d'améliorer la spécificité spatiale des diverses méthodes en profitant des forces de l'une pour compenser les faiblesses d'une autre.

Plusieurs chercheurs ont mis à l'épreuve des stratégies multidimensionnelles pour la localisation d'enquêtés ou des problèmes semblables (Norris et Kirk 1989; Schneider 1987; Yergen 1987; Drew, Armstrong et Dibbs 1987). Les recherches en cours élargissent les méthodes générales employées par ces auteurs pour englober le recours aux indicatifs de central téléphonique et les désignations officielles du territoire. On fait intervenir l'insensibilité aux erreurs dans le processus d'identification des candidats individuels en incorporant des règles destinées à vérifier et à raffiner les indicatifs régionaux et les noms de province qui manquent ou qui comportent une erreur. En outre, à titre d'essai en marge du recensement de l'agriculture, on utilise une variable distincte comme indicateur de probabilité pour le départage final des candidats lorsqu'il s'agit de localiser une exploitation agricole ou un agriculteur.

La première étape du processus consiste à raffiner les données d'entrée afin de garantir qu'on dispose des éléments essentiels voulus pour maximiser l'utilité ou la pertinence des données aux fins des opérations subséquentes. Il faut d'abord retenir l'adresse postale de l'enquêté, son code postal, son numéro de téléphone et la désignation officielle du territoire où il vit. On raffine ces données en corrigeant les données provinciales manquantes, étape critique du traitement de l'adresse; ainsi, le code postal et l'indicatif régional servent à extraire le nom de la province d'un fichier de référence. L'indicatif régional est vérifié à partir d'une liste de tous les indicatifs connus, tels qu'ils ont été réunis par les compagnies de téléphone. On corrige un indicatif régional inexact (ou on établit un indicatif manquant) en utilisant le nom du bureau de poste et de la province pour extraire l'indicatif régional de l'enquêté à partir d'un fichier de référence. On procède à une analyse syntaxique de l'adresse et on extrait le code postal approprié au moyen du programme PCODE (Sous-division de la recherche et des systèmes généraux 1987). Les désignations officielles du territoire déclarées à part font aussi l'objet d'une analyse syntaxique en vue d'en éliminer le bruit et de faciliter l'identification des composantes les plus significatives aux fins d'un appariement ultérieur.

La deuxième étape consiste à extraire les lieux ou SD où on peut retrouver l'enquêté au moyen de chacun des éléments localisateurs: code postal, adresse, indicatif de central et désignation officielle du territoire. Il suffit de procéder à une série de consultations directes de fichiers où les codes postaux, les indicatifs de central et les comtés-cantons-concessions-lots sont recoupés avec les SD (dans le premier cas, il s'agit du fichier de conversion des codes postaux).

Lorsqu'on a récupéré jusqu'à quatre séries de lieux/SD, on procède à une contre-vérification afin d'identifier les candidats qui se retrouvent dans le nombre le plus élevé de séries de données d'entrée. La série d'intersections qui en découle définit les lieux où devrait se trouver l'enquêté. Cette contre-vérification sert à éliminer les données erronées, car elle exige que plusieurs critères géographiques indépendants pointent vers les mêmes lieux avant d'être inclus dans la série d'intersections.

Le tableau 3 illustre la résolution spatiale d'un tel procédé multidimensionnel.

TABLEAU 3: NOMBRE MOYEN DE SD ASSOCIÉS À UNE COMBINAISON UNIQUE DE CODES POSTAUX ET D'INDICATIFS DE CENTRAL TÉLÉPHONIQUE PAR PROVINCE[2]

	Can.	C-B	Alb.	Sask.	Man.	Ont.	Qué.	N-B	N-É	Î-P-É	T-N
Urbains	---------- Données non disponibles ----------										
Ruraux	2.65	2.80	2.82	2.81	2.07	2.61	1.99	2.49	2.28	2.93	1.50

Nota: L'estimation est fondée sur les données du recensement de l'agriculture de 1986, qui sous-estime les régions rurales non agricoles.

Lorsqu'il s'agit de localiser un enquêté, on a souvent besoin d'un lieu définitif. Pour les études sur les migrations, la résolution voulue peut correspondre à une municipalité. Pour une étude sur la qualité des données, on peut préférer l'exactitude d'un SD; d'autre part, l'établissement d'un ordre prioritaire des SD où risque de se trouver l'enquêté est utile pour réduire l'étendue de la recherche, qu'il s'agisse d'une recherche manuelle approfondie ou d'un appariement automatisé de certaines caractéristiques de l'enquêté telles que le sexe et la date de naissance. Par conséquent, lorsqu'on identifie de multiples candidats dans la série d'intersections, on a besoin d'un mécanisme ou d'un procédé pour déterminer l'éventualité de retrouver l'enquêté dans chacun des lieux retenus. Le procédé adopté fait intervenir une variable associée à titre d'indicateur de la probabilité de retrouver l'enquêté dans un lieu donné. Par exemple, lorsqu'il s'agit de localiser un agriculteur donné, le nombre d'exploitations agricoles dans chacun des SD candidats peut servir d'indicateur de la probabilité qu'une recherche plus rigoureuse s'impose pour un SD donné.

À l'avenir, on devrait améliorer le rendement du procédé en élargissant la série de règles dans le but de tenir compte d'autres caractéristiques clés de l'enquêté et, de ce fait, de mieux définir les lieux candidats qui partagent les mêmes caractéristiques précises. Dans le cas du recensement de l'agriculture, cette série de règles élargies pourrait englober les cultures d'un agriculteur donné ou d'autres caractéristiques telles que la taille de l'exploitation agricole.

4. L'APPRENTISSAGE-MACHINE POUR AMÉLIORER LE RENDEMENT DU SYSTÈME

Il est coûteux et fastidieux de compiler un fichier de référence des unités officielles de territoire (canton-concession-lot) pour le Canada. Dans le cadre de nos recherches, nous mettons au point un procédé qui s'inspire de la faculté qu'a le personnel affecté à la vérification manuelle d'acquérir des connaissances au travail.

Ce procédé assimile, à partir d'enregistrements sans erreurs, des désignations d'unités territoriales inconnues jusqu'ici, ainsi que les SD auxquels elles se rapportent; il établit ensuite un rapprochement entre l'information existante et les parties inconnues des descripteurs géographiques et tient un cumul du nombre de fois qu'est signalé chaque rapprochement. Lorsque le cumul dépasse une valeur plancher, qu'on est en train de définir empiriquement à partir de données des recensements précédents, le rapprochement en question est versé dans le fichier de recoupement du SD avec l'unité territoriale afin d'améliorer la couverture.

À la lumière d'un certain nombre d'essais, ce procédé convient aux régions où les données comportent relativement peu d'erreurs et où chaque unité territoriale pour laquelle on cherche à préciser le rapport spatial comporte un nombre d'enquêtés assez élevé. Dans les régions où les désignations territoriales sont très variables et où les réponses fluctuent beaucoup, comme le Québec, on trouve rarement le nombre de réponses répétitives voulu pour faire passer un rapport spatial donné à la «base des connaissances». La poursuite des recherches en vue d'améliorer la connaissance des caractéristiques des réponses et, de ce fait, la mise au point de stratégies améliorées d'analyse syntaxique et de substitution d'expressions, amélioreraient vraisemblablement le rendement de ce processus d'«apprentissage».

5. CONCLUSIONS

Dans bien des secteurs d'enquête, de collecte et d'analyse, il faut localiser les sujets et relier les données à un ensemble de régions géostatistiques. Les recherches en cours ont montré qu'une approche multidimensionnelle peut améliorer la résolution spatiale des vérifications localisatrices et la robustesse du procédé. Cette approche, en voie d'élaboration, s'inspire des stratégies et méthodes utilisées dans l'extraction automatisée de textes, le couplage des enregistrements et le travail d'autres spécialistes dans le domaine du géocodage multidimensionnel. Certaines applications ont déjà été mises en oeuvre aux fins de la vérification géographique du recensement de l'agriculture et à celles des études qualitatives menées dans le cadre du recensement de la population et du logement.

À l'heure actuelle, les systèmes mis en oeuvre servent toujours, de sorte qu'on ne dispose pas encore de résultats définitifs quant à leur rendement. D'après les indications provisoires, celui-ci se rapproche beaucoup des prévisions. On prévoit d'analyser les résultats dans un avenir rapproché.

Selon l'expérience à ce jour, un certain nombre d'occasions de recherche futures se sont manifestées. Ainsi, il faudra faire d'autres recherches sur l'établissement des limites des codes postaux dans les régions rurales. La pénurie de connaissances détaillées des noms de lieu et des désignations territoriales à l'échelle du pays constitue aussi un obstacle au rendement du système. Par ailleurs, il serait utile de consulter le personnel sur le terrain afin d'étudier d'autres sources éventuelles de cartes et de renseignements géographiques et d'intégrer ces renseignements au processus automatisé. Si nous voulons élargir les règles d'élimination des candidats superflus, nous devons procéder à une recherche précise susceptible de mieux nous renseigner sur les rapports entre les enquêtés et les caractéristiques de divers lieux géographiques, puis mettre au point, pour diverses interactions, des critères appropriés d'élimination des enquêtés selon la distance. L'adoption de certains concepts et approches inhérents à l'imputation par enregistrement donneur promet d'ailleurs d'être fort utile. Enfin, des recherches sur les capacités de l'intelligence artificielle pourraient ouvrir la voie à l'élaboration et à l'amélioration de règles en temps réel et à la bonification globale des procédés.

REMERCIEMENTS

La plus grande partie de cette recherche a été menée en vertu du programme interne de congé sabbatique de Statistique Canada. Le recensement de l'agriculture a fourni des données et contribué au financement. Nous remercions sincèrement Catherine Cromey, Marcelle Dion, Dave Dolson, Doug Drew, Rennie Molnar, Mary-Jane Norris et Peter Schut de leurs précieux conseils. Nous ne pouvons pas passer sous silence le travail acharné de l'équipe de recherche et d'élaboration des systèmes, dirigée par Danny Wall et Joe Gomboc. Les auteurs sont profondément reconnaissants du soutien et de l'appui manifestés par chacun de ses membres.

BIBLIOGRAPHIE

Ashford, J., et Willet, P. (1988). Text Retrieval and Document Databases, Chartwell-Bratt.

Bouchard, D., (1979). Combined top-down and bottom-up algorithms for using context in text recognition. Unpublished MSc. thesis, McGill Univ., Dept. of Computer Science, Montreal.

Canada Post (1983). *Postal Code Manual*. Canada Post Corporation, Operational Services Directorate, Mail Collection and Delivery Branch.

Deguire, Y. (1988). Postal address analysis. *Techniques d'enquête*, 14, 2, 335-343.

Development Division (1989). *ACTR - Automated Coding and Text Recognition System Manual*. Statistics Canada, Methodology and Informatics Branch, Development Division, General Systems Subdivision, Ottawa.

Drew, J.D., Armstrong, J.B. et Dibbs, R., (1987). Research into a register of residential addresses for urban areas of Canada, *Proceedings of the Annual Meetings of the American Statistical Association, Section on Survey Research Methods*, San Francisco, California, Aug. 17-20, 1987, 300-305.

Dueker, K.J., (1974). Urban geocoding. *Annals of the Association of American Geographers*, 64, 2.

ESRI (1989). *ARC/Info network manual*. Environment Systems Research Institute, Redlands, California.

Geography Division (1989a). *Detailed user guide, Postal Code Conversion File*, January 1989 Version. Statistics Canada, Ottawa.

Geography Division (1989b). *AMF user's guide*. Statistics Canada, Ottawa.

Giles, P. (1988). A model for generalized edit and imputation of survey data, *The Canadian Journal of Statistics*, 16, supplement, 57-73.

Hart, S.A. (1983). The development and use of postcodes for population information systems, in Jones H. (ed.) *Population change in contemporary Scotland*. Geo Books, Norwich, England. ISBN-0-86094-153-1.

Mapping Information Systems Corporation (1989). *MapInfo user's guide*. Mapping Information Systems Corporation, Troy, New York.

Nadwodney, R. (1989). *The canadian postal code system and postal code applications*. Geography Division, Statistics Canada, Ottawa.

Norris, M.J., et Kirk, J. (1989). Research and testing of an automated coding system for the mobility status variable using the ACTR System, analysis report. Internal report, Statistics Canada, Demography Division, 1991 Census Automated Coding Research Task. Ottawa.

Ramamohanarao, K., Loyld, J.W., et Thom, J.A. (1983). Partial-match retrieval using hashing and descriptors. *ACM Transactions on Database Systems*, 8, 4, 552-576.

Research and General Systems Subdivision (1987). *PCODE (Automated Postal Coding System) user guide*. Statistics Canada, Ottawa.

Saffady, W. (1989). *Text storage and retrieval systems*, Meckler Corp., London.

Schneider, P.J. (22 juillet, 1987). Memo to Robert W. Marx on "1986 Test Census: Analysis of Migration Coding and place-of-Work Workplace File Sources". US Bureau of the Census, Population Division.

Sellis, T.K. (1988). Multiple-query optimization, *ACM Transactions on Database Systems*, 13, 1, 23-52.

Stanfil, C., et Kahle, B. (1986). Parallel free-text search on the connection machine system, *Computing Practices*, 29, 12, 1229-1239.

Wilkins, R. (1988a). Using postal codes for analysis of socio-economic inequities in health outcomes, paper presented at the 14th Annual Health Administration Forum, University of Ottawa, Ottawa, August 1988.

Wilkins, R., et MacDonald, R. (1988). *Potential uses of postal codes with vital statistics data*. Health Division, Statistics Canada, Ottawa.

Wu, C.T., et Burkhard, W.A. (1987). Associative searching in multiple storage units, *ACM Transactions on Database Systems*, 12, 1, 38-64.

Yergen, W. (1987). 1990 test geocoding experience, *Building on the past-shaping the future, proceedings of URISA 25th Annual Conference*, Vol. II.

Dept. of the Environment (1987). *Handling Geographic Information: Report of the Committee of Enquiry chaired by Lord Chorley*, Her Majesty's Stationary Office, London.

SESSION 5

Innovations géographiques au niveau de la collecte des données

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

TIGER ET SES APPLICATIONS POUR LE RECENSEMENT DE 1990: AVANTAGES SUR LE PLAN DE L'ANALYSE DE DONNÉES ET APPLICATIONS ÉVENTUELLES POUR LES ENQUÊTES

R.W. Marx[1]

RÉSUMÉ

Le système TIGER n'évoque pas la même chose pour tous. Au cours des quatre dernières années, ce système a exécuté les fonctions de soutien géographique pour lesquelles il avait été conçu par la Division de la géographie du U.S. Bureau of the Census en vue du recensement décennal de 1990. Il est aussi à l'origine de la plupart des produits géographiques et cartographiques qui sont nécessaires pour totaliser les données recueillies et les rendre accessibles aux nombreuses organisations présentes dans les divers domaines d'activité économique. Parallèlement, nous sommes à tracer les plans d'avenir de ce nouveau système révolutionnaire -- et cet avenir est prometteur. Des travaux semblables sont en cours dans des organismes et des institutions aux États-Unis et dans le monde entier. Ces travaux portent plus spécialement sur les applications du système d'information géographique (SIG) qui mettent en évidence les produits ordinolingues du système TIGER et les produits de données du recensement décennal de 1990.

MOTS CLÉS: Cartographie automatique; recensement décennal; avenir; systèmes d'information géographique; système TIGER; Census Bureau des É.-U.

1. INTRODUCTION

Le système TIGER n'évoque pas la même chose pour tous. C'est tout d'abord une innovation (Carbaugh et Marx, 1990). La plupart des articles qui ont été écrits jusqu'à maintenant sur le système TIGER portent surtout sur la conception du système et son utilisation au sein du United States Bureau of the Census (Marx et coll., 1990); on y décrit une série d'opérations massives de production géographique et cartographique exécutées en même temps que le recensement décennal de 1990. Dans la présente communication, nous résumons les points saillants de cette évolution et essayons d'imaginer quel devrait être l'avenir du système TIGER au sein du Census Bureau et comme outil pour les géographes, les cartographes et les autres spécialistes qui sont à mettre au point des systèmes de géographie et de cartographie automatiques. Nous examinons aussi la possibilité d'utiliser ce nouveau système, de même que les données pertinentes du recensement de 1990, pour effectuer des analyses spatiales qui mettent à contribution la technologie du système d'information géographique (SIG).

2. QU'EST-CE QUE LE SYSTÈME TIGER?

La Division de la géographie, qui est une composante du Census Bureau, a élaboré le système TIGER à la suite d'une décision importante qui a été prise par l'organisme statistique en 1981 et selon laquelle il fallait créer un système qui aurait pour but

[1] R.W. Marx, Geography Division, Bureau of the Census, Département du commerce des É.-U., Washington (DC) 20233, É.-U.

«[...] d'automatiser toutes les opérations de soutien géographique et cartographique dans les délais voulus pour faciliter la collecte, la totalisation et la diffusion des données du recensement décennal de 1990, année du bicentenaire des recensements aux États-Unis.» (TRADUCTION)

La Division de la géographie disposait donc de six petites années pour construire une base de données informatique qui contiendrait le nom de chaque rue et de chaque route connue aux États-Unis (tâche qui a pu être accomplie uniquement grâce à la précieuse collaboration du United States Geological Survey -- USGS) de même que les tranches d'adresses qui correspondent à chaque segment de rue pour chacune des rues des 345 plus grands centres urbains des États-Unis; six petites années pour inclure aussi dans cette base toutes les voies ferrées des États-Unis plus tous les plans d'eau d'importance; enfin, six petites années pour saisir et vérifier le nom, le code numérique et les limites de toutes les unités géographiques dont s'est servi le Census Bureau pour totaliser les résultats des recensements décennaux de 1980 et de 1990. Bon nombre de ces unités servent aussi à la totalisation des résultats du recensement de l'agriculture et du recensement sur l'économie (voir tableau 1).

L'étape de l'élaboration est terminée; l'objectif global a été atteint...dans les délais prescrits! La base de données TIGER est-elle parfaite? Bien sûr que non! Mais la majeure partie de l'information qu'elle contient est exacte et cette information est beaucoup plus à jour que celle qui figure sur les cartes classiques ou celle contenue dans les fichiers GBF/DIME dont s'est servi le Census Bureau lors du recensement de 1980. Et le plus important, c'est que toute cette information est stockée désormais dans l'ordinateur! Le Census Bureau peut donc modifier facilement cette information au fur et à mesure qu'on lui signale de nouvelles données.

3. DE QUELLE MANIÈRE TIGER SERT-IL AU CENSUS BUREAU?

Le succès d'un recensement ou d'une enquête dépend non seulement de la collecte de données mais aussi de la capacité d'associer ces données à des régions géographiques. Dans ces circonstances, le Census Bureau a utilisé les produits et services du système TIGER pour établir, valider et géocoder la liste d'adresses des unités de logement et des habitations collectives pour le recensement de 1990, exécuter et évaluer les opérations de collecte des données de ce recensement, totaliser les données recueillies et les rendre accessibles à la nombreuse clientèle qui, par obligation légale ou par choix, s'intéresse aux résultats des recensements décennaux afin d'exécuter les nombreuses tâches qui façonnent notre quotidien (Tomasi, 1990), et offrir une vaste gamme de produits "à grande diffusion", sur support papier (voir tableau 2) et support informatique (voir tableau 3), qui sont directement en rapport avec les unités géographiques qui ont servi à produire les données du recensement de 1990 (LaMacchia, Tomasi et Piepenburg, 1987). Ces produits du système TIGER sont compatibles avec l'usage que font du SIG les individus et les organismes qui utiliseront les produits de données du recensement de 1990 à mesure que ceux-ci deviendront disponibles auprès du Census Bureau dans les deux prochaines années.

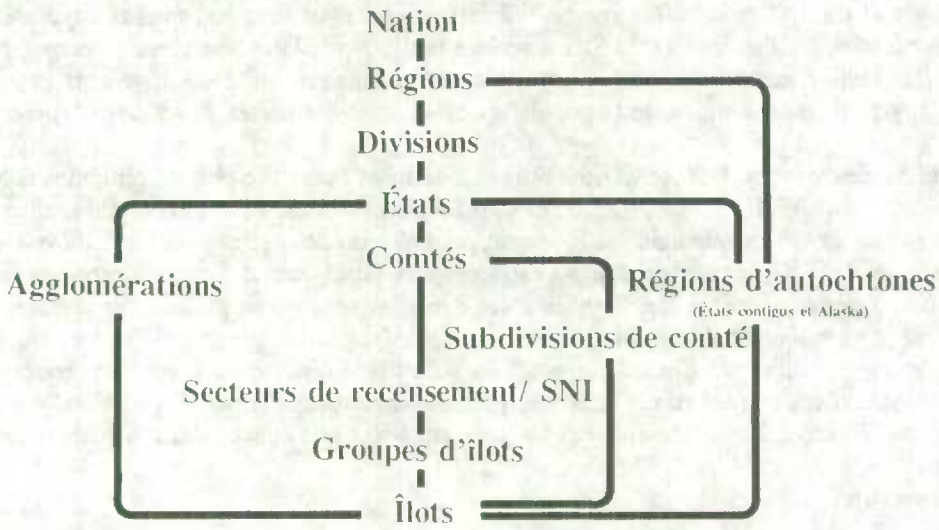
4. QUEL NIVEAU DE DÉTAIL GÉOGRAPHIQUE OBTIENT-ON AVEC TIGER?

Le Census Bureau utilise une hiérarchie d'unités géographiques pour la plupart de ses programmes statistiques (voir figure 1 et tableau 1). De façon générale, les unités de cette hiérarchie "s'emboîtent" les unes dans les autres, c'est-à-dire que chaque grande unité renferme généralement plusieurs petites unités.

4.1 Unités "de niveau supérieur"

Dans le haut de la hiérarchie se trouvent les États-Unis, c'est-à-dire l'ensemble des 50 États plus le District de Columbia. Les États-Unis sont divisés en 4 régions (qui sont des groupes d'États), puis en 9 divisions (qui sont aussi des groupes d'États). Ensuite viennent les 50 États et Washington (DC). Le Census Bureau a conçu la base de données TIGER de manière qu'elle comprenne aussi Porto Rico et les autres régions périphériques où se font les recensements de la population et du logement sous les auspices du Census Bureau: Samoa américaines, Guam, Mariannes du Nord, Palau et les îles Vierges américaines.

Figure 1 : Hiérarchie des unités géographiques du Census Bureau



Une fois divisés les États et ces unités statistiquement équivalentes, on obtient plus de 3 200 unités que l'on appelle généralement "comtés". La plupart de ces unités correspondent à un territoire assez vaste, ce qui représente un niveau de "résolution géographique" assez faible pour l'analyse des caractéristiques des individus, des unités de logement, des exploitations agricoles et des entreprises aux États-Unis; ce niveau est utile surtout pour les analyses à l'échelle nationale. Les comtés sont des unités très importantes dans l'analyse de données car le Census Bureau produit des chiffres par comté à partir des résultats du recensement de l'agriculture et du recensement sur l'économie, qui ont lieu à tous les cinq ans; ces recensements portent sur des sujets aussi variés que les entreprises manufacturières, le commerce de détail, les industries de services, le commerce de gros, les industries de la construction, les industries minérales, les transports, les statistiques d'entreprises, les entreprises appartenant à des membres de minorités ou à des femmes et tous les aspects de la vie agricole. Beaucoup d'autres organismes publics, de même que des organismes privés, recueillent et publient des données pour les comtés, ce qui concourt à produire une "riche" base de données sur ces unités.

Les quelque 3 200 comtés se subdivisent à leur tour en plus de 60 000 unités d'administration locale et unités statistiquement équivalentes -- townships, villes, villages, agglomérations de recensement, et ainsi de suite. Le rapport entre les agglomérations de recensement et le reste de la hiérarchie n'est pas toujours clairement défini car les agglomérations ne sont pas des unités géographiques "complètes en soi" comme peuvent l'être les comtés. Le mode de division mentionné ci-dessus implique vingt fois plus de cases de données que le mode fondé sur les comtés. Néanmoins, le niveau de "résolution géographique" est encore assez faible pour un SIG parce que de nombreuses administrations de subdivision de comté exercent leur compétence sur de très grands territoires. Malgré cela, il existe des données démographiques pour toutes les unités décrites ci-dessus et des données économiques pour les unités les plus populeuses.

4.2 Unités "de niveau inférieur"

Plus bas dans la hiérarchie on trouve les secteurs de recensement et leurs cousins, les secteurs de numérotation d'îlots, dont le nombre dépassait 62 000 pour le recensement de 1990. Le nombre de ces unités s'est accru sensiblement depuis le recensement de 1980 puisque ces unités couvrent maintenant tous les États-Unis et leurs territoires. Ces secteurs "découpent" les grandes unités administratives en "blocs" géographiques qui comptent en moyenne quelque 4 000 habitants. À ce niveau, la "résolution géographique" des unités de totalisation du recensement de 1990 est déjà beaucoup plus utile dans la plupart des applications du SIG. Par exemple, au niveau de la subdivision de comté ou de l'agglomération dans la hiérarchie des unités géographiques du Census Bureau, on retrouve une entité qui s'appelle "ville de Los Angeles"; au niveau du secteur de recensement, Los Angeles compte 737 unités pour lesquelles le Census Bureau produit des données détaillées. Ces secteurs de recensement sont des régions géographiques relativement petites à l'intérieur de la grande ville de Los Angeles.

Les secteurs de recensement et les SNI sont eux-mêmes divisés en quelque 229 000 groupes d'îlots qui morcellent davantage les unités administratives pour les besoins de la présentation de données. Pour le niveau du secteur de recensement et du SNI et celui du groupe d'îlots (comme pour tous les niveaux supérieurs dont il a été question précédemment), l'utilisateur du SIG a accès à tout l'éventail des données du recensement décennal -- les données recueillies auprès de chaque personne et sur chaque unité de logement et celles recueillies uniquement auprès d'un échantillon de la population et au sujet des unités de logement correspondantes.

Enfin, pour le recensement de 1990, le Census Bureau a défini et numéroté plus de 7 millions d'îlots -- polygones d'individus -- à la grandeur des États-Unis. Le Census Bureau a totalisé les données qu'il avait recueillies auprès de chaque personne et sur chaque unité de logement pour chacun de ces îlots. Comme dans le cas des secteurs de recensement et des SNI, le nombre d'îlots s'est accru très sensiblement depuis le recensement de 1980; à ce moment-là, on ne pouvait obtenir des données d'îlot que pour les noyaux urbains des régions métropolitaines essentiellement. Ces millions d'îlots permettent un niveau de "résolution géographique" très élevé pour les séries de données démographiques du Census Bureau. Ils sont délimités par les routes, les cours d'eau, les voies ferrées et les frontières des unités administratives, qui sont les éléments "géométriques" fondamentaux de la base de données TIGER, ces mêmes éléments qui forment au moins une couche dans la plupart des SIG.

4.3 Unités spéciales

La structure géographique utilisée par le Census Bureau contient plusieurs autres classes d'unités pour lesquelles le Bureau totalise bon nombre des données élémentaires qu'il recueille; ces unités correspondent à des niveaux de couverture et de "résolution géographique" variés (voir tableau 1).

- Les régions métropolitaines se composent généralement d'un comté renfermant une grande ville ou un centre urbain important et de comtés limitrophes qui ont des liens économiques avec le comté principal. Dans les États de la Nouvelle-Angleterre, les régions métropolitaines sont des groupements de villes plutôt que des groupements de comtés. Collectivement, les habitants des régions métropolitaines constituent la "population métropolitaine des États-Unis"; les habitants des autres régions constituent la "population non métropolitaine".

- Les régions urbanisées sont quelque chose de tout à fait différent. Alors que les régions métropolitaines sont définies comme des groupes d'unités administratives et qu'elles peuvent différer beaucoup entre elles par leur étendue et la densité de leur population, les régions urbanisées sont définies strictement en fonction de la densité de population. Ainsi, chaque région métropolitaine renferme au moins une région urbanisée; certaines en comptent plus d'une. Les régions urbanisées sont généralement beaucoup moins étendues que les régions métropolitaines mais ont une densité de population moyenne beaucoup plus grande. On retrouve aussi des régions urbanisées dans des comtés qui ne font pas nécessairement partie d'une région métropolitaine. Collectivement, tous les habitants des régions urbanisées de même que toutes les personnes qui vivent dans les autres municipalités ou agglomérations de recensement de 2 500 habitants ou plus constituent la "population urbaine des États-Unis"; le reste de la population est désigné comme la "population rurale".

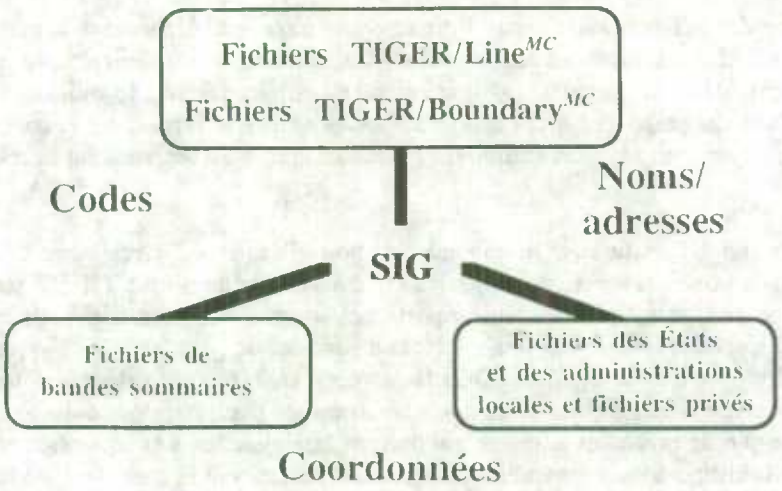
- Les produits du recensement de 1990 comprennent aussi des totalisations pour les régions habitées par les autochtones (États contigus et Alaska) et pour les districts électoraux; ils comprendront bientôt des totalisations par code postal et pour un grand nombre d'autres unités.

5. EN QUOI TIGER EST-IL UTILE POUR LES APPLICATIONS DU SIG?

Le plus gros avantage de la base de données TIGER du point de vue du SIG est la création de codes numériques (pour chaque unité géographique) qui correspondent à ceux utilisés dans les produits statistiques du Census Bureau. Grâce à ces codes, un SIG peut associer directement les données statistiques au réseau d'éléments cartographiques qui délimitent les millions d'îlots de recensement utilisés en 1990 et peut, par conséquent, raccorder ces données aux nombreux autres ensembles de données que l'on produit en quantité de plus en plus grande: données provenant des administrations des États et des administrations locales et données tirées des recensements décennaux ou économiques ou des recensements de l'agriculture (voir figure 2).

Antérieurement, lorsque les analystes de données se servaient des noms et des codes d'unités géographiques pour fabriquer des cartes qui montraient la distribution de divers éléments d'information, ils faisaient souvent leur analyse à l'aide de cartes en papier et de crayons de couleur! Dans un univers SIG où l'on utilise les produits du système TIGER, l'ordinateur fait lui-même le "coloriage" au moyen des codes numériques des produits TIGER dans le but d'afficher et d'analyser les données des recensements décennaux, des recensements sur l'économie et des recensements de l'agriculture et d'établir des liens entre ces données et le réseau d'objets cartographiques. Les codes numériques permettent à l'analyste d'étudier les caractéristiques des personnes qui occupent le territoire : leur habitation, leur ferme, leur entreprise et leur activité industrielle. Grâce à ces codes, le SIG peut présenter les données en question en fonction des domaines de compétence des divers niveaux de gouvernement et peut étudier les caractéristiques des habitants en lien avec d'autres ensembles de données réparties géographiquement -- classes de sols, lieux d'élimination de déchets dangereux, qualité de l'eau, occupation du sol, données sur les ventes, etc.

Figure 2 : SIG et le système TIGER



6. QUEL EST L'AVENIR DU SYSTÈME TIGER?

Pendant que le Census Bureau exploite le système TIGER pour la totalisation et la diffusion des données du recensement de 1990, nous sommes à tracer les plans d'avenir de ce nouveau système révolutionnaire; et cet avenir est prometteur. Un volet de cet exercice de planification a été de documenter le fait que le système TIGER et le fichier de contrôle des adresses du recensement de 1990 sont des "ressources nationales". Bien que la base de données TIGER puisse produire un plan des rues assimilable par ordinateur pour tous les États-Unis, plan dont se sert le Census Bureau pour la dimension géographique de tous ses programmes statistiques, TIGER n'offre que la moitié de ce qui est nécessaire pour chaque recensement et chaque enquête. En effet, il faut aussi une liste d'adresses; le recensement décennal est, parmi tous, celui qui nécessite la liste la plus imposante, peu importe la méthodologie que l'on choisira pour le recensement de l'an 2000. De plus, cette même liste sert de base de sondage pour les enquêtes démographiques du Census Bureau.

Toutefois, contrairement aux vins fins, ni la base de données TIGER ni la liste d'adresses du recensement de 1990 ne s'améliorent avec le temps. Les deux perdent de leur actualité à mesure que les béliers mécaniques ouvrent de nouvelles rues au milieu des terres agricoles et des terrains boisés et que les constructeurs mettent en chantier à chaque année des millions d'unités de logement. De plus, ces deux immenses fichiers existent toujours indépendamment l'un de l'autre. À l'heure actuelle, le personnel du Census Bureau doit effectuer deux mises à jour distinctes; une pour la base de données TIGER et une autre pour la liste d'adresses.

Un autre volet important de l'exercice de planification a été la rédaction d'un énoncé des perspectives d'avenir qui vise justement à orienter le processus; cet énoncé se lit comme suit :

«Nous devrons gérer selon un mode automatisé une base cartographique pour les États-Unis et leurs territoires; cette base, qui sera mise à jour continuellement et gagnera progressivement en précision, contiendra l'adresse postale et le code d'emplacement géographique de chaque unité de logement, logement collectif, exploitation agricole, entreprise et établissement industriel. Ce fichier informatique devra constamment dépasser les attentes de notre clientèle, interne comme externe, en produisant des cartes le plus à jour possible et des données correctement géocodées. Il devra aussi produire des listes d'adresses complètes qui serviront de base de sondage dans des opérations statistiques autorisées.»

«À cette fin, nous tenterons d'établir des rapports de coopération avec d'autres organisations préoccupées, comme nous, de la nécessité de disposer de données géographiques et de fichiers d'adresses récents et exacts. Nous serons alors en mesure de pousser plus loin l'innovation de sorte que les produits et les méthodes à venir répondent mieux aux besoins des enquêtés, des clients et des partenaires du Census Bureau.» (TRADUCTION)

6.1 Liste d'adresses exhaustive

Il semble que la nature des activités du Census Bureau demeurera essentiellement la même dans un avenir prévisible. C'est pourquoi le Census Bureau continuera vraisemblablement d'utiliser des listes d'adresses comme cadre pour ses recensements et ses enquêtes -- que ces listes proviennent de fournisseurs ou qu'elles soient établies par des membres du personnel à l'occasion d'opérations sur le terrain ou encore qu'il s'agisse d'un fichier interne mis à jour par comparaison automatisée pendant que d'autres assurent la mise à jour des listes proprement dites.

Tant et aussi longtemps que le Census Bureau obtiendra de nouvelles listes d'adresses ou des listes mises à jour pour ses programmes statistiques, la fonction d'appariement d'adresses du système TIGER sera toujours un outil précieux. Cette fonction se révèle l'une des plus importantes sources d'information pour la mise à jour de la base de données TIGER, spécialement pour des systèmes d'adresses de type "numéro civique/nom de rue". La résolution des cas (adresses) que le système TIGER ne peut traiter convenablement met en évidence de nouveaux objets, de nouveaux noms d'objet et de nouvelles tranches d'adresses qui doivent être ajoutés dans la base TIGER, de même que de nouvelles adresses qui doivent être ajoutées à la liste maîtresse des adresses du Census Bureau. En définitive, ces renseignements additionnels enrichiront la base de données existante de telle manière qu'on pourra accroître l'appariement automatisé d'adresses et rendre plus efficace la production cartographique.

6.2 Cartes quadrillées

De par l'existence, dans la base de données TIGER, d'un système de coordonnées qui définit l'emplacement de chaque croisement et de chaque extrémité de rue, le système TIGER offre une fonction qui peut remplacer avantageusement la méthode à forte intensité de main-d'oeuvre et sujette à l'erreur utilisée habituellement pour créer un produit très en demande, à savoir un index des noms de rues qui renvoie à une grille cartographique. Le Census Bureau a produit une version plus générale de cet index au moyen de processus automatisés lorsqu'il a créé son répertoire des noms de rue et zones d'affectation des recenseurs par comté dans le cadre du programme de cartographie du recensement de 1990. En utilisant le système TIGER pour produire les cartes pour les états imprimés du recensement de 1990, les préposés ont appliqué une grille uniforme, comportant des lettres ou des chiffres ou les deux, sur tout le territoire des États-Unis. Cela étant fait, le logiciel peut produire de l'information sur chaque maille de grille qui contient un tronçon de rue (ou tout autre objet reproduit sur la carte), peut dresser la liste des noms de rue dans un ordre alphanumérique et peut créer les listages voulus. Ce produit sera particulièrement utile dans les enquêtes du Census Bureau où le personnel sur le terrain ne doit visiter qu'un échantillon d'unités de logement. Non seulement le système TIGER produit les cartes et les index qui indiquent au représentant sur le terrain l'emplacement d'une unité d'échantillon, mais aussi il permet, grâce à la technologie du SIG, de déterminer le chemin le plus court pour se rendre à cette unité.

6.3 Fichiers de référence géographique

De même, bien qu'elle soit essentiellement une fonction non cartographique, la création de fichiers de référence géographique à partir de la base de données TIGER demeurera une fonction majeure du système TIGER. Ces

fichiers nous renseignent sur la hiérarchie des unités géographiques qui servent à la totalisation des données du Census Bureau et sur les relations entre ces unités. Ils permettent d'identifier l'unité géographique dans les états imprimés et les fichiers de bandes sommaires issus de chaque recensement et de chaque enquête, y compris ceux qui seront produits après le recensement de 1990.

6.4 Structure de la base de données TIGER

La structure de la base de données TIGER demeure l'une des plus grandes caractéristiques de pointe du système TIGER. Elle établit des liens directs entre les unités géographiques pour lesquelles le Census Bureau totalise des données. Elle établit aussi des liens directs entre la représentation d'objet de ces unités et la représentation d'objet du réseau de caractéristiques correspondant qui forme la base cartographique de nombreuses cartes et de nombreux produits visuels du système d'information géographique. Fait révélateur, elle établit ces liens sans qu'il y ait superposition multiple de polygones limitrophes, qui est le fait de nombreux systèmes de cartographie automatiques.

La méthode de stockage-extraction de la base de données TIGER élimine tous les problèmes que l'on rencontre habituellement lorsqu'un seul trait représente plusieurs objets cartographiques à la fois; par exemple, la frontière d'un État qui est, à la fois, une limite de comté, une limite de township, une limite de municipalité et une route. Peu importe l'échelle qu'un cartographe définit pour une carte (image graphique), et peu importe les catégories d'information qu'il choisit de présenter, les relations sont constantes. Ainsi, il est impossible que la représentation de plus d'une catégorie d'objets à la fois (par exemple, une route qui longe une frontière) produise des généralisations ou des tracés différents par rapport aux objets fondamentaux.

Néanmoins, il est nettement justifié de revoir la structure des fichiers qui constituent la base de données TIGER en se fondant sur l'expérience du traitement des données du recensement de 1990. Bien que la fonction d'appariement d'adresses et la fonction de création de fichiers de référence géographique du système TIGER soient toujours des éléments indispensables dans l'élaboration des logiciels d'application futurs pour le système TIGER, les données dont nous disposons sur l'utilisation des ressources informatiques pour chaque catégorie d'applications révèlent que les activités les plus exigeantes au point de vue calcul ont été la production de cartes et l'organisation des données de la base TIGER en fonction de cette opération. Il faudra donc s'attacher à reconsidérer la décision qui avait été prise à l'origine et qui visait à réduire la redondance dans le stockage des données cartographiques au prix de frais de traitement plus élevés.

6.5 Projets de mise à jour menés en collaboration

Le Census Bureau a toujours utilisé des méthodes de comparaison de cartes à forte intensité de main-d'oeuvre pour mettre à jour les objets qui figurent sur ses cartes (et, désormais, dans la base de données TIGER). La diffusion de plus en plus grande des systèmes de cartographie automatisée et de la technologie du SIG, conjuguée à la sophistication des techniques de production de cartes et à la complexité des besoins des utilisateurs à tous les niveaux de l'administration publique et dans le secteur privé, rend très intéressante la perspective de participer à des programmes de coopération et d'appliquer des méthodes de transfert automatisé.

Une des grandes réussites du système TIGER jusqu'à ce jour a été l'établissement de liens de coopération entre le Census Bureau et le USGS. Celui-ci maintiendra fort probablement ses liens étroits avec le Census Bureau et il se pourrait que d'autres organismes joignent les rangs. Par exemple, le Census Bureau cherche actuellement, en collaboration avec le United States Postal Service (USPS), des moyens de mettre à jour les données sur les adresses et les voies de communication à l'intérieur d'une base de données qui serait le produit de la fusion de la base TIGER et d'une liste nationale d'adresses (voir tableau 4). Un autre projet est en cours en Caroline du Nord, où de nombreuses organisations se sont donné la main pour élaborer et tenir à jour des fichiers cartographiques numériques plus détaillés; parmi ces organisations on compte un organisme de l'État, les administrations locales, l'American Association of State Highway and Transportation Officials, le USGS et le Census Bureau. Un projet semblable est en cours au Texas, où le Census Bureau travaille en collaboration avec les organismes chargés d'établir les réseaux téléphoniques d'urgence "911". L'innovation et la coopération ont bel et bien leur place au Census Bureau.

6.6 Produits numériques et SIG

Pour que le système TIGER soit jugé vraiment utile en dehors du Census Bureau, le même travail de planification devra se faire dans toutes les organisations aux États-Unis et à l'étranger, plus spécialement dans les organisations étrangères qui souhaitent exercer des activités commerciales aux États-Unis. Grâce aux outils analytiques que la technologie du SIG, désormais accessible au secteur privé, met à notre disposition pour étudier la corrélation entre divers ensembles de données, nous pouvons vraiment avoir une meilleure compréhension des choses.

BIBLIOGRAPHIE

Carbaugh, L.W., et Marx, R.W. (1990). The TIGER System: A Census Bureau innovation serving data analysts. *Government Information Quarterly*, 7, 3, 285-306.

LaMacchia, R.A., Tomasi, S.G., et Piepenburg, S.K. (1987). The TIGER File: Proposed Products. Paper distributed at the fall meeting of the National Conference of State Legislatures, Hartford, Connecticut.

Marx, R.W. (collaborateur spécial) et coll. (1990). Special Content: The Census Bureau's TIGER System. *Cartography and Geographic Information Systems*, 17, 1, 9-113.

Tomasi, S.G. (1990). Why the nation needs a TIGER System. *Cartography and Geographic Information Systems*, 17, 1, 21-26.

Tableau 1. Unités géographiques utilisées pour le recensement de 1990 et d'autres recensements récents

Niveau de détail	Genre d'unité géographique	Nombre d'unités géographiques avec données* — Recensements décennaux 1980[2]	Recensements décennaux 1990[3]	Recensement de 1987 — Économie[4]	Recensement de 1987 — Agriculture[5]
Très grossier (études nationales)	Nation (les États-Unis)[1]	1	1	1	1
	Régions (des États-Unis)[1]	4	4	4	—
	Divisions (des États-Unis)[1]	9	9	9	—
	États et régions statistiquement équivalentes	57[7]	57[7]	55[6]	53[6]
Grossier (études nationales/études à l'échelle des États)	Comtés et régions statistiquement équivalentes	3 231	3 248	3 228[6]	3 179[7]
	Subdivisions de comté et agglomérations	59 451	60 228	7 287[10,13]	—
Subdivisions de comté	Division civiles secondaires - DCS	30 450	30 386	—	—
	Sous-div. civiles secondaires	265	145	—	—
	Divisions de recensement - DR	5 512	5 581	—	—
	Territoires non organisés - TN	274	282	—	—
	Autres régions statistiquement équivalentes	41[8]	40[9]	—	—
Agglomérations	Agglomérations constituées en municipalité	19 176[10]	19 365[10]	6 776[11]	—
	Regroupements de municipalités	—	6	—	—
	Agglomérations de recensement - AR	3 733[10]	4 423[10]	44[10]	—
Autres régions apparentées	Zones urbaines économiques spéciales - ZUES	—	—	433[13]	—
	Parties restantes de régions métropolitaines	—	—	34[13]	—
	Régions d'autochtones (États contigus et Alaska)	499	579	—	—
États contigus	Réserves indiennes (à l'exclusion des terres sous tutelle)	241	259	—	—
	Territoires indiens, y compris les terres sous tutelle	37	52	—	—
	Zones statistiques de compétence tribale - ZSCT	—	17	—	—
	Zones statistiques à désignation tribale - ZSDT	—	19	—	—
Alaska	Villages d'autochtones - VA	209	(Voir ZSVA)	—	—
	Zones statistiques de villages d'autochtones - ZSVA	(Voir VA)	217	—	—
	Corporations régionales d'autochtones - CRA	12	12	—	—
	Régions métropolitaines (RM)/régions urbanisées (RU)	—	—	—	—
"Métropolitaine"[17]	Zones statistiques métropolitaines normalisées - ZSMN	323	—	—	—
	Zones statistiques regroupés normalisées - ZSRN	17	—	—	—
	Zones statistiques métropolitaines - ZSM	—	267	265	—
	Zones statistiques métropolitaines regroupées - ZSMR	—	21	21	—
"Urbaine"[18] toutes les agglomérations de 2 500 habitants ou plus et toutes les RU	Zones statistiques métropolitaines majeures - ZSMM	—	73	73	—
	Régions urbanisées - RU	373	405	—	—
	Unités spéciales	—	—	—	—
	Districts électoraux fédéraux	435	435	—	—
	Districts scolaires	16 075	16 000[E]	—	—
Moyen (études à l'échelle du comté/études à caractère local)	**Secteurs de recensement/secteurs de numérotation d'îlots (SNI)**	47 114	62 276	—	—
	Unités de totalisation	n.d.	145 035	—	—
	Groupes d'îlots (GI)	258 398[14]	229 192	—	—
	Unités de totalisation	300 192	356 742	—	—
	Unités spéciales	—	—	—	—
	Unités de voisinage	28 381	—	—	—
	Zones d'analyse de la circulation - ZAC	160 000[E]	200 000[E]	—	—
	Districts électoraux - DE	36 361[15]	148 874[15]	—	—
	Zones postales	37 000[E]	40 000[E]	31 000[E]	31 000[E]
Fin (études à caractère local/études de quartier)	Îlots	2 545 416[16]	7 017 427	—	—

* Situation au 1er janvier de l'année de recensement indiquée.

1-18 Voir page ci-contre pour des explications détaillées sur chaque élément affecté d'un renvoi.

— Sans objet E = estimation n.d. = non disponible

Tableau 1 (suite). Unités géographiques utilisées pour le recensement de 1990 et d'autres recensements récents

[1] Officiellement, les "États-Unis" se composent des 50 États et du District de Columbia. À chaque recensement, le Census Bureau produit des totalisations détaillées pour les 51 unités qui constituent les États-Unis et pour plusieurs unités statistiquement équivalentes (voir les notes 2 à 5 pour des précisions concernant chaque recensement récent); ces dernières unités sont souvent appelées collectivement "Porto Rico et les régions périphériques". Tous les chiffres de ce tableau représentent un nombre d'unités géographiques contenues dans ces "États". La base de données TIGER renferme également, au même titre:

- Les États fédérés de Micronésie et les îles Marshall (anciennement du Trust Territory of the Pacific Islands). Le Census Bureau a inclus ces unités dans la base pour le cas où on lui demanderait de prêter son concours pour un recensement dans l'une ou l'autre de ces unités.

- Les îles Midway. Le Census Bureau a inclus ces îles dans la base de données afin de couvrir tout le territoire contenu dans les limites de l'État d'Hawaii.

Les chiffres du tableau ont trait seulement aux unités géographiques contenues dans les "États" mentionnés à la note 3. S'il fallait tenir compte des trois unités mentionnées ci-dessus, le nombre d'unités géographiques pour chaque niveau à la date du recensement décennal de 1990 serait le suivant: 60 "États", 3 286 "comtés", 60 420 subdivisions de comté et agglomérations (y compris 30 536 DCS, 186 sous-divisions civiles secondaires et 4 423 AR), 62 392 secteurs de recensement et secteurs de numérotation d'îlots (qui comptent 145 196 unités de totalisation), 229 466 groupes d'îlots (qui comptent 357 038 unités de totalisation) et 7 020 924 îlots.

[2] Outre les 50 États et le District de Columbia (les États-Unis), les unités suivantes étaient visées par le recensement décennal de 1980: Samoa américaines, Guam, Mariannes du Nord, Porto Rico, ce qui reste du Trust Territory of the Pacific Islands, et les îles Vierges américaines.

[3] Outre les 50 États et le District de Columbia (les États-Unis), les unités suivantes étaient visées par le recensement décennal de 1990: Samoa américaines, Guam, Mariannes du Nord, Palau, Porto Rico et les îles Vierges américaines.

[4] Outre les 50 États et le District de Columbia (les États-Unis), les unités suivantes étaient visées par les recensements de 1987 sur l'économie: Guam, Mariannes du Nord, Porto Rico et les îles Vierges américaines.

[5] Outre les 50 États (dans ce cas-ci, le District de Columbia est exclu car il n'est pas touché par le recensement de l'agriculture), les unités suivantes étaient visées par le recensement de l'agriculture de 1987: Guam, Porto Rico et les îles Vierges américaines. Le Census Bureau a profité du recensement décennal de 1990 pour faire un recensement de l'agriculture dans les Samoa américaines et les Mariannes du Nord.

[6] Lors des recensements de 1987 sur l'économie, on a totalisé des données non seulement pour les comtés et les unités équivalentes qui forment les "États" visés par les recensements (voir note 4), mais aussi pour les sept entités côtières suivantes, qui étaient considérées comme les équivalents statistiques de comtés:

Alaska	Louisiane	Atlantique	Nord du golfe du Mexique
Californie	Texas	Pacifique	

[7] En ce qui concerne le recensement de l'agriculture de 1987, on a totalisé des données pour les comtés et les unités équivalentes qui forment les "États" visés par ce recensement (voir note 5), sauf quelques exceptions:

- On a groupé les 23 municipalités et zones de recensement que comptait l'Alaska en 5 unités géographiques que l'on considérait comme les équivalents statistiques de comtés;

- On n'a pas produit de données spécifiquement pour les villes autonomes ni pour la plupart des comtés dont le territoire correspond à celui d'une agglomération constituée en municipalité;

- On a exclu quelques autres comtés et unités statistiquement équivalentes.

[8] Ce nombre comprend les 37 "sous-zones de recensement" de l'Alaska et les 4 "quadrants" du District de Columbia.

[9] Ce nombre comprend les 40 "sous-zones de recensement" de l'Alaska. À la demande de l'administration du District de Columbia, le Census Bureau n'a pas fait de totalisations pour les "quadrants" en 1990; ceux-ci ont été remplacés par une DCS appelée "Washington".

[10] Conformément à la volonté de l'État d'Hawaii, le Census Bureau ne reconnaît pas la ville d'Honolulu – dont le territoire se confond avec celui du comté d'Honolulu – comme une agglomération constituée en municipalité pour les besoins de la présentation des données. L'État définit plutôt des AR qui correspondent aux diverses communautés du comté d'Honolulu, et pour lesquelles le Census Bureau totalise des données.

Tableau 1 (suite). Unités géographiques utilisées pour le recensement de 1990 et d'autres recensements récents

[11] Les recensements de 1987 sur l'économie ne visaient que les agglomérations constituées en municipalités de 2 500 habitants ou plus; cependant, on note trois exceptions à cette règle: il s'agit d'agglomérations de taille moindre mais dont la vocation est surtout commerciale ou industrielle.

[12] Dans les recensements de 1987 sur l'économie, les divisions civiles secondaires des six États de la Nouvelle-Angleterre, du New Jersey et de la Pennsylvanie qui comptaient 10 000 habitants ou plus étaient incluses parmi les "agglomérations".

[13] Dans les six États de la Nouvelle-Angleterre, on a regroupé les données des portions de comté qui n'étaient pas reconnues comme des équivalents statistiques d'agglomérations dans une région métropolitaine.

[14] Ce nombre comprend 102 235 unités que le Census Bureau appelait "districts de recensement" (DR) en 1980. Pour le recensement de 1990, les groupes d'îlots couvraient tout le territoire des États-Unis.

[15] Ce nombre comprend seulement les unités visées par les dispositions de la loi 94-171.

[16] En 1980, le Census Bureau a produit des totalisations par îlot pour un nombre limité d'unités: les régions urbanisées et leurs environs, les autres agglomérations constituées en municipalité de 10,000 habitants ou plus et d'autres unités qui avaient choisi de confier au Census Bureau le mandat de totaliser des données par îlot. (Parmi ces unités, on retrouvait cinq États: Georgie, Mississippi, New York, Rhode Island et Virginie.)

[17] La population "métropolitaine" des États-Unis est l'ensemble des personnes qui vivent dans des régions métropolitaines (les ZSM et les ZSMR); la population "non métropolitaine" comprend tous les autres habitants des États-Unis.

[18] La population "urbaine" des États-Unis est l'ensemble des personnes qui vivent dans les régions urbanisées et dans les agglomérations (constituées en municipalité et AR) de 2 500 habitants ou plus qui ne font pas partie de régions urbanisées; la population "rurale" comprend tous les autres habitants des États-Unis.

Tableau 2. Produits sur support papier du Système TIGER

Cartes "comtés-îlots" (produites par traceur électrostatique). Cartes montrant pour chaque comté les régions d'autochtones (États contigus et Alaska), les subdivisions de comté (DCS ou DR), les agglomérations (celles constituées en municipalité et les agglomérations de recensement), les secteurs de recensement ou les secteurs de numérotation d'îlots (SNI) et les îlots de recensement pour 1990.

Carte "comtés-îlots" P.L. 94-171 (1990) (disponible) -- montre les districts électoraux par État.

Carte "comtés-îlots" (1990) (disponible) -- ne montre **pas** les districts électoraux.

Cartes "îlots" axées sur l'unité (1990) (fin 1991) -- Quatre séries de cartes: Carte "îlots" pour les régions d'autochtones (États contigus) (1990); Carte "îlots" pour les régions d'autochtones (Alaska) (1990); Carte "îlots" pour les subdivisions de comté (1990) et Carte "îlots" pour les agglomérations (1990) -- mettent l'accent sur certaines unités administratives et unités statistiquement équivalentes qui ne sont pas des comtés -- contiennent les mêmes renseignements que la carte "comtés-îlots" (1990).

Carte "comtés-îlots" (1990), Version 2 (automne 1992) -- feuilles de carte produites lorsque l'opération informatique qui vise à créer les fichiers Meta de la Carte "comtés-îlots" (1990) donne un tracé qui diffère du tracé original de la Carte "comtés-îlots" (1990); cette version est assortie aux fichiers Meta sur CD-ROM.

Cartes à grandes lignes (produites par traceur électrostatique). Ces cartes montrent les régions d'autochtones (États contigus et Alaska), les États, les comtés, les subdivisions de comté (DCS et DR), les agglomérations, le domaine de la série de cartes et certains objets cartographiques de base à une petite échelle.

Cartes à grandes lignes de subdivisions de comté (janvier 1992) -- par État.

Cartes à grandes lignes de districts électoraux (disponibles) -- par comté; existent uniquement pour les comtés des États qui ont participé au programme de délimitation des districts électoraux du Census Bureau.

Cartes à grandes lignes de secteurs de recensement et de secteurs de numérotation d'îlots (disponibles) -- par comté.

Cartes de limites des régions urbanisées (décembre 1991) -- par région urbanisée

Cartes à grande lignes (imprimés). Ces cartes montrent les régions d'autochtones (États contigus et Alaska), les États, les comtés, les subdivisions de comté (DCS et DR), les agglomérations, le domaine de la série de cartes et certains objets cartographiques de base à une petite échelle.

Cartes à grandes lignes de régions d'autochtones (États contigus et Alaska) (printemps-été 1992) -- par région d'autochtones.

Cartes à grandes lignes d'États et de régions métropolitaines (printemps-été 1992) -- par État.

Cartes à grandes lignes d'États et de comtés (disponibles) -- par État.

Cartes à grandes lignes de districts électoraux fédéraux -- 103è. Congrès (automne 1992) -- par État.

Cartes à grandes lignes de subdivisions de comté (disponibles) -- par État.

Cartes à grandes lignes de régions urbanisées (printemps-été 1992) -- par région urbanisée.

Cartes à grandes lignes de secteurs de recensement et de secteurs de numérotation d'îlots (automne 1992) -- par comté.

Tableau 3. Produits sur support informatique du système TIGER

Fichiers Meta de la Carte "comtés-îlots" (1990) (sur CD-ROM seulement -- automne 1992.) Fichiers texte (ASCII) qui renferment des commandes graphiques simples et complexes (par ex.: «plume levée», «plume abaissée» et «remplir polygone») qui servent à produire les feuilles de carte de la Carte "comtés-îlots" (1990). L'information indique aussi le territoire qui correspond à chaque image cartographique. Ce produit permet de visualiser l'information qui figure sur les cartes produites par traceur électrostatique.

Fichiers TIGER/Line[MC] pour le recensement de 1990 (disponibles -- sur CD-ROM, par État ou groupe d'États, et sur bande magnétique, par comté ou groupe de comtés d'un même État, selon les précisions de l'acheteur). La version sur CD-ROM contient aussi le fichier GRF-N pour les États qui se trouvent sur le disque compact.

Fichiers TIGER/SDTS[MC] (sur bande magnétique seulement - milieu de 1992; prototype sur CD-ROM pour quelques comtés d'essai - début de l'hiver 1991-1992). Ce fichier présente l'information ponctuelle, linéaire et aréolaire tirée de la base de données TIGER sous une forme qui respecte la norme de conversion de données spatiales (Spatial Data Transfer Standard - SDTS) de la FIPS.

Fichiers TIGER/GRF-N[MC] (Geographic Reference File - Names) (disponibles -- sur bande magnétique; viennent aussi avec les fichiers TIGER/Line[MC] pour le recensement de 1990 sur CD-ROM). Fichiers d'État des noms et des codes géographiques pour les unités géographiques du recensement de 1990 classées par catégorie.

TIGER/GICS[MC] (sur bande magnétique - printemps 1992; sur support papier - milieu de 1992). Fichiers d'État des noms et des codes géographiques pour les unités géographiques du recensement de 1990 au niveau de l'agglomération et aux niveaux supérieurs, présentés selon un ordre hiérarchique. La version sur bande magnétique contiendra aussi de l'information aréolaire et ponctuelle pour toutes les unités géographiques contenues dans le fichier.

Fichiers TIGER/Boundary[MC] (sur bande magnétique - dès le milieu de 1992). Six fichiers qui renferment une représentation numérique des frontières des unités spécifiées dans le titre de chaque produit. Ces fichiers contiennent aussi le tracé littoral des principaux plans d'eau. Ils seront disponibles en deux versions: l'une avec un ensemble complet de coordonnées, l'autre avec un ensemble "réduit" de coordonnées, qui convient mieux au micro-ordinateur.

État/Comté (fichier national)

DR (fichier national)

Région d'autochtones/subdivision de comté/agglomération (fichiers d'État)

Région d'autochtones (fichier national)

Secteur de recensement/SNI/GI (fichiers d'État)

RU (fichier national)

Fichier TIGER/Census Tract Comparability[MC] (sur bande magnétique - fin 1991). Par groupe de comtés d'un même État, selon les précisions de l'acheteur.

TIGER/Census Tract Street Index[MC] (sur bande magnétique et sur imprimé - fin 1991; version mise à jour sur CD-ROM - fin 1992). Fichier axé sur le comté et comportant les numéros des secteurs de recensement/SNI pour les rues identifiées avec des tranches d'adresses.

Fichier TIGER/Map Sheet - Geography[MC] (fichier national sur bande magnétique - hiver 1991-1992). Ce fichier définit les feuilles de la Carte "comtés-îlots" (1990) qui sont nécessaires pour décrire les régions d'autochtones, les subdivisions de comté, les agglomérations et les secteurs de recensement/SNI.

Fichier TIGER/Map Sheet Corner Point Coordinate[MC] (fichier national sur bande magnétique - disponible). Ce fichier renferme les coordonnées sphériques de points (latitude et longitude) pour chaque feuille de carte de la série des Cartes "comtés-îlots" (1990).

Tableau 4. Projet de coopération TIGER entre le Census Bureau et le U.S. Postal Service

Forces actuelles
(et leurs faiblesses)

Système TIGER dans sa forme actuelle

Forces

- La base de données géographiques automatisée la plus complète au pays à ce jour
- Utilisé largement par d'autres organismes gouvernementaux et des organismes du secteur privé
- Offre des produits et des services géographiques pour les besoins des recensements sur l'économie et du recensement de l'agriculture, pour le remaniement des enquêtes courantes et pour les besoins des autres programmes du Census Bureau

Faiblesses

- Absence des tranches d'adresses pour 40% des adresses au pays
- Absence des codes de zone postale à 5 chiffres pour 40% des adresses au pays
- Absence des codes de zone postale à 9 chiffres pour toutes les adresses
- Retard de 2 à 8 ans dans la mise à jour des rues
- Représentation inexacte des rues des grands centres urbains sur les cartes du système TIGER

Liste d'adresses du recensement de 1990

Forces

- Liste automatisée complète des adresses domiciliaires au pays en date de 1990
- Chaque adresse est géocodée en fonction des zones de totalisation du recensement de 1990

Faiblesses

- N'est pas intégrée à TIGER
- Le format des adresses n'est pas conforme à celui exigé par le U.S. Postal Service
- Absence des codes de zone postale à 9 chiffres
- La liste d'adresses est déjà vieille de 2 ans
- Adresses ajoutées par les recenseurs souvent sur des rues qui ne figurent pas sur les cartes du système TIGER

Grande expertise du U.S. Bureau of the Census en matière de géographie

PROJET DE COOPÉRATION TIGER ENTRE LE CENSUS BUREAU ET LE U.S POSTAL SERVICE

Distribution quotidienne du courrier à l'échelle du pays

Produits-Avantages-Possibilités
(découlant du projet de coopération)

Système TIGER amélioré comprenant une liste nationale permanente d'adresses

- Cartes TIGER plus exactes et plus à jour pour tous les clients internes et externes
- Si l'on dispose de tous les codes de zone postale et de toutes les tranches d'adresses, on obtiendra des taux de géocodage plus élevés pour les recensements sur l'économie et les recensements de l'agriculture et un meilleur soutien pour les enquêtes courantes et les programmes d'estimation
- On disposera d'une liste d'adresses à jour qui répond aux normes du USPS pour le recensement décennal de l'an 2000 et tous les autres programmes du Census Bureau
- Élimination des opérations liées à la création de la liste d'adresses du recensement décennal; ces opérations, coûteuses et sujettes à l'erreur, sont exécutées normalement par une main-d'oeuvre très éparse, difficile à gérer et qui est en majeure partie temporaire
- La façon la plus rentable de tenir à jour la base de données TIGER et la liste nationale permanente d'adresses qui s'y rattache
- Il sera possible de faire des totalisations par district électoral fédéral à partir des données du recensement sur l'économie et du recensement de l'agriculture
- On pourra établir des liens géographiques entre les totalisations tirées du recensement sur l'économie et celles tirées du recensement décennal
- Le USPS jouira désormais des systèmes de géographie automatiques dont il a besoin pour améliorer ses opérations de gestion, de planification et de contrôle
- Mise à jour continue de la base géographique nécessaire à l'établissement des réseaux téléphoniques d'urgence (E-911) à la grandeur du pays
- L'utilisation quotidienne de la base de données TIGER et de la liste d'adresses par les facteurs du USPS favorisera une meilleure assurance de la qualité; on déclarera les erreurs et apportera les corrections nécessaires quotidiennement
- En partageant le coût des opérations de mise à jour de TIGER et de la liste d'adresses, le Census Bureau et le USPS peuvent jouir d'un produit de qualité supérieure que ni l'un ni l'autre ne pourraient s'offrir seuls

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

LA CRÉATION D'UN REGISTRE D'ADRESSES RÉSIDENTIELLES À STATISTIQUE CANADA

L. Swain, J.D. Drew, B. Lafrance, K. Lance[1]

RÉSUMÉ

Le registre des adresses est une base de sondage d'adresses résidentielles pour les centres urbains de moyenne et de grande dimension qui figurent dans le Fichier principal de région (FPR) de la Division de la géographie de Statistique Canada. Pour la Colombie-Britannique, le registre des adresses a été augmenté afin d'inclure des agglomérations urbaines plus petites ainsi que certaines régions rurales. Dans la présente communication, on présente un aperçu historique du projet, ses objectifs comme moyen de réduire le sous-dénombrement lors du recensement du Canada de 1991, ses sources et produit, la méthodologie requise pour sa mise en application initiale, l'évaluation postcensitaire proposée et des perspectives pour l'avenir.

MOTS CLÉS: Registre des adresses; sous-dénombrement du recensement; systèmes d'information géographique (SIG).

1. OBJECTIF

Le concept d'un registre des adresses à Statistique Canada remonte aux années 60. Fellegi et Krotki (1967) ont, pour la première fois, considéré la création d'un de ces registres pour le recensement de 1971, en se basant sur des fichiers administratifs. Leur méthode était surtout manuelle et a donné un ensemble très complet d'adresses avec un sous-dénombrement et un surdénombrement minimum. Au milieu des années 70 (Booth 1976), l'idée est réapparue lors de la planification du recensement de 1981. Cette fois, la méthode utilisée commençait avec la saisie des adresses recueillies au cours du recensement précédent et on ajoutait à ces données des renseignements fournis par Postes Canada. Dans les deux cas, les listes d'adresses produites étaient considérées comme une base de sondage pour un recensement avec envoi par la poste. Cependant, les coûts de création étaient élevés et, pour être efficaces, ils auraient entraîné des réductions compensatrices dans d'autres opérations du recensement. De plus, on considérait que les risques associés au fait de changer la méthode traditionnelle de dénombrement étaient trop élevés. C'est pourquoi, la production d'un registre des adresses a été suspendue dans chaque cas.

Il y a eu un renouveau d'intérêt pour le concept de registre des adresses à la suite de la Conférence internationale sur la planification du recensement de 1991 (Royce 1986, 1987) qui s'est tenue en octobre 1985. Cet intérêt découlait de la possibilité d'automatiser la méthode de Fellegi et Krotki suite aux progrès technologiques, comme la disponibilité de fichiers administratifs, renfermant les adresses et les codes postaux, sur support exploitable par une machine et l'élaboration de logiciels internes pour décomposer les adresses en éléments standard, pour affecter les codes postaux et pour coupler ces derniers aux éléments géographiques du recensement. Cet intérêt découle aussi de l'élaboration d'une théorie statistique portant sur le couplage d'enregistrements (Fellegi et Sunter 1969) et de systèmes informatiques basés sur cette théorie (Hill et Pring-Mill 1985).

---

[1] L. Swain et B. Lafrance, Division des méthodes d'enquêtes sociales; J.D. Drew, Division des enquêtes des ménages; K. Lance, Division de la géographie, Statistique Canada, Ottawa (Ontario), Canada K1A 0T6.

Suite à cet intérêt, un projet a été lancé en 1986, la première recherche (Gamache-O'Leary et coll. 1987) portant sur l'étude de l'utilisation d'un registre des adresses pour un recensement réalisé par envoi postal plutôt qu'avec la méthode traditionnelle de livraison. Cette étude a conclu que la nouvelle méthode de collecte des données du recensement serait moins dispendieuse seulement si la qualité du registre des adresses était telle qu'on ne devrait effectuer qu'un minimum de mises à jour sur le terrain avant le recensement. Deux petits registres-pilotes créés au début de 1987 ont permis d'établir la couverture du registre des adresses à 90-95%, ce qui était inacceptable sans mise à jour sur le terrain (Drew et coll. 1987), éliminant ainsi l'utilisation d'un registre des adresses pour un recensement par envoi postal.

Toutefois, les deux registres-pilotes ont permis d'établir les possibilités qu'offrait un registre des adresses pour aider à améliorer la couverture quand il était employé avec la méthode traditionnelle de livraison. Cela cadrait bien avec l'apparition de l'amélioration de la couverture comme un des éléments prioritaires pour le recensement de 1991. Les résultats de la contre-vérification des dossiers pour le recensement de 1986 avaient montré une augmentation considérable dans le taux de sous-dénombrement comparativement aux recensements antérieurs (de 2.01% en 1981 à 3.21% en 1986 pour la population totale du pays; de 2.08% en 1981 à 3.28% en 1986 pour la population urbaine à l'échelle nationale) (Statistique Canada 1990). Il a donc été décidé que *le projet de recherche devrait se concentrer sur l'élaboration du registre des adresses comme méthode d'amélioration de la couverture pour le recensement de 1991.*

Dans la section ci-après on décrit les deux principaux essais effectués pour élaborer et améliorer les procédures utilisées afin de créer le registre des adresses pour le recensement de 1991. De plus, dans la deuxième section on décrit sommairement l'accord conclu avec la province de la Colombie-Britannique visant à augmenter le registre des adresses. Dans la troisième section on présente les sources administratives et géographiques utilisées lors du processus de production ainsi que la structure et le contenu des cahiers du registre des adresses, le produit final utilisé par les recenseurs sur le terrain. Dans la quatrième section on décrit la méthodologie employée pour exploiter les sources disponibles afin de produire les cahiers du registre des adresses. Dans la cinquième section on discute de l'évaluation postcensitaire proposée alors que dans la dernière section on présente les perspectives futures pour le registre des adresses. Un rapport distinct renfermant une évaluation détaillée de la méthodologie sera produit plus tard.

2. DONNÉES DE BASE

2.1 L'essai de novembre 1987 de méthodes d'amélioration de la couverture

Un essai important de l'utilisation du registre des adresses (RA) comme outil pour améliorer la couverture a été réalisé en novembre 1987 dans cinq grandes villes où l'on trouve un bureau régional. Cet essai avait été conçu pour estimer à la fois le sous-dénombrement et le surdénombrement des logements pour la méthode traditionnelle de listage du recensement et pour deux méthodes expérimentales utilisant un registre des adresses: le post-listage et le pré-listage. Quand il appliquait la méthode du post-listage, le recenseur compilait la liste des logements selon les méthodes habituelles du recensement (en créant un registre des visites) puis il la conciliait avec une liste de logements pour le secteur de dénombrement (SD) produite à partir du RA. Des suivis sur le terrain ont été effectués lorsqu'il y avait des écarts, au niveau des adresses, entre les deux listes. Pour la méthode du pré-listage, on a remis le RA au recenseur à l'avance et ce dernier en a effectué la mise à jour lors d'une prospection du SD afin de créer la liste finale des logements.

Les résultats (van Baaren 1988) concluaient que la méthode du post-listage était la méthode la plus efficace pour améliorer la couverture. Cette méthode appliquée comme simple ajout au processus normal de dénombrement du recensement était totalement sûre. Si pour une raison quelconque nous ne produisions pas le RA (soit en entier, soit en partie) à temps pour le recensement de 1991, l'étape de conciliation à l'aide du RA pouvait tout simplement être éliminée sans que cela ait un effet sur le processus de dénombrement traditionnel. Les données d'essai ont aussi fourni des estimations de l'importance de l'amélioration de la couverture et des coûts (Royce et Drew 1988). On a estimé que 34 000 logements occupés et 68 000 personnes seraient ajoutés, à la suite de la production du RA, aux centres urbains, qui comptent une grande population ou une population moyenne, pour lesquels ce RA serait produit (ces centres urbains représentant les régions pour lesquelles le fichier principal de région existe, c.-à-d. qu'ils renferment environ 65% de la population canadienne). Cela représenterait une

amélioration de la couverture de 0.26 point (le taux national de sous-dénombrement en 1986 étant estimé à 3.21%). Par rapport aux deux tentatives antérieures pour produire un RA, on a prouvé que les coûts, pour le recensement, étaient faibles à cause de la méthode très automatisée et de l'avantage démontré. De plus, le risque était minimisé puisque la méthode de collecte traditionnelle serait encore utilisée. Basé sur ce coût, sur l'avantage offert et sur l'évaluation du risque, on a autorisé la création d'un RA pour le recensement de 1991.

Deux questions se sont posées après l'essai de novembre 1987. Premièrement, le classement des adresses dans les cahiers du RA produits pour chaque secteur de dénombrement (SD) ne correspondait pas à leur ordre dans les Registres des visites, ce qui faisait de la conciliation une tâche ennuyeuse et prenant beaucoup de temps. Deuxièmement, le surdénombrement global qui s'établissait à 17% semblait encore trop élevé et il fallait plus d'efforts pour éliminer les enregistrements mal classés ou en double. On s'est attaqué à ces deux problèmes en améliorant les méthodes utilisées pour apparier le RA aux éléments géographiques du recensement. Plutôt que de coupler les adresses seulement aux SD comme on l'avait fait lors de l'essai de novembre, on a élaboré des procédures pour apparier le RA aux côtés d'îlot qui figurent dans le fichier principal de région (FPR) (Statistique Canada 1988). Un algorithme a été produit afin de trier les adresses par îlot et à l'intérieur d'un l'îlot dans l'ordre dans lequel on les rencontrerait si l'on parcourait l'îlot à pied.

2.2 L'essai de septembre 1989 visant à améliorer les procédures

Un autre essai important a été réalisé en septembre 1989, il portait sur quatre villes de taille différente: Moncton, Laval, Brampton et Calgary. Chacune de ces villes a été choisie à cause des difficultés particulières qui pourraient s'y présenter d'après les renseignements obtenus lors de l'essai de novembre 1987. Les résultats (Dick 1990) ont montré une diminution importante dans la couverture qui est passée de 84% lors de l'essai de 1987 à 73%, un résultat décourageant. Par contre, cet essai a fait ressortir une réduction considérable du surdénombrement qui a diminué de 17% à 8%. Il est important de remarquer qu'en dépit de la couverture réduite du RA, le rendement de ce dernier comme outil pour améliorer la couverture du recensement était encore acceptable. Après analyse, on a trouvé que la nouvelle opération de géocodage était problématique, tant pour ce qui est des coûts élevés puisqu'elle comportait beaucoup d'opérations effectuées par des employés de bureau que pour sa qualité. Les étapes du géocodage ont donc été améliorées en vue de la production du RA. L'adoption du logiciel de couplage d'enregistrements CANLINK (Statistique Canada 1989b) afin d'améliorer la qualité et de réduire les coûts du couplage du RA avec le FPR a été un aspect clé de cette amélioration.

2.3 Accord avec la province de la Colombie-Britannique

Au Ministry of Finance and Corporate Relations de la Colombie-Britannique on s'inquiétait du taux élevé de sous-dénombrement dans la province lors du recensement de 1986 (4.49% en 1986, en hausse par rapport à 3.16% en 1981, pour l'ensemble de la population de la province) (Statistique Canada 1990). Statistique Canada a conclu à un accord avec la Planning and Statistics Division (l'organisme statistique provincial) du ministère relativement à un projet pour aider à réduire le sous-dénombrement en Colombie-Britannique lors du recensement de 1991. Le contrat portait sur deux domaines importants: premièrement, une collaboration avec la Division de la géographie pour que le fichier principal de région (FPR) couvre un territoire plus grand en Colombie-Britannique afin qu'il comprenne des centres urbains plus petits, augmentant ainsi la population qui figure dans ce fichier de 62% à 88% et, deuxièmement, la production du RA pour ces régions urbaines (Stewart 1991).

Le projet était un effort de collaboration avec les responsabilités et le travail tant pour le prolongement du FPR que pour la création du RA partagés entre les organismes. Par exemple, pour la création du RA, Statistique Canada a fourni à la Planning and Statistics Division du logiciel pour effectuer la normalisation et la fusion des dossiers administratifs ainsi que pour éliminer les données en double; c'est cette division qui était chargée d'acquérir les fichiers administratifs et de produire le RA jusqu'à l'étape du géocodage (sans inclure cette dernière étape). La division a effectué ces travaux tant pour les régions existantes du FPR que pour les régions agrandies.

3. SOURCES ET PRODUIT

La production du registre des adresses (RA) a commencé en avril 1990 et s'est terminée avec l'agrafage du dernier cahier à la mi-mai 1991. Nous avions alors compilé 22 756 cahiers renfermant 6.6 millions d'adresses pour les opérations de collecte des données du recensement.

3.1 Sources administratives

Suite à l'essai de septembre 1989, on a conclu que, dans la mesure du possible, l'on devrait utiliser les quatre sources administratives mentionnées ci-après comme sources d'adresses lors de la création du RA: fichiers de facturation des sociétés de téléphone, rôles d'évaluation des municipalités, fichiers de facturation des sociétés d'électricité et le fichier des déclarations d'impôt sur le revenu des particuliers (T1). Toutefois, ce n'est qu'en Nouvelle-Écosse, au Nouveau-Brunswick et dans huit grands centres urbains de l'Ontario (Ottawa, Toronto, Brampton, Etobicoke, London, Mississauga, Hamilton et Windsor) que l'on pouvait utiliser ces quatre sources de données. À cause de la multiplicité des fichiers, de leur coût et des refus, seulement trois sources ont été utilisées pour Terre-Neuve, pour le Québec, pour le Manitoba, pour l'Alberta (fichiers des sociétés de téléphone, des sociétés d'électricité et de l'impôt) ainsi que pour Regina et pour le reste de l'Ontario (fichiers des sociétés de téléphone, rôles d'évaluation et fichier de l'impôt). Pour Saskatoon, seuls les fichiers des sociétés de téléphone et de l'impôt sur le revenu étaient disponibles. Les principaux fichiers sources utilisés par le gouvernement de la Colombie-Britannique étaient ceux des sociétés de téléphone et des sociétés d'électricité, bien que l'on ait aussi employé les fichiers des immatriculations de véhicules, ceux des câblodistributeurs et les listes électorales (Stewart 1991).

3.2 Sources de données géographiques

Lors de la création du RA, nous avons largement utilisé certains produits de la Division de la géographie.

i. Le Fichier principal de région (FPR) (Statistique Canada 1988) est un réseau numérisé d'éléments (rues, voies ferrées, fleuves et rivières, etc.) pour de grandes et de moyennes régions urbaines, dont la population est généralement de 50 000 personnes ou plus. Nous étions intéressés par les données sur les rues qui comprenaient le nom de la rue et les gammes de numéros de voirie qui pouvaient être utilisés pour situer des adresses particulières sur un côté d'îlot, notre principal élément pour effectuer le couplage.

ii. Le système de cartographie assistée par ordinateur (CAO) groupe les côtés d'îlot en îlots et les îlots pour former un secteur de dénombrement (SD) du recensement. Le système de CAO a été utilisé pour classer les adresses dans les cahiers du RA. Les recenseurs qui ont travaillé au recensement de 1991 ont utilisé les cartes de SD produites par le système de CAO. Dans le cas du RA, les cartes pour toutes les régions couvertes par un FPR ont été utilisées lors de la deuxième opération effectuée par les employés de bureau.

iii. Le fichier de conversion des codes postaux (FCCP) de 1990 (Statistique Canada 1991) est un fichier qui contient tous les codes postaux au pays, chacun étant couplé à un SD ou à une série de SD du recensement de 1986. Ces données en entrée ont été utilisées pour effectuer le couplage secondaire des adresses au niveau du SD.

iv. Le fichier de correspondance entre SD de 1986/1991 établit le lien entre les éléments géographiques des SD de 1986 et ceux de 1991. Ce fichier a été utilisé pour effectuer le couplage secondaire au niveau du SD et pour la deuxième opération effectuée par les employés de bureau.

3.3 Cahiers du registre des adresses

Le produit final était un ensemble de cahiers d'adresses résidentielles, un pour chaque secteur de dénombrement, qui englobait toutes les régions urbaines du Canada pour lesquelles un fichier principal de région existait. La figure 1 renferme un spécimen (en format réduit) d'une page d'un cahier du RA.

Figure 1: Spécimen de page d'un cahier du RA (en format réduit)

ADDRESS REGISTER — Protected — PROVINCE 24 — EA - SD 261 — Page 21 of-de 22
REGISTRE DES ADRESSES — Protégé — FED - CÉF 038 — VN - NV 0

Block No. No d'îlot	Address - Adresse			Hhld No. No de ménage	Not Listed at Drop-off Non inscrit à la livraison	Field Follow-up Required Suivi sur place requis	Invalid - Non valide			AR Ref No. No de réf. du RA	Telephone Number Numéro de téléphone
	Civic No. No de voirie	Street Rue	Apt. No. No d'app.				Duplicate En double	Outside EA En dehors du SD	Other Autre		
1	2	3	4	5	6	7	8	9	10	11	12
4	23	PRINCIPALE RU								1044566	5551111
4	19	PRINCIPALE RU								1044564	5561234
4	15	PRINCIPALE RU								1044562	5552321
4	11	PRINCIPALE RU								1044559	
4	9	PRINCIPALE RU								1044583	7475739
4	7	PRINCIPALE RU								1044581	5552222
5	30	CENTRE BV								1019615	5561029
5	34	CENTRE BV								1019617	
5	34	CENTRE BV	S-SOL							1019618	5564261
5	60	CENTRE BV								1019627	
5	64	CENTRE BV								1019629	7478765
5	68	CENTRE BV								1019634	5556942
5	72	CENTRE BV								1019636	
5	76	CENTRE BV								1019640	
5	80	CENTRE BV								1019642	7476789
5	84	CENTRE BV								1019644	5568765
5	88	CENTRE BV								1019646	5559999
5	92	CENTRE BV								1019579	7473456
5	96	CENTRE BV								1019581	7450987
5	100	CENTRE BV								1019648	
5	108	CENTRE BV								1019579	5557171
5	112	CENTRE BV								1019581	5558888
5	116	CENTRE BV								1019583	7462009
5	120	CENTRE BV								1019586	7450235
5	124	CENTRE BV								1019588	5569630

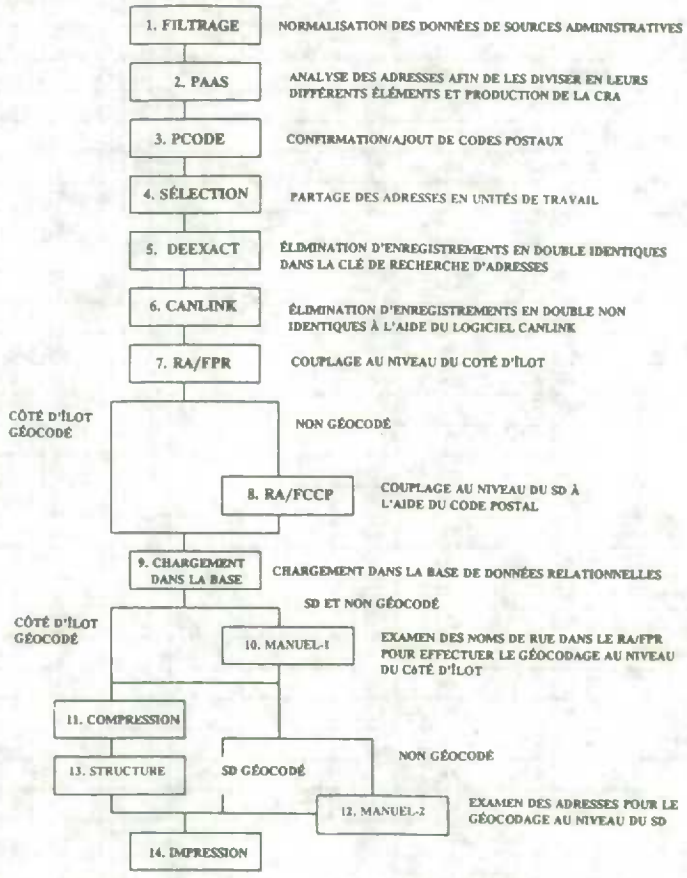
Chaque cahier était divisé en deux sections: une partie structurée et une partie non structurée. La partie structurée renfermait toutes les adresses liées à un côté d'îlot et tous les côtés d'îlot y étaient classés en îlots dans le SD. Le classement correspondait aux renseignements qui figuraient sur la carte utilisée par le recenseur pour dresser la liste du SD dans son Registre des visites (RV). La partie non structurée contenait les adresses qui ne pouvaient être liées qu'au SD plutôt qu'à un côté d'îlot. Ces adresses étaient classées par numéros de voirie impairs/pairs pour un même nom de rue. Le nombre d'adresses était réparti dans la proportion suivante: 90%-10% entre les données structurées et les données non structurées.

En plus des données sur les adresses, chaque page d'un cahier de RA comprenait une série de colonnes à utiliser lors de l'opération de conciliation entre le RA et le RV. Au cours de la conciliation, le recenseur comparait manuellement le Registre des visites au RA afin de déterminer où il y avait correspondance et où il n'y en avait pas. Si l'adresse n'était que dans le RV, elle était ajoutée au RA (sous-dénombrement dans le RA). Si l'adresse ne figurait que sur le RA, le recenseur devait habituellement régler le problème sur le terrain. Cette adresse était donc désignée soit comme une nouvelle adresse que le recenseur devait dénombrer dans le cadre du recensement (sous-dénombrement au niveau du recensement), soit comme une adresse invalide classée selon le genre d'erreur (surdénombrement dans le RA). Les adresses étaient considérées invalides s'il s'agissait d'adresses en double, si elles se trouvaient à l'extérieur du SD ou pour toute autre raison. Pour toutes les adresses valables le recenseur inscrivait le numéro de ménage du recensement dans le cahier. Un numéro de téléphone correspondant à l'adresse, s'il était disponible, était imprimé à l'avance dans la dernière colonne du cahier afin que le recenseur puisse l'utiliser, au besoin, pour le travail de suivi du recensement.

4. MÉTHODOLOGIE

Dans la présente section, nous décrivons la création du registre des adresses (RA). La figure 2 donne un aperçu des étapes que comprend cette opération.

FIGURE 2 : ENCHAÎNEMENT DES OPÉRATIONS



4.1 Aperçu de la méthodologie

On a tout d'abord effectué la normalisation des éléments constituants des adresses à structure non imposée contenues dans les fichiers sources (étapes 1 et 2) en vue de leur utilisation par le logiciel qui effectue les étapes ultérieures. Puis, les codes postaux ont été confirmés ou corrigés (étape 3) afin que les régions pour lesquelles le RA allait être créé puissent être choisies à partir de toutes les adresses et emplacements contenus dans les fichiers sources (étape 4). Parce que la même adresse pouvant être contenue dans plus d'un fichier ou plus d'une fois dans le même fichier, on a effectué la suppression des adresses en double en se basant sur des appariements tant exacts que probabilistes (étapes 5 et 6).

Ensuite, on a couplé, de façon automatisée, les adresses au côté d'îlot à l'aide du Fichier principal de région (étape 7) ou au secteur de dénombrement (SD) à l'aide du Fichier de conversion des codes postaux (étape 8). Après avoir chargé les adresses dans un système de gestion de base de données (étape 9), des couplages manuels ont été effectués entre les adresses et les côtés d'îlot (étapes 10 et 11) ou entre les adresses et les SD (étape 12). Les adresses dans chaque SD ont ensuite été classées par côté d'îlot et à l'intérieur des côtés d'îlot (étape 13) avant d'être imprimées et assemblées en cahiers (étape 14) pour utilisation dans le cadre du recensement.

4.2 Normalisation des adresses (étapes 1, 2 et 3)

Le système d'analyse des adresses postales (PAAS - étape 2 de la figure 2) (Statistique Canada 1989c) remplissait deux tâches: tout d'abord il séparait les adresses à structure non imposée provenant des fichiers sources en leurs éléments constituants (nom de rue, numéro de voirie, genre de rue, sens de la rue, numéro d'appartement, municipalité, province, code postal) et composait la clé de recherche d'adresses (CRA). La CRA est un enchaînement ordonné de tous les éléments qui composent une adresse et elle est utilisée pendant les opérations visant à éliminer les adresses en double.

Bien que le PAAS fut un excellent produit, l'analyse des résultats du prototype de 1989 avait révélé certains défauts qui, selon nous, pourraient être réglés en filtrant le contenu des fichiers administratifs avant d'utiliser le logiciel général. Cette étape de FILTRAGE (étape 1) portait sur les tâches suivantes: élimination des caractères spéciaux que le PAAS refusait de traiter, regroupement des éléments de l'adresse afin de la rendre compatible avec le PAAS, traduction des indications abrégées de genre de rue en genres acceptables, introduction de virgules entre les éléments de la rue et de la municipalité des adresses à structure non imposée afin que le PAAS puisse mieux les comprendre, élimination des zéros en tête dans les numéros de voirie et les noms d'adresses numériques et ajout du nom de la municipalité et de la province.

Les étapes de FILTRAGE et PAAS travaillaient ensemble de façon itérative. Tout d'abord, nous avons évalué quelles anomalies devaient être filtrées pour chaque source de données administratives. Si le taux d'erreur du PAAS après le filtrage était supérieur à 5%, nous examinions les enregistrements erronés afin de chercher des problèmes qui se répétaient en vue de les éliminer par un filtrage additionnel jusqu'à ce qu'un taux d'erreur inférieur à 5% soit atteint. Comme tout enregistrement d'adresses qui était rejeté lors de la normalisation des adresses était éliminé du système, il était essentiel que le taux de réussite du PAAS soit le plus élevé possible.

L'étape PCODE (étape 3) utilisait le progiciel du système automatisé d'établissement des codes postaux (PCODE) (Statistique Canada 1989a) pour confirmer et pour produire les codes postaux. Ce logiciel n'était pas tout à fait aussi efficace que le logiciel PAAS pour analyser les adresses et il ne pouvait confirmer ou ajouter des codes postaux que pour 84% des sorties du PAAS. Ce progiciel a confirmé 78% des codes postaux et il en a modifié un autre 6%. Seulement 0.003% des enregistrements administratifs d'origine avaient été fournis sans code postal. Il était essentiel de disposer des codes postaux exacts parce que ces derniers allaient être utilisés pour le choix de l'unité de travail au cours de l'étape suivante.

Deux problèmes liés au temps de traitement ont surgi lors de l'étape PCODE au moment de la production du RA. S'il n'y avait pas d'élément municipalité ou province pour une adresse, le logiciel continuait de tenter de trouver un code postal plutôt que de suspendre le traitement. Par conséquent, des temps de traitement considérables pouvaient être passés à essayer de trouver des codes postaux. Pour empêcher que cela ne se produise, une tâche additionnelle a été incluse dans l'étape de FILTRAGE pour ajouter le nom de la municipalité et celui de la province. Le deuxième problème était dû au fait que lorsqu'un nom de rue était numérique, le temps de traitement par adresse quadruplait. Pour corriger ce problème, il faudra modifier le logiciel PCODE.

4.3 Choix de l'unité de travail (étape 4)

Au cours de cette étape le pays a été divisé, au moyen des codes postaux, en unités de travail administrables, dont la taille était basée sur l'efficience du logiciel CANLINK pour coupler de nombreux gros fichiers. Cela nous a amenés à adopter des unités de travail géographiques comprenant entre 100 000 et 150 000 logements d'après les données du recensement de 1986. Les unités de travail étaient formés à partir d'un seul FPR (pour une ville de taille moyenne), à partir d'ensembles de FPR adjacents (pour de petites villes/cantons), ou à partir de parties d'un FPR (pour une grande ville). Quand un FPR était divisé, cela se faisait en fonction des trois premiers caractères du code postal (la région de tri d'acheminement ou RTA) et l'on suivait généralement les éléments physiques (p. ex., un fleuve ou une rivière) dans la région urbaine. Le Fichier de conversion des codes postaux (FCCP) de la Division de la géographie qui couple les codes postaux avec des éléments détaillés de géographie du recensement a été utilisé pour effectuer ce partage lors de l'étape de SÉLECTION (étape 4). Une fois le partage terminé, il y avait 105 unités de travail distincts et le nombre d'adresses originales avait été

ramené de 43.4 millions à 20.5 millions, les adresses éliminées ayant des codes postaux à l'extérieur des régions couvertes par les FPR (c.-à-d. dans les plus petites villes et dans les régions rurales).

4.4 Élimination des adresses en double (étapes 5 et 6)

Afin de supprimer les adresses qui figuraient plus d'une fois dans les fichiers sources, une opération d'élimination d'adresses en double a été effectuée en deux étapes: un appariement exact avec DEEXACT (étape 5) et un appariement probabiliste à l'aide de CANLINK (étape 6).

Au cours de l'étape DEEXACT on utilisait la clé de recherche d'adresses (CRA) produite par le logiciel PAAS et tous les enregistrements avec une CRA identique ont été comprimés en un seul enregistrement. Avec DEEXACT, les 20.5 millions d'enregistrements dont on disposait après l'étape de SÉLECTION ont été ramenés à 10.1 millions d'enregistrements. Ces chiffres font ressortir l'importance d'effectuer la normalisation des adresses.

Dans l'étape 6 on utilisait le logiciel général de couplage d'enregistrements CANLINK (Statistique Canada 1989b) qui avait été utilisé au cours de l'essai. Ce logiciel stratifie les enregistrements en groupes appelés "poches" et seuls les enregistrements dans la même poche sont effectivement appariés. Pour cette application, c'est le numéro de voirie qui a été utilisé comme poche. Les éléments constituants de l'adresse (nom de la rue, nom de la municipalité, code postal, etc.) ont été utilisés à des fins d'appariement et des poids ont été attribués quand il y a accord ou désaccord pour chaque élément. L'élaboration de niveaux d'accord partiel pour le nom de la rue, pour le nom de la municipalité et pour l'unité de distribution locale (les trois derniers caractères du code postal) tenait compte des variations orthographiques et des transpositions de lettres dans les champs. L'étape CANLINK a entraîné une autre réduction du nombre d'enregistrements, il n'y en avait que 6.7 millions après cette étape. On trouve plus de détails sur l'utilisation de CANLINK pour l'élimination des adresses en double dans Drew et coll. (1988), où l'on décrit l'application de ce logiciel dans le cadre de l'essai de novembre 1987.

4.5 Couplage RA/FPR (étape 7)

La stratégie utilisée pour coupler les adresses à leur côté d'îlot respectif constituait la principale préoccupation après l'essai de 1989. À cause de la diminution de 11% dans la couverture qui est passée de 84% à 73% comparativement à l'essai de 1987, il fallait effectuer une étude complète et peut-être employer une nouvelle approche. L'autre préoccupation évidente en matière de procédure était le fait que l'appariement automatisé n'avait permis d'obtenir que 80% des enregistrements appariés alors que le reste (20%) de l'appariement était attribuable à du travail de bureau. Cette situation aurait représenté une charge importante de travail manuel quand la production du RA battrait son plein. En 1989, le processus automatisé ne permettait qu'un appariement exact du nom de la rue/du genre de rue/du sens de la rue entre un enregistrement du RA et un enregistrement du fichier principal de région (FPR). Tout appariement inexact était effectué à la main. Afin de surmonter ces deux préoccupations, on a élaboré une autre application de CANLINK pour le couplage RA/FPR (étape 7) à l'aide de techniques d'appariement probabilistes.

Les fichiers originaux utilisés dans le cadre de l'essai de 1989 pour Brampton existaient encore, c'est donc sur cette localité qu'a porté l'essai visant à élaborer cette étape. La méthode révisée a donné 10% de plus d'appariements, ce qui a ramené la couverture aux niveaux de 1987. De plus, l'appariement automatisé permettait de réaliser 97% des appariements, 3% étant effectués par des employés de bureau, une amélioration importante par rapport à la répartition antérieure de 80%-20%. Suite à ces résultats, la méthode CANLINK a été adoptée pour la réalisation des travaux pour le recensement.

Lors de l'élaboration de la nouvelle stratégie d'appariement, le premier domaine d'étude comprenait une comparaison du contenu des champs qui seraient utilisés pour effectuer l'appariement. Cette opération a révélé certaines anomalies qui pouvaient être corrigées avant l'utilisation des fichiers afin d'améliorer le nombre de couplages. Les modifications du traitement des champs existants portaient sur les domaines suivants: suppression des espaces entre les noms de rue composés; alignement des sens des rues et des numéros de voirie; conversion des noms de rue numériques en chiffres (dans le FPR); suppression des caractères spéciaux dans les noms de rue (dans le FPR); correction des variantes orthographiques dans les noms de municipalité (dans le RA) et une

reconstitution de certaines traductions effectuées par le PAAS pour des noms de rue (dans le RA). Plusieurs nouveaux champs ont aussi été créés: des versions NYSIIS (New York State Identification and Intelligence System) et SOUNDEX du nom de rue, à l'aide de deux progiciels de codage phonétique utilisés pour éliminer les effets des erreurs d'orthographe courantes (Statistique Canada 1989d); un drapeau de nom de rue en double (dans le FPR) pour signaler les situations où un nom de rue n'est pas unique; un drapeau de rue unidirectionnelle (dans le FPR) pour signaler les rues pour lesquelles un seul sens de rue avait été codé et un drapeau de nom de rue officiel (dans le RA) pour signaler que le nom de la rue correspondait à un nom officiel de rue dans le FPR. Les enregistrements du FPR ne contenaient que des données sur les rues, nous leur avons donc ajouté le nom de la subdivision de recensement ainsi qu'un code de province puis nous avons tenté d'attribuer des codes postaux aux numéros de voirie dans des côtés d'îlot. Quand les codes postaux pour les numéros de voirie "de départ" et "d'arrivée" différaient, nous avons produit des sous-côtés d'îlot pour chaque code postal unique.

Pour cette application, trois poches distinctes ont été créées pour chaque enregistrement produisant, effectivement, trois exemplaires des fichiers. La poche primaire était celle à laquelle s'appliquaient les conditions les plus rigoureuses et elle était conçue pour trouver rapidement toutes les possibilités d'un bon appariement lors du premier passage des fichiers. Cette poche était composée du nom de la rue/de la région de tri d'acheminement (RTA)/du drapeau de numéros de voirie impairs ou pairs. La deuxième poche comprenait le code postal/le drapeau de numéros de voirie impairs ou pairs, ce qui permettait d'apparier d'après le code postal les adresses mal analysées. La troisième poche comprenait la version NYSIIS du nom de la rue/du drapeau de numéros de voirie impairs ou pairs, ce qui permettait d'apparier éventuellement des enregistrements renfermant des variations orthographiques du nom de la rue ou qui n'ont pas de code postal.

Les règles de fonction établies dans le cas des appariements partiels pour le nom de la rue, pour le nom de la municipalité et pour l'unité de distribution locale (les trois derniers caractères du code postal) ont été tirées directement de notre application CANLINK existante utilisée pour l'élimination interne des données en double, où elles avaient déjà démontré leur efficacité.

Une option du logiciel CANLINK permet d'introduire des éléments de programme adaptés à l'application à différents points clés du processus de couplage. Un de ces points clés, l'étape de "sélection-appariement", permet d'exclure des couplages non désirés avant que les enregistrements soient appariés, réduisant ainsi les coûts du traitement et augmentant la qualité des couplages. Ces couplages non désirés sont ceux dont la pondération est assez élevée pour qu'il y ait appariement mais qui, à cause de circonstances particulières, ne sont pas réellement des appariements valables et qui devraient être exclus de la comparaison. Voici certains cas qui ont été exclus lors du processus de production: quand le numéro de voirie dans l'enregistrement du RA tombait à l'extérieur de la gamme de numéros de voirie dans l'enregistrement du FPR; quand un nom de rue dans le FPR n'était pas unique et quand les indicateurs de genre de rue étaient différents ou n'existaient pas dans l'enregistrement du RA; quand les noms de rue dans le RA et dans le FPR étaient différents et que le RA renfermait un nom de rue officiel du FPR.

Au cours de l'étape de la production du RA, nous avons éprouvé de la difficulté à apparier les enregistrements contenus dans trois FPR: Red Deer, St. Thomas et Charny. Dans les trois cas, c'est le manque de données sur les numéros de voirie dans le FPR qui a créé les difficultés. Sachant qu'il faudrait effectuer beaucoup de travail de bureau pour éliminer ces difficultés, une opération sur le terrain a été lancée en décembre 1990 afin de mettre à jour les cartes produites par CAO. La Division de la géographie a envoyé les cartes produites par CAO aux bureaux régionaux où des employés ont ajouté les gammes de numéros de voirie manquants. Ces cartes mises à jour ont ensuite été retournées à la Division de la géographie pour être incluses dans la prochaine série de mises à jour des FPR. Pour la création du RA, les gammes de numéros de voirie pour les trois FPR ont été utilisées dans des opérations manuelles effectuées par des employés de bureau.

Au niveau de l'appariement, le succès était fort semblable dans toutes les provinces sauf le Québec. Au Québec, l'appariement automatique au côté d'îlot a diminué d'environ 10 à 12% pour atteindre 73%, puisque cette opération n'était pas aussi efficace lors du traitement des adresses en français qu'elle ne l'était dans le cas des adresses en anglais. On a trouvé trois situations qui causaient la diminution du taux d'appariement automatique: l'utilisation ou la non utilisation d'articles dans les noms de rue (p. ex., Savane, de la Savane, la Savane), l'utilisation d'un nom complet de personne comme nom de rue avec beaucoup de variantes orthographiques

(p. ex., Jean-François Bélanger, J.F. Bélanger, Jean F. Bélanger) et le manque d'indicateurs de genre de rue. C'est pourquoi, les opérations de bureau décrites ci-dessous, particulièrement la première, avaient une importance accrue pour l'appariement au Québec par rapport à la situation dans les autres provinces.

Lors du traitement du RA/FPR à l'aide du logiciel CANLINK, il n'y a eu qu'une difficulté soit le fait que le nombre maximum d'enregistrements permis dans une poche interne à CANLINK pouvait être dépassé. La solution adoptée consistait à déterminer les rues qui étaient la source de la difficulté à partir du rapport sur la poche (il s'agissait toujours d'artères principales) et à préparer des programmes spéciaux de pré-traitement qui ajouteraient le cinquième caractère du code postal lors du calcul de la valeur de la poche pour ces rues afin qu'une distinction puisse être établie plus facilement. Cela avait pour effet de réduire le nombre d'enregistrements dans la poche.

4.6 Couplage RA/FCCP (étape 8)

Dans cette étape (étape 8) on tentait d'obtenir un couplage automatisé avec le secteur de dénombrement (SD) approprié pour les adresses qui n'avaient pu être appariées, à l'aide du FPR, aux côtés d'îlot pendant l'étape 7.

Les principales données en entrée étaient le Fichier de conversion des codes postaux (FCCP), qui donnait la correspondance entre les codes postaux et les SD de 1986 et le fichier de correspondance entre SD de 1986 et de 1991. En faisant l'appariement de ces deux fichiers nous pouvions déterminer les codes postaux qui ne correspondaient qu'à un seul SD de 1991, ainsi que les codes postaux qui correspondaient à au moins deux SD possibles pour 1991 et dont le cas allait être réglé, par des opérations manuelles, au cours de l'étape 12.

À nouveau, c'est Brampton qui a été utilisé pour l'essai. L'analyse de l'appariement code postal/SD a permis d'établir que 38% des codes postaux pouvaient être attribués, de façon unique, à un SD de 1991. Le couplage de ces codes postaux à des enregistrements du RA qui n'étaient pas appariés à un côté d'îlot a permis d'augmenter d'un autre 5% le nombre total d'appariements. Dans l'ensemble, le taux des appariements automatisés a augmenté pour atteindre 89% (84% avec le côté d'îlot et 5% avec le SD), en hausse par rapport à 64% lors de l'essai de septembre 1989, ce qui réduit de près de la moitié l'importance des opérations manuelles à effectuer.

4.7 Chargement dans la base (étape 9)

Lors de l'essai de 1989, pour faciliter les consultations et en prévision d'une utilisation ultérieure, nous avions utilisé ORACLE comme système de gestion de base de données et nous l'avons employé à nouveau pour la production du RA en 1991. L'étape du chargement des données dans ORACLE (étape 9) comprenait la transformation du fichier jusqu'ici séquentiel en quatre fichiers de composantes distincts, un pour chacun des éléments suivants: municipalité, côté d'îlot, rue et adresse.

4.8 Travaux de bureau (étapes 10, 11 et 12)

Lors de l'essai de 1989, le travail de bureau consistait à examiner toutes les combinaisons uniques de nom de rue/d'indicateur de genre de rue/de sens de rue tirés des enregistrements tant du FPR que du RA avec un relevé du nombre d'enregistrements dans le RA pour chaque combinaison de rues. L'objectif de ce travail était de remplacer une combinaison de rues du RA pour laquelle il n'y avait pas d'appariement par la combinaison valable du FPR. La comparaison de combinaisons de rues semblables et la détermination de celles qui auraient, en fait, dû être identiques, permettait d'apparier manuellement des enregistrements du RA non codés jusqu'à ce moment avec un côté d'îlot particulier. Cette procédure avait donné de bons résultats en 1989 et s'était révélée utile pour régler deux situations qui posaient des difficultés: les cas où il y avait des écarts considérables dans l'orthographe du nom de la rue et ceux où le champ du nom de la rue dans le RA contenait à la fois le nom de la rue et une forme abrégée d'indicateur de genre de rue que le logiciel PAAS n'avait pas compris lors de l'analyse de l'adresse.

Nous avons accru la portée de ce travail de bureau (étape 10) afin de comparer des combinaisons de rues du RA avec d'autres combinaisons de rues du RA semblables pour traiter les cas où il se pourrait que l'on trouve

un certain nombre de variations orthographiques du nom d'une rue particulière dans le RA sans équivalent dans le FPR.

Suite à cet accroissement de la portée du premier travail de bureau (Manuel-1), nous avons ajouté une étape de compression (étape 11). Pour chaque valeur unique de nom de rue/indicateur de genre de rue/sens de rue pour une unité de travail, nous avons vérifié tous les enregistrements d'adresse correspondants afin de nous assurer qu'ils sont uniques en utilisant le numéro de voirie/le numéro d'appartement comme clé. Quand des enregistrements multiples étaient rencontrés, ils étaient regroupés avec toutes les données pertinentes combinées en un seul enregistrement.

Au cours de l'étape 12, nous traitions les adresses restantes qui n'avaient pu être appariées avec un seul SD, mais qui pouvaient l'être avec au moins deux SD (les SD candidats à un appariement) au cours de l'étape 8. Un ensemble complet de cartes conçues par cartographie assistée par ordinateur (CAO) a été produit pour le projet du RA. Au cours de l'étape Manuel-2, on examinait ces cartes pour trouver ces SD candidats afin d'affecter ces adresses restantes au SD approprié, dans la mesure du possible.

En général, le rapport entre l'appariement automatisé et l'appariement manuel était de 91% à 9%. La partie automatisée se répartissait de la façon suivante: 87% provenant du couplage RA/FPR au côté d'îlot et 4% du couplage RA/FCCP au SD. Pour le traitement manuel, la répartition des appariements était la suivante: 3% provenant du couplage avec le côté d'îlot au cours de l'opération Manuel-1 et 6% du couplage avec le SD au cours de l'opération Manuel-2.

Bien que ORACLE ait constitué un véhicule approprié pour l'essai de 1989, son utilisation s'est révélée coûteuse et il a éventuellement constitué un goulot d'étranglement quand la production du RA battait son plein alors que ce dernier n'était qu'un des utilisateurs de cette base de données employée par tous les services du Bureau. ORACLE ne permettait d'avoir que de 8 à 10% des unités de travail accessibles en direct à un moment quelconque et le système devait exporter et importer continuellement des données sur les unités de travail afin de libérer de l'espace pour continuer le traitement. Nous avons donc constitué une seconde base de données ORACLE à l'usage exclusif de l'équipe travaillant au RA. En toute justice pour ORACLE, ce n'est pas tout le traitement à effectuer qui donnait des conditions propices à quelque système de gestion de base de données que ce soit. Nous étions en train de créer notre produit et, par conséquent, nous consultions de grandes parties des tables pour apporter des modifications radicales à des champs, pour éliminer des enregistrements en double et pour choisir des enregistrements à imprimer. ORACLE offrait une flexibilité considérable pour changer les procédures logicielles rapidement et pour en produire de nouvelles à mesure que l'étape de production se déroulait.

4.9 Utilisation du système de cartographie assistée par ordinateur (étape 13)

Le système de cartographie assistée par ordinateur (CAO) était une nouvelle initiative de recherche pour le recensement de 1991 dont l'élaboration s'est faite concurremment avec celle du RA. Le système a produit toutes les cartes de secteurs de dénombrement dans les régions pour lesquelles il existe des FPR. Cela constituait un changement considérable par rapport au processus manuel de production de cartes utilisé auparavant. La CAO a aussi fourni, pour les SD, une structure dans laquelle les côtés d'îlot étaient situés dans les îlots et les îlots classés dans le SD (étape 13). Aux fins du RA, on a préparé un programme qui dérive de la CAO pour classer les logements dans un côté d'îlot. Cette opération était nécessaire pour structurer les listes d'adresses afin qu'elles correspondent plus étroitement à la façon dont les recenseurs dressent leurs listes.

Le système de CAO n'était pas disponible pour l'essai de 1989, on a donc produit les cartes utilisées pour l'essai en les faisant tracer par ordinateur, un SD à la fois, afin de simuler le produit fourni par le système de CAO. Les données sur la structure qui servaient à classer les côtés d'îlot dans le SD et les adresses dans les côtés d'îlot ont été produites à l'aide d'un logiciel élaboré spécialement à cette fin (Schut et Haythornthwaite 1990). Les résultats de l'essai de 1989 ont fait ressortir certains domaines qui causaient des inquiétudes. Premièrement, les cartes de SD seraient plus utiles si elles fournissaient des renseignements sur les SD environnants. Deuxièmement, il y a eu des problèmes à propos des données sur la structure: codage des côtés d'îlot au SD approprié quand un SD était contenu dans un autre SD; affectation des numéros d'ordre aux côtés d'îlot dans

les rues en forme d'hameçon (impasses); données manquantes sur les côtés d'îlot et données en double sur les côtés d'îlot. Par la suite, lors de l'élaboration du système de CAO, on a réglé ces problèmes.

Le système de CAO était entièrement mis en place au moment de la production du RA. Afin d'assurer la compatibilité avec ce logiciel, nous avons utilisé la version du FPR qui était employée par le système de CAO. Toutefois, aucune donnée sur la structure n'était affectée à une petite portion des côtés d'îlot. Pour tout SD où ce pourcentage était supérieur à 5%, nous reprenions le traitement effectué par le système de CAO pour cette unité de travail, si le temps le permettait, ou nous utilisions un autre système, celui qui emploie l'algorithme d'affectation des points dans un polygone (Point in Polygon Assignments (PIPA)), qui situe les côtés d'îlot dans leur SD. Bien que le système PIPA faisait passer les adresses de la partie structurée du cahier du RA (basée sur le codage des côtés d'îlot) à la partie non structurée (basée sur le codage des SD), les adresses posant des problèmes lors du processus de sélection pour l'impression pouvaient au moins être conservées, ce qui n'était pas le cas quand les données sur l'ordonnancement manquaient.

4.10 Impression et production des cahiers (étape 14)

Les essais antérieurs ne donnaient aucun renseignement sur le volume d'impression et sur la production des cahiers (étape 14) pour les environ 23 000 secteurs de dénombrement renfermant, à ce moment, 6.6 millions d'adresses. Les principales inquiétudes portaient sur la rapidité et la qualité de l'impression, sur la résistance des cahiers et sur les coûts de la compilation.

Avec le système prototype nous avions utilisé une imprimante à feuilles séparées qui était près de 10 fois plus lente qu'une imprimante à papier en continu. Nous avons donc adapté la procédure d'impression pour l'imprimante plus rapide après nous être assurés que la qualité de l'impression était entièrement satisfaisante. L'éclatement était une autre opération prenant beaucoup de temps. Nous avons décidé d'éliminer cette opération et de laisser les listes imprimées sous forme de feuilles à pliage paravent.

Pour que les cahiers puissent durer plus longtemps lors des opérations sur le terrain, des couvertures avant et arrière ont été produites. La couverture avant comprenait une fenêtre qui permettait au personnel du recensement de lire le code de la province, le numéro de la circonscription électorale fédérale, le numéro du SD ainsi que le numéro du district de commissaire au recensement afin de faciliter l'organisation et la distribution pendant les opérations de collecte des données du recensement. L'autre inquiétude en matière de résistance était le genre de reliure. Après étude, on a trouvé que la reliure collée était trop fragile et qu'elle ne pouvait supporter l'ouverture et la fermeture constantes qui étaient nécessaires. Nous avons donc choisi d'utiliser des agrafes.

Plusieurs options se présentaient pour la compilation des cahiers: sous-traiter le travail, faire assembler les cahiers par l'imprimerie ou assembler les feuillets et les agrafer nous-mêmes. À la fin, c'est la dernière option qui a été choisie pour des raisons de coût et de délai. Afin de fournir un cahier dont la qualité serait acceptable au personnel de la Division des opérations des enquêtes travaillant sur le terrain, nous avons conçu et fait construire un bâti d'agrafage en bois auquel pouvait être adaptées trois agrafeuses avec des guides de document. Nous étions alors en mesure de garantir la largeur des marges et que les agrafes seraient bien placées le long du dos du cahier. Le prototype a fonctionné à merveille. Suite aux estimations du temps requis, quatre postes d'agrafage ont été construits pour répondre aux besoins de production. À cause du volume d'impression à traiter, nous avons établi avec le Centre principal des ordinateurs des calendriers d'impression et de traitement différé en entrée/sortie afin de minimiser l'incidence de nos travaux sur le traitement effectué par d'autres services du Bureau.

5. ÉVALUATION POSTCENSITAIRE

L'évaluation postcensitaire peut être classée, en gros, en quatre domaines d'étude: opérations sur le terrain, saisie des données sur les cahiers du RA, mise à jour du RA et détermination de l'apport du RA aux améliorations de la couverture.

L'évaluation des opérations sur le terrain se concentrera sur l'efficacité de la formation, sur jusqu'à quel point le travail de conciliation a été complet et sur les causes des erreurs en vue d'améliorer la méthodologie pour les recensements à venir.

La saisie des données donnera deux résultats distincts. Premièrement, les adresses imprimées dans les cahiers seront supprimées si elles sont erronées et si elles sont valides, leur numéro de ménage du recensement sera saisi. Deuxièmement, les nouvelles adresses ajoutées par les recenseurs seront saisies. Il sera alors possible de calculer les taux de surdénombrement et de sous-dénombrement du RA ainsi que la contribution du RA à la couverture du recensement. On pourra examiner le cas des adresses classées dans le mauvais SD et remonter à la source de l'erreur. Le numéro de ménage du recensement nous permet d'étudier le nombre de personnes ajoutées ainsi que les caractéristiques des logements et des personnes.

Du point de vue coût, on calculera le coût unitaire par logement ajouté à cause de l'utilisation du RA en vue de déterminer le coût de création du RA et de son utilisation dans le cadre du recensement.

6. ORIENTATIONS FUTURES

Le registre des adresses (RA), bien qu'il ait été conçu à l'origine comme une des procédures visant à réduire le sous-dénombrement du recensement, est un projet évolutif avec un effet possible sur d'autres programmes de Statistique Canada ainsi que sur ceux d'autres organismes gouvernementaux.

Voici les objectifs plus immédiats pour le développement futur du RA, selon les directives de la haute direction: incorporer les adresses relevées pendant le dénombrement du recensement; évaluer l'efficacité du RA pour ce qui est de l'amélioration de la couverture du recensement de 1991; consigner par écrit et évaluer les activités de production et élaborer un plan à plus long terme pour le RA s'attaquant à son rapport coût-efficacité comme base de sondage de ménages, à la stratégie optimale de mise à jour et aux possibilités qu'offre son emploi par des organismes de l'extérieur du Bureau.

Compte tenu de ces lignes directrices, on a préparé un plan de projet qui est présenté ci-après sous six sujets principaux.

6.1 Rapports entre le recensement et le registre des adresses

En plus des possibilités qu'il offre pour améliorer la couverture, nous étudierons d'autres façons qui permettraient de mettre le RA à contribution pour le recensement. Certaines idées préliminaires à ce sujet comprennent la possibilité d'utiliser le RA comme fichier de contrôle du traitement, pour obtenir les numéros de téléphone qui seront employés lors d'un suivi éventuel, pour créer des chiffres de contrôle des logements dans un secteur de dénombrement, pour certifier les totaux de logements pour le traitement ou pour des analyses de la migration. On étudiera si le RA devrait être utilisé avant ou après le jour du recensement et comment il pourrait être employé dans le cas des adresses pour lesquelles on ne peut établir des éléments géographiques qu'à un niveau supérieur au SD.

6.2 Rapports entre la Division de la géographie et le registre des adresses

Comme cela est manifeste dans la description de la méthodologie, lors de la création du RA on s'est basé beaucoup sur de nombreux produits de la Division de la géographie (p. ex., le Fichier principal de région, le Fichier de conversion des codes postaux). Nous examinerons la contribution de ces produits à l'établissement du RA ainsi que leurs limitations. On étudiera, pour tous les nouveaux produits élaborés par la Division de la géographie, l'utilisation possible de ces produits dans le RA en vue d'incorporer les besoins du RA directement dans les nouveaux produits. De plus, le RA sera intégré dans le système d'information géographique (SIG) de la Division de la géographie.

Il se peut que le RA puisse fournir des indicateurs de mise à jour pour le fichier principal de région (FPR) ou pour la délimitation des secteurs de dénombrement. Le RA pourrait être utilisé pour établir des ordres de priorité, particulièrement dans les régions à forte croissance ou dans les régions où les gammes de numéros de

voirie dans les FPR sont de qualité médiocre. On pourrait utiliser des combinaisons de code postal/secteur de dénombrement ou de code postal/côté d'îlot provenant du RA pour effectuer la mise à jour du Fichier de conversion des codes postaux. Après chaque recensement, tous les ménages du recensement sont codés avec des centroïdes de côté d'îlot. Puisque la majeure partie des enregistrements du RA ont déjà été géocodés avant le recensement, le fait de coupler le RA avec le numéro de ménage du recensement réduira l'importance du travail de géocodage à effectuer manuellement après le recensement. Ce dernier projet est déjà en cours de réalisation.

6.3 Évaluation et amélioration des procédures et production de la documentation connexe

On est en train de préparer, pour le travail effectué jusqu'ici, un guide de l'utilisateur où l'on décrit les procédures et un guide technique qui renferme de la documentation relative aux programmes, à des problèmes types ainsi qu'à leurs solutions et à l'assurance de la qualité.

Comme dans le cas de tout nouveau projet, beaucoup de choses sont apprises pendant le processus de création et les procédures sont élaborées au besoin et en fonction du temps ainsi que du budget disponibles. Après le fait, on peut habituellement réaliser des gains d'efficience en examinant ces procédures.

Pour les procédures automatisées, les projets déjà en cours comprennent une utilisation plus efficace de ORACLE ou le choix d'un autre système, l'emploi d'ordinateurs de bureau plutôt que du gros ordinateur de Statistique Canada, la normalisation du filtre, des améliorations au PAAS, la fusion des unités de travail en bases de données provinciales, l'élimination de certains champs plus tôt dans le traitement, l'étude d'autres logiciels pour déterminer les codes postaux, l'amélioration de l'appariement d'une adresse avec un nom de lieu et l'amélioration du couplage entre le Fichier principal de région et les adresses en français.

Pour les procédures manuelles, on continuera d'améliorer le traitement des secteurs de dénombrement adjacents de chaque côté des limites de circonscriptions électorales fédérales et les cas où des numéros de voirie manquent sur les cartes produites par CAO. Le système de contrôle pour corriger les adresses sera aussi examiné afin de l'améliorer si cela est possible.

Les numéros de téléphone ont été ajoutés à une étape ultérieure de la production du RA. Une évaluation complète de la couverture qu'ils offrent et de leur exactitude sera entreprise particulièrement en vue des utilisations possibles des numéros de téléphone dans le cadre du recensement et d'autres enquêtes de Statistique Canada. Pour ces dernières, l'accent initial sera placé sur la réalisation d'essais dans le contexte du remaniement imminent de l'enquête sur la population active.

Dans une large mesure, on a déjà épuré les systèmes informatiques élaborés pour la production initiale du RA afin d'augmenter l'efficience de l'utilisation du gros ordinateur, des programmes, du stockage sur disques et sur bandes, de la manipulation des fichiers, des sorties, de l'accès aux bibliothèques ainsi qu'aux fichiers. De meilleurs contrôles des systèmes seront préparés.

Ce RA n'a été produit que pour les régions urbaines. Au cours d'un développement méthodologique ultérieur, on examinera la possibilité de l'étendre aux régions rurales.

6.4 Méthodologie de mise à jour

Le RA a été créé à partir de quatre ensembles de fichiers administratifs: les fichiers des sociétés de téléphone, les fichiers des rôles d'évaluation des municipalités, les fichiers des sociétés d'électricité ainsi que le fichier de déclaration de revenu des particuliers (T1) de Revenu Canada. De plus, la mise à jour du RA est actuellement en cours afin que ce dernier soit compatible avec le recensement de 1991, de sorte que le recensement est aussi une source de données. Les contributions relatives de ces fichiers sources, tant en volume qu'en qualité, seront étudiées afin qu'on puisse prendre une décision pour ce qui est de l'acquisition de fichiers en vue d'effectuer la mise à jour.

L'élaboration d'une méthodologie pour effectuer la mise à jour constitue une partie intégrante de la stratégie de mise à jour. On aura besoin de la définition d'une mise à jour ainsi que d'un système de mise à jour. On

considérera aussi le rapport coût-efficacité d'une mise à jour continue, selon les divers besoins qui découlent des projets définis dans toutes ces orientations futures. La mise à jour continue est-elle rentable quand on la compare à la mise à jour effectuée seulement à temps pour le recensement? Quelles exigences découleront d'autres utilisations possibles? Les réponses à ces questions nous amèneront à élaborer une stratégie de mise à jour.

6.5 Autres utilisations du registre des adresses à Statistique Canada

En plus du recensement et des rapports géographiques présentés plus haut, un certain nombre d'autres utilisations sont proposées à l'intérieur de Statistique Canada. L'utilisation possible du RA pour l'enquête sur la population active (EPA) sera étudiée dans le cadre du projet de remaniement de l'EPA. La possibilité d'utiliser le RA dans les régions urbaines soit pour améliorer l'échantillonnage en fonction de la base de sondage existante, soit comme liste pour réduire le nombre d'étapes dans le plan d'échantillonnage est un domaine important mis au premier plan pour faire l'objet de recherches. Avec les numéros de téléphone qui figurent sur le RA, il serait possible de réaliser plus d'interviews téléphoniques.

L'utilisation du RA comme base de sondage pour d'autres enquêtes de Statistique Canada sera examinée. De plus, puisqu'on utilise actuellement les fichiers des sociétés de téléphone comme principale source de renseignements pour produire le RA, on dispose déjà de ces fichiers pour exploitation ultérieure. Le programme des enquêtes spéciales, l'enquête sociale générale et l'enquête existante sur la population active sont des enquêtes qui utilisent des fichiers des sociétés de téléphone ou qui en ont besoin.

Une autre application possible du RA à Statistique Canada consisterait à l'employer comme base de données sur le logement si on lui ajoutait les données sur le logement tirées du recensement de 1991 et les données provenant des rôles d'évaluation des municipalités, par exemple. L'existence d'une telle base de données pourrait réduire la quantité de renseignements sur le logement que l'on devrait recueillir dans les recensements à venir. Les besoins en données et la disponibilité de ces dernières doivent être étudiés.

6.6 Utilisations du registre des adresses à l'extérieur de Statistique Canada

La réalisation du RA est considérée comme une des entreprises conjointes possibles dans les discussions présentement en cours avec la Société canadienne des postes. Il vaut la peine d'étudier davantage comment Postes Canada pourrait contribuer à la mise à jour future du RA (comme source additionnelle ou peut-être comme seule source de renseignements) et comment le RA pourrait être utilisé par Postes Canada. Pour commencer, nous comparerons le RA à la base de données de Postes Canada. De plus, on considère la possibilité de faire du logiciel utilisé pour la normalisation des adresses du RA un projet conjoint.

Comme on l'a déjà mentionné, dans le cadre de la production du RA une entreprise conjointe a été réalisée avec le gouvernement de la Colombie-Britannique pour étendre le RA à des centres urbains plus petits afin de réduire le sous-dénombrement du recensement. D'autres entreprises avec la Colombie-Britannique ou avec d'autres provinces pourraient être réalisées.

Si le RA doit être utilisé à l'extérieur de Statistique Canada, il faut s'attaquer aux questions portant sur la confidentialité des fichiers sources. Certains fichiers sources ont été fournis à Statistique Canada à titre confidentiel, soit dans le cadre d'un contrat (p. ex., certains fichiers de l'Alberta et l'entreprise conjointe avec le gouvernement de la Colombie-Britannique), soit en vertu de la loi (le fichier des déclarations de revenus des particuliers (T1) de Revenu Canada).

6.7 Conclusion

L'ampleur et la diversité des idées mentionnées plus haut dans les orientations futures démontrent les possibilités qu'offre le registre des adresses comme produit géographique avec applications méthodologiques dans de nombreux domaines qui intéressent Statistique Canada et des organismes de l'extérieur.

REMERCIEMENTS

Les auteurs désirent remercier les nombreuses personnes qui font partie des services mentionnés ci-après pour leur dévouement et leur persévérance lors de la création du registre des adresses: Phillip Reed ainsi que la sous-section de la production du RA, Division de la géographie; l'unité du contrôle de l'échantillon de l'enquête sur la population active, Méthodes de recensement, Division des opérations des enquêtes; le Centre principal des ordinateurs et la Division des enquêtes des ménages. Les auteurs désirent aussi remercier Gordon Deecker, Peter Schut, Dick Carter, Phillip Reed et Carol Sol pour leurs suggestions utiles lors de la réalisation de la présente communication.

BIBLIOGRAPHIE

Booth, J.K. (1976). A summary report of all address register studies completed to date, Statistique Canada, Rapport interne.

Dick, P. (1990). Address register - September 1989 test, Statistique Canada, Ébauche interne.

Drew, J.D., Armstrong, J.B., et Dibbs, R. (1987). Research into a register of residential addresses for urban areas of Canada, American Statistical Association, *Proceedings of the Section on Survey Research Methods*, 300-305.

Drew, J.D., Armstrong, J., van Baaren, A., et Deguire, Y. (1988). Méthodologie de la construction d'un registre d'adresses à partir de plusieurs sources administratives, Statistique Canada, *Recueil du Symposium sur les utilisations statistiques des données administratives*, 209-219.

Fellegi, I.P., et Krotki, K.J. (1967). The testing programme for the 1971 Census in Canada, American Statistical Association, *Proceedings of the Social Statistics Section*, 29-38.

Fellegi, I.P., et Sunter, A.B. (1969). A theory for record linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Gamache-O'Leary, V., Nieman, L., et Dibbs, R. (1987). Cost implications of mail-out of Census questionnaires using an address register, Statistique Canada, Rapport interne.

Hill, T., et Pring-Mill, F. (1985). Generalized iterative record linkage system, *Proceedings of the Workshop on Exact Matching Methodologies*, Arlington, Virginia, 327-333.

Royce, D. (1986). Address register research for the 1991 Census of Canada, *Journal of Official Statistics*, 2, 4, 447-455.

Royce, D. (1987). Applications du registre des adresses au recensement du Canada, *Actes de la conférence internationale sur la planification du recensement de 1991*, Statistique Canada, 227-237.

Royce, D., et Drew, J.D. (1988). Address register research: Current status and future plans, Statistique Canada, Rapport interne.

Schut, P.H., et Haythornthwaite, T.W. (1990). Locating street addresses within a GIS, Canadian Institute of Surveying and Mapping, *Proceedings of the Conference on GIS for the 1990s*, 1055-1064.

Statistique Canada (1988). Area Master File (AMF), User guide, Rapport interne.

Statistique Canada (1989a). Automated Postal Coding System (PCODE), User and retrieval guide, Rapport interne.

Statistique Canada (1989b). Generalized Iterative Record Linkage System, Concepts guide, Rapport interne.

Statistique Canada (1989c). Postal Address Analysis System (PAAS), User guide, Rapport interne.

Statistique Canada (1989d). Record linkage software, Reference guide, Rapport interne.

Statistique Canada (1990). Guide à l'intention des utilisateurs sur la qualité des données du recensement de 1986: Couverture, 99-135F.

Statistique Canada (1991). Postal Code Conversion File (PCCF), the January 1991 version, User guide, Rapport interne.

Stewart, A. (1991). Joint Census undercount project: B.C. address register creation, Ministry of Finance and Corporate Relations, Gouvernement de la Colombie-Britannique, Rapport interne avec une diffusion restreinte.

van Baaren, A. (1988). Report on the November 1987 address register test, Statistique Canada, Rapport interne.

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

APPLICATIONS ACTUELLES ET FUTURES DE LA TÉLÉDÉTECTION AU NIVEAU DE LA COLLECTE DES DONNÉES SPATIALES

R. Ryerson et M. Manore[1]

RÉSUMÉ

La couverture des recensements de la population peut être évaluée soit par des méthodes démographiques, soit à l'aide d'estimations basées sur des enquêtes portant sur les erreurs de couverture, ou par une combinaison de ces deux genres de méthodes. Dans la présente communication on examine les progrès prévus au cours des cinq à dix prochaines années dans le domaine de l'imagerie par télédétection, de son interprétation et de son intégration dans la collecte courante de données spatiales et de données ponctuelles.

Les facteurs étudiés comprennent la disponibilité d'imageries diverses obtenues de l'espace et d'imagerie optique à plus haute résolution obtenue tant à partir de l'espace qu'à partir d'avions. Dans la communication on étudie aussi la nature des renseignements spatiaux dans le contexte du développement historique de la télédétection. Pour ce qui est du traitement des images, on passe en revue l'utilisation de systèmes experts afin d'effectuer l'interprétation et l'intégration, souvent connexe, de renseignements obtenus par télédétection dans l'environnement des systèmes d'information géographique (SIG) pour mieux interpréter les changements. À la fin de la communication, on jette un coup d'oeil sur l'application possible de la télédétection à divers domaines qui pourraient intéresser les personnes s'occupant de la collecte de statistiques spatiales.

MOTS CLÉS: Télédétection; données spatiales; systèmes d'information géographique.

1. INTRODUCTION

Dans la présente communication on résume brièvement la situation de la télédétection aujourd'hui et on la projette dans l'avenir en fonction des applications probables et des questions qui intéressent les personnes travaillant dans le domaine des renseignements géographiques ou spatiaux. La télédétection est définie ici comme la collecte de renseignements sur les richesses naturelles à partir d'une imagerie obtenue à l'aide de capteurs aéroportés ou spatiaux. L'emploi de la télédétection présente de nombreux avantages, mais ses limitations peuvent aussi être importantes. La présente communication ne vise pas à présenter en détail tous ces avantages et toutes ces limitations, mais un résumé avec l'accent mis sur les renseignements spatiaux est présenté dans l'annexe A.

2. SYSTÈMES DE TÉLÉDÉTECTION SPATIAUX ET AÉROPORTÉS

C'est en 1972 que les États-Unis ont lancé, pour la surveillance des richesses naturelles, le premier satellite d'observation de la terre (Landsat 1). Jusqu'à une période aussi récente que le milieu des années 80, le programme Landsat était le seul système spatial conçu spécialement pour la surveillance des richesses naturelles. Le système, basé sur le balayeur multibande (S.M.B.) à résolution spatiale de 80 mètres n'était pas

[1] R. Ryerson et M. Manore, Centre canadien de télédétection, Énergie, Mines et Ressources, 1547 rue Merivale, Ottawa (Ontario), Canada K1A 0Y7.

particulièrement utile pour une utilisation étendue à l'extérieur de certaines applications dans le domaine agricole, cartographique et océanographique (Ryerson et Howarth 1983; Thompson et coll. 1982 et Alfoldi et Munday 1978). La série de satellites météorologiques de la National Oceanographic and Atmospheric Administration (NOAA) des États-Unis avec leur radiomètre perfectionné à très haut pouvoir de résolution (AVHRR) qui a une résolution d'un kilomètre, a aussi été utilisée pour couvrir de grandes régions à des fins d'évaluation de la végétation.

L'arrivée de l'appareil de cartographie thématique (TM) Landsat à haute définition, augmente considérablement la gamme d'applications. (Voir le numéro spécial de *Geocarto International 1990* (1)). En 1986, la France a fait son entrée dans le domaine avec son système SPOT. Aujourd'hui, l'Europe, l'Inde, l'Union Soviétique et le Japon offrent tous de l'imagerie sur une base commerciale ou quasi-commerciale. Les caractéristiques des capteurs dont on peut facilement obtenir les données sont présentées au tableau 1.

Les caractéristiques des systèmes SPOT, Landsat et AVHRR de la NOAA sont bien résumées dans plusieurs notes d'information fournies par l'organisme qui emploie leurs auteurs pour ce qui est de l'application de ces systèmes en sylviculture, en agriculture et en géologie (Anon. 1986, 1987, 1988).

Les systèmes de télédétection aéroportés ont été améliorés pendant la même période, ils couvrent maintenant de plus grandes superficies avec des résolutions spatiales et spectrales accrues. En plus des capteurs optiques et à infra-rouge on trouve des systèmes actifs qui font appel aux lasers et aux radars (Till 1987 et Raney et coll. 1991). Pour diverses raisons décrites plus loin, ces systèmes présentent un défi particulier aux personnes qui s'occupent de la collecte et du traitement de données spatiales.

Comparativement aux capteurs disponibles dans les années 70 et au début des années 80, les systèmes de télédétection qui peuvent être utilisés deviennent beaucoup plus spécialisés et ils sont beaucoup plus produits en fonction des applications. Les caractéristiques spatiales et spectrales des nouveaux capteurs sont étroitement assorties à des besoins identifiables des organismes qui s'occupent de la gestion des ressources ou de la surveillance. Conjointement à ce que l'on appelle l'évaluation des besoins, on effectue habituellement des essais et des évaluations complets afin de s'assurer qu'il existe un marché pour tout nouveau capteur avant qu'il soit effectivement lancé ou utilisé.

3. LA NATURE SPATIALE DE L'IMAGERIE OBTENUE PAR TÉLÉDÉTECTION

Nous supposons ici que les renseignements spatiaux sont utilisés pour expliquer, mesurer ou mieux comprendre le monde dans lequel nous vivons. L'élaboration de modèles du monde réel constitue une application des renseignements spatiaux. Beaucoup d'efforts ont été consacrés à l'élaboration de tels modèles à la fin des années 60. En fait, certains des géographes qui ont pris part à la discussion à ce moment (y compris l'auteur principal de la présente communication) ont constaté qu'ils entraient dans le nouveau domaine de la télédétection afin de disposer d'un moyen d'obtenir de meilleurs renseignements spatiaux à propos de caractéristiques qui présentaient un intérêt pour leurs sous-disciplines géographiques particulières.

Le fait que les géographes se soient intéressés tôt à la télédétection ne constitue qu'un prolongement naturel de l'étude géographique traditionnelle. On peut constater ce cheminement plus clairement en suivant l'étude effectuée par Haggett sur la construction de modèles (Haggett 1965; 19-20) dans laquelle on voit plusieurs niveaux d'abstraction du monde réel pour un système particulier. Le premier niveau d'abstraction qu'il a mentionné pour un réseau routier, par exemple, serait une photographie aérienne de ce réseau. Le deuxième niveau serait une carte du réseau routier alors que le troisième, serait une mesure de la densité des routes. Essentiellement, on peut considérer ces niveaux d'abstraction comme un continuum tel que représenté ci-dessous:

Monde réel → Image → Renseignements spatiaux → Modèle → Théorie → Application

Le fait que ces abstractions ont été tirées d'images ressemblant à des photos aériennes faciles à comprendre dans le modèle, basé sur la vue, qu'utilisent les géographes a été très important dans le développement de la télédétection. Les géographes ont appliqué depuis longtemps des principes d'organisation spatiale pour extraire des renseignements à partir d'images à l'aide de clés et d'outils semblables (voir, par exemple, l'étude présentée

dans Ryerson 1989). On pourrait soutenir que si le géographe ne pouvait comprendre facilement les images et si ces dernières ne s'accordaient pas avec la vision du monde qu'a le géographe ou avec la façon dont ce dernier construit ses modèles, d'autres personnes travaillant dans d'autres disciplines élaboreraient les applications puisque ces personnes seraient mieux équipées pour comprendre les images et, par conséquent, mieux en mesure d'extraire des renseignements.

On pourrait même soutenir que les complexités associées à l'imagerie radar ont entraîné une participation accrue de personnes oeuvrant dans d'autres domaines, comme le génie électrique et la physique, à l'élaboration d'applications ou aux tentatives en ce sens. Cela est facile à comprendre puisque ces nouvelles formes d'imagerie sont le plus facilement interprétées en des termes qui sont les mieux compris par le physicien ou l'ingénieur en électricité. Ces nouvelles formes d'images plus complexes représentent une abstraction du monde réel qui est éloignée d'un niveau du modèle visuel si bien compris par le géographe traditionnel. Cette situation a une conséquence malheureuse car il se peut que les personnes formées en génie électrique et en physique comprennent bien l'imagerie et même la nature physique de ce dont on obtient l'image, mais pas le contexte spatial ou géographique dans lequel on trouve ces objets. Par conséquent, certains problèmes intéressants ont vu le jour dans le domaine de l'extraction des renseignements.

4. EXTRACTION DES RENSEIGNEMENTS

Pendant quelques années, des lignes directrices ou des clés ont été élaborées et utilisées pour l'interprétation d'imageries comme les photographies aériennes. Ces moyens permettaient à un interprète peu expérimenté de suivre un modèle élaboré par une personne possédant plus d'expérience. Cette forme d'extraction des renseignements n'a pas été populaire au cours des quelque quinze dernières années.

Depuis 1972, on a accordé plus d'attention à l'extraction de renseignements multispectraux. La méthode utilisée était basée sur l'hypothèse que des objets semblables auront des caractéristiques spectrales semblables dans plusieurs régions spectrales et sur le corollaire que des objets différents auront des caractéristiques spectrales différentes. Cette simplification excessive a entraîné certaines erreurs ahurissantes - comme le fait de confondre des roches nues avec des régions urbaines, par exemple. (Dans un cas, les projections basées sur des erreurs de ce genre ont entraîné l'obtention de résultats sans signification et une dépense d'argent considérable dans le cadre d'un gros programme de cartographie dans la partie américaine du bassin des Grands Lacs.) Poursuivant la même idée, on a décidé qu'il serait possible d'ajouter une dimension en superposant des images obtenues à des dates différentes afin de tirer profit des variations saisonnières qui pourraient permettre de séparer des objets différents. Bien que cette façon d'agir se soit révélée utile dans certaines applications, l'élément le plus important pour la présente discussion est qu'il existe maintenant une série de méthodes qu'on peut utiliser pour superposer des ensembles de données différents d'imageries obtenues par télédétection. Une bonne partie de ce travail peut maintenant être réalisée à l'aide des systèmes d'information géographique.

Cependant, comme cela a été décrit plus en détail ailleurs (Ryerson 1989), le problème fondamental lié à l'approche multispectrale (même quand elle fait appel à des données pluritemporelles) est qu'on n'utilise qu'un des nombreux éléments d'interprétation de base disponibles. Ce type de méthode ne tient pas compte de la texture, du contexte, de la structure et d'autres renseignements spatiaux. C'est pourquoi on a récemment consacré plus d'efforts à la question des renseignements spatiaux.

Il y a aussi eu un renouveau d'intérêt à propos des clés dans le contexte des systèmes experts et de l'intelligence artificielle. On élabore maintenant des systèmes qui tiendront compte de l'expertise spéciale d'interprètes expérimentés, tout en permettant à l'analyste de faire une utilisation appropriée des algorithmes spectraux et spatiaux disponibles (Ryerson 1989).

C'est à titre d'expert en données spatiales que le géographe trouvera un nouveau rôle et de nouvelles utilisations pour les données obtenues par télédétection. De plus, la nouvelle «sorte» de géographes possédant de l'expertise dans le traitement informatique des données spatiales sera bien placée pour faire progresser la technologie de l'extraction de renseignements à partir de sources obtenues par télédétection.

5. LA TÉLÉDÉTECTION COMME SOURCE DE RENSEIGNEMENTS SPATIAUX

En général, la télédétection n'est pas très utile pour fournir des mesures socio-économiques *directes* (sauf pour les piscines!). Cette méthode est beaucoup mieux adaptée aux enquêtes sur les richesses naturelles, sur la couverture végétale ou l'utilisation des sols et sur le temps de changement (dans les forêts, l'utilisation des sols, les pratiques agricoles, les plans d'eau, etc.). Par exemple, on a démontré l'utilité de la télédétection pour des études sur la population dans les pays en voie de développement (Ryerson et Lo 1990). On ne pourrait faire la même démonstration au Canada.

Bien qu'il soit évident que la télédétection offre des possibilités considérables pour fournir des renseignements spatiaux utiles dans le cas d'enquêtes sur les richesses naturelles, la télédétection seule ne peut permettre de répondre qu'à peu de besoins au niveau des renseignements spatiaux. Cela est particulièrement le cas dans un pays développé comme le Canada où il existe une infrastructure bien élaborée pour effectuer la collecte de données avec laquelle la télédétection doit entrer en concurrence. Par exemple, les superficies en culture peuvent être facilement déterminées pour certaines cultures spécialisées au Canada mais, pour les cultures sur de grandes superficies, d'autres méthodes se sont révélées de beaucoup supérieures à la télédétection. Toutefois, quand ces autres méthodes n'existent pas, il se peut bien que la télédétection soit concurrentielle.

La télédétection peut souvent fournir un genre nouveau ou différent d'informations qui s'ajoutent à l'ensemble des renseignements existants. On peut citer un certain nombre d'exemples d'imageries satellites utilisées pour la collecte de données spatiales. La bibliothèque du CCT contient plus de 80,000 documents, dont plus de la moitié portent sur une forme ou l'autre d'extraction de renseignements spatiaux. Par exemple, une mesure spécialisée de l'état des cultures a été élaborée à l'aide de données obtenues par les satellites météorologiques de la NOAA pour, à un coût relativement faible, aider à la prévision du rendement des cultures sur de grandes superficies. Ces renseignements ont été produits sous forme d'un rapport hebdomadaire par Statistique Canada. D'autres applications au Canada comprennent la mise à jour de renseignements thématiques comme la couverture forestière et le développement urbain sur des cartes topographiques. Les inventaires forestiers dans tout le pays sont mis à jour régulièrement à l'aide d'imageries satellites. Reconnaissant les possibilités qu'offre l'imagerie satellite pour ce qui est des vues d'ensemble, on utilise aussi cette imagerie pour aider à planifier le travail sur le terrain et à choisir les placettes d'échantillonnage. En Australie, l'emploi de l'imagerie satellite a permis aux vulgarisateurs dans le domaine agricole de réduire de 15 pourcent leurs frais de déplacement (communication personnelle, K. McLoy, NSW Dept. of Agriculture, Sydney).

6. LES ANNÉES 90 ET PLUS TARD - PROBLÈMES ET OCCASIONS

Un nouveau problème important fait son apparition dans le domaine de la télédétection. Les données recueillies par les nouveaux imageurs qui utilisent des parties du spectre électromagnétique avec lesquelles nous sommes peu familiers sont plus difficiles à interpréter. On peut dire, tout simplement, que les nouveaux systèmes fournissent une imagerie que nous ne pouvons relier au monde comme nous le voyons généralement. Notre perception est entachée d'un certain biais associé à nos systèmes de capteurs visuels (nos yeux) qui fonctionnent dans la partie bleu-vert-rouge (c.-à-d. la partie visible) du spectre.

Pour comprendre et utiliser les photographies aériennes et les premières imageries satellites, il suffisait de posséder une compréhension générale du système environnemental dont on obtenait une image - qu'il s'agisse de l'urbanisation dans le sud de l'Ontario, d'un champ de pommes de terre au Nouveau-Brunswick ou d'une forêt en Colombie-Britannique.

De nos jours, cela ne suffit plus. On doit comprendre ces systèmes et aussi pouvoir saisir les interactions complexes entre le capteur actif et les objets captés pour une gamme complète de variables. De plus, la diversité des genres de capteurs disponibles augmente.

Pour résoudre ces problèmes, on utilise diverses méthodes. On a créé des équipes de projets multidisciplinaires dans lesquelles on retrouve un large savoir-faire afin de s'attaquer au problème qui consiste à trouver la meilleure façon de transformer les données en renseignements. L'accent a été mis, de plus en plus, sur la préparation appropriée avant le lancement de nouveaux capteurs. Une fois que l'on a disposé de l'imagerie,

d'autres innovations ont été appliquées. L'intégration de la télédétection à d'autres ensembles de données, souvent à l'aide d'un SIG, constitue un facteur important dans l'accroissement récent de l'efficience pour ce qui est de l'utilisation de l'imagerie obtenue par télédétection.

La carte-image constitue un exemple simple qui illustre comment divers facteurs se sont combinés pour donner un nouveau produit qui contient des renseignements spatiaux. Pour certaines personnes, le travail a commencé avec, comme objectif, la production d'affiches esthétiquement plaisantes. Pour d'autres personnes, on voulait combiner les renseignements donnés par une carte avec le détail d'une image. Le résultat de ces travaux est un ensemble entièrement nouveau de produits cartographiques qui combinent l'esthétique des cartes-affiches et l'intégrité cartographique d'une carte. Maintenant, sur la majorité des cartes-images ordinaires, on superpose certains renseignements topographiques de base (comme le réseau routier) aux informations fournies par l'image. L'emploi d'une clé-légende simple, permet à l'utilisateur d'extraire de ces cartes-images les renseignements dont il a besoin. Ce produit particulier semble répondre à divers besoins et à un marché de masse.

La tendance en vue de fournir des produits améliorés «à valeur ajoutée» comme la carte-image est accompagnée de l'établissement de groupes ou d'organismes qui se spécialisent dans le prétraitement de données brutes obtenues par télédétection et dans leur conversion en produits d'information utilisables. Par exemple, pour surveiller les conditions de végétation dans l'ouest du Canada, le Centre manitobain de télédétection géocode et traite, de façon opérationnelle, une imagerie AVHRR de la NOAA brute pour obtenir, sur une base hebdomadaire, un produit normalisé pratiquement exempt de couverture nuageuse. Ces données sont, à leur tour, achetées par divers organismes, y compris la Division de l'agriculture de Statistique Canada où un traitement ultérieur permet d'obtenir, pour les utilisateurs finals, une série de produits cartographiques, statistiques et graphiques faciles à comprendre. À chaque étape, il faut disposer d'un niveau d'expertise technique qui n'est pas généralement disponible afin d'obtenir le produit à valeur ajoutée, ce qui amène le besoin de bureaux spécialisés pour traiter les données obtenues par télédétection. On s'attend à ce que ce genre de spécialisation augmente à cause de la plus grande compréhension technique exigée pour interpréter l'imagerie radar, à mesure qu'un plus grand nombre de telles données deviennent disponibles.

7. CONCLUSION

La télédétection a déjà eu une incidence importante sur la façon dont nous recueillons les renseignements spatiaux et dont nous pensons à ces renseignements. Elle nous a fourni une vue de notre monde qui nous en facilite la compréhension tout en soulignant les rapports réciproques complexes qui existent à l'interface entre l'environnement construit par l'homme et le milieu naturel. Bien que la technologie devienne plus complexe, elle est aussi présentée sous une forme qui la rend beaucoup plus facile et beaucoup moins coûteuse à utiliser que ce n'était le cas il y a à peine dix ans. On peut dire, avec une certaine confiance, que l'élaboration de nouvelles applications relatives à la collecte de données spatiales est limitée plus par l'imagination de l'utilisateur éventuel que par la technologie.

Tableau 1 - Caractéristiques des capteurs à bord des satellites

Capteurs des satellites Landsat

1. Balayeur multibande

Largeur de couloir couvert	185 km	Résolution spatiale 80 m
Bandes spectrales:	1.	0.50 - 0.60 micromètre (vert)
	2.	0.60 - 0.70 micromètre (rouge)
	3.	0.70 - 0.80 micromètre (infrarouge proche)
	4.	0.80 - 1.10 micromètre (infrarouge proche)
Résolution radiométrique	64 niveaux de gris	

2. Appareil de cartographie thématique (TM)

Largeur de couloir couvert	185 km	
Résolution spatiale	30 m (sauf pour la bande 6)	
Bandes spectrales:	1.	0.45 - 0.52 micromètre (bleu)
	2.	0.52 - 0.60 micromètre (vert)
	3.	0.63 - 0.69 micromètre (rouge)
	4.	0.76 - 0.90 micromètre (infrarouge proche)
	5.	1.55 - 1.75 micromètre (infrarouge ondes courtes)
	6.	10.5 - 12.5 micromètres (infrarouge thermique - 120 m)
	7.	2.08 - 2.35 micromètres (infrarouge ondes courtes)
Résolution radiométrique	256 niveaux de gris	

Capteurs du système SPOT

1. PLA

Largeur de couloir couvert	60 ou 117 km
Résolution spatiale	10 m
Bandes spectrales:	0.51 - 0.73 micromètre
Résolution radiométrique	64 niveaux de gris

2. MLA

Largeur de couloir couvert	60 ou 117 km
Résolution spatiale	20 m
Bandes spectrales:	0.50 - 0.59 micromètre (vert)
	0.61 - 0.68 micromètre (rouge)
	0.79 - 0.89 micromètre (infrarouge proche)
Résolution radiométrique	256 niveaux de gris

Capteur AVHRR de la NOAA

Largeur de couloir couvert	2,500 km
Résolution spatiale	1.1 km

Bandes spectrales:	0.58 - 0.68 micromètre (rouge)
	0.725 - 1.10 micromètre (infrarouge proche)
	3.55 - 3.93 micromètres (infrarouge ondes courtes)
	10.5 - 11.3 micromètres (infrarouge thermique)
	11.5 - 12.5 micromètres (infrarouge thermique)

BIBLIOGRAPHIE

Alfoldi, T.T., et Munday, J.C. (1978). Water quality analysis by digital chromaticity mapping of landsat data, *Canadian Journal of Remote Sensing*, 4, 108-126.

Anon. (1986). *Remote Sensing for Forestry*, CCRS, EMR, Ottawa, Ontario.

Anon. (1987). *Remote Sensing for Agriculture*, CCRS, EMR, Ottawa, Ontario.

Anon. (1987). *Remote Sensing for Geology*, CCRS, EMR, Ottawa, Ontario.

Haggett, P. (1965). *Locational Analysis in Human Geography*, London: Edward Arnold.

Raney, R.K., Luscombe, A.P., Langham, E.J., et Ahmed, S. (1991). Radarsat, *IEEE Proceedings*.

Ryerson, R.A., et Howarth, P.J. (1983). Canadian landsat studies for monitoring agricultural intensification and urbanization: A Summary, *Advanced Space Research*, 2, 8, 147-150.

Ryerson, R.A. (1989). Image interpretation concerns for the 1990s and lessons from the past, *Photogrammetric Engineering and Remote Sensing*, 55, 10, 1427-1430.

Ryerson, R.A., et Lo, C.P. (1990). Remote sensing for demographic studies related to global change, (Invited Paper), ISPRS Commission VII, (Interpretation of Data), Mid Term Symposium.

Thompson, M.D. (Editor-in-Chief) (1982). *Landsat for Mapping the Changing Geography of Canada*, Special Publication to mark the COSPAR Meeting in Ottawa, 1982. CCRS, EMR.

Till, S.M. (1987). Airborne electro-optical sensors for resource management, *Geocarto International*, 2, 3, 13-23.

Annexe A

AVANTAGES ET LIMITATIONS DE LA TÉLÉDÉTECTION

Avantages:

Contrairement à une entrevue ou à une visite sur le terrain, une fois qu'une image est obtenue, elle peut être réinterprétée afin d'en tirer de nouveaux renseignements. Cela n'est pas possible dans le cas des visites sur le terrain ou des entrevues.

Une image contient un aperçu de tout ce qui était dans la champ de vision du capteur.

Les données numériques, corrigées géométriquement, peuvent être combinées avec d'autres données spatiales à l'intérieur d'un SIG.

Les données peuvent être utilisées afin d'aider à effectuer une stratification pour des enquêtes spéciales.

Accroissement du détail au niveau spatial par rapport aux sources cartographiques.

Limitations:

La résolution spatiale et les parties du spectre électromagnétique captées sont limitées par le groupe-capteur utilisé.

La couverture ne peut être effectuée en temps opportun pour de grandes superficies avec des résolutions fines.

L'obtention de données à des fins d'échantillonnage coûte aussi cher que l'achat d'une couverture complète.

Coût élevé pour la couverture de grandes superficies.

SESSION 6

Qualité des produits de données spatiales

Recueil du Symposium 91 de Statistique Canada
Questions spatiales reliées aux statistiques
Novembre 1991

LE TRAITEMENT DES ERREURS DANS LES BASES DE DONNÉES SOCIO-ÉCONOMIQUES: RÉSULTATS CHOISIS D'UNE INITIATIVE NATIONALE DE RECHERCHE

U. Deichmann, M.F. Goodchild et L. Anselin[1]

RÉSUMÉ

L'utilisation croissante des systèmes d'informations géographiques dans toute une gamme d'applications en recherche et en gestion attire de plus en plus l'attention sur le problème de la précision des bases de données spatiales. Nous examinons des résultats choisis d'études menées dans le cadre d'une initiative de recherche du National Center for Geographic Information and Analysis. Cette question est abordée du point de vue de l'utilisateur de données socio-économiques publiées. Il est soutenu que la tendance croissante à l'utilisation de grandes bases de données intégrées et à l'automatisation de la manipulation de données spatiales soulève des questions reliées à la qualité des produits géographiques, questions qui souvent ne sont pas adéquatement traitées en analyse appliquée. L'importance de l'analyse de sensibilité en analyse spatiale est démontrée à l'aide d'un exemple de modélisation multirégionale intégrée.

MOTS CLÉS: Systèmes d'informations géographiques; précision des bases de données spatiales; modélisation socio-économique.

1. INTRODUCTION

Les systèmes d'information géographiques (SIG) sont de plus en plus appliqués par des géographes, des statisticiens, des planificateurs et des scientifiques de la cognition régionale à titre d'instrument facilitant toute une gamme de formes d'analyses spatiales faisant intervenir des données sociales, économiques et démographiques. Les méthodes nouvelles et l'extension d'applications d'études axées sur des recherches spécifiques à des analyses générales de politiques soulèvent des questions quant à la précision des opérations avec des bases de données spatiales qui n'ont fait l'objet que d'une attention restreinte en analyse spatiale classique. Le problème de la précision des bases de données spatiales est souvent décrit de manière simpliste comme étant un problème du domaine cartographique. Du point de vue de la cartographie, c'est la précision de la position des objets topologiques stockés dans la base de données qui constitue la principale préoccupation, p. ex. dans quelle mesure les lignes de la base de données correspondent-elles aux "vraies" lignes à la surface de la Terre. La précision des bases de données spatiales est cependant également importante du point de vue de la modélisation spatiale ou de la cognition régionale. L'imprécision des bases de données spatiales pose pour plusieurs raisons un problème plus sérieux en traitement numérique des données que ce n'est le cas en cartographie classique. La plus importante de ces raisons tient peut-être au fait que les SIG permettent la création de très grandes bases de données puisant à toute une gamme de sources de données souvent hétérogènes. Une faisabilité accrue sur le plan technique et un partage croissant de l'information spatiale par les institutions rendent nécessaire un examen davantage minutieux des effets de l'intégration de grands ensembles de données spatiales.

Le fait que l'exploitatiuon des SIG soit essentiellement indépendante de l'échelle constitue une deuxième question connexe. L'aptitude à utiliser des données recueillies à diverses échelles entraîne facilement l'introduction d'erreurs reliées à l'échelle. Si un processus spatial est présenté à une échelle, mais est étudié à

---

[1] U. Deichmann, M.F. Goodchild et L. Anselin, National Center for Geographic Information and Analysis, Department of Geography, University of California, Santa Barbara, Californie, 93106-4060, États-Unis.

partir d'ensembles de données basés sur des échelles différentes qui ne conviennent pas, les résultats de l'analyse peuvent être erronés. Enfin, les SIG permettent d'automatiser de manière à les rendre très souples les opérations de l'analyse spatiale. Cela signifie qu'un nombre beaucoup plus grand d'étapes peuvent être franchies dans le cadre d'un projet donné d'analyse spatiale et qu'ainsi il peut devenir impossible de suivre la propagation des erreurs.

Même si cette liste est loin d'être complète (d'autres points sont mentionnés par Goodchild et Gopal 1989), elle montre que l'utilisation des SIG pour la gestion et l'exploitation de données spatiales aggrave en réalité les problèmes associés aux d'erreurs que doivent régler les analystes de données spatiales. Pour ces raisons, une initiative en recherche a été prise au National Center for Geographic Information and Analysis (NCGIA) dans le domaine de la précision des bases de données spatiales. Elle a pris forme lors d'une réunion de spécialistes tenue en décembre 1988 à Montecito en Californie et ou le programme spécifique de recherche a été élaboré. Bien que cette initiative ait formellement pris fin en novembre 1990 lors d'une série de séances spéciales à la conférence sur les SIG/SIT, un grand nombre des projets individuels de recherche se poursuivent. Dans cet article nous décrivons brièvement certains des travaux accomplis dans le cadre de cette initiative et nous abordons les problèmes de précision dans la perspective de l'analyse de données socio-économiques. Finalement, nous présentons un exemple tiré d'un effort de modélisation multirégionale. D'après ces expériences, il est soutenu que bien que les SIG peuvent introduire davantage de problèmes, en termes de précision des bases de données spatiales, ils pourraient également constituer un cadre souple permettant à l'analyste de traiter ces problèmes fréquents.

2. INITIATIVE 1 DU NGCIA, PRÉCISION DES BASES DE DONNÉES SPATIALES

Du point de vue de l'analyse spatiale, les recherches qui ont été menées dans le cadre de l'Initiative 1 du NCGIA sur la précision des bases de données spatiales peuvent être regroupées en trois grands domaines:

- Compréhension de la nature des erreurs dans les bases de données spatiales.
- Compréhension des effets des erreurs dans les bases de données spatiales.
- Mise au point de méthodes de réduction ou de gestion des erreurs.

La compréhension de la nature des erreurs exige en tout premier lieu l'élaboration d'une classification commune des erreurs qui tienne compte des implications particulières de diverses structures de données (par exemple tracés en mode-trame versus tracés en mode-vecteur) ainsi que des différents types d'erreurs, comme les erreurs de position versus les erreurs d'attribut (voir Veregin 1989). Une classification commune est une condition préalable à la mise au point de méthodes fiables de modélisation des erreurs dans les divers modèles de données, méthodes par lesquelles le modèle d'erreur ne devrait pas seulement comporter une définition conséquente du type d'erreur, mais devrait également englober des mesures et des statistiques permettant de résumer l'incertitude associée à un ensemble de données (voir Goodchild 1990).

À titre d'exemple, examinons le cas d'une carte pédologique stockée sous forme d'un ensemble de données rastrées. Les limites pédologiques ne sont habituellement pas très bien définies parce que résultant d'une interpolation basée sur des échantillons ponctuels. Cela signifie qu'une incertitude considérable est rattachée à chaque limite. L'une des manières permettant de reconnaître explicitement cette incertitude consiste à attribuer à chaque pixel un vecteur de probabilité d'appartenance à une classe particulière de sol (Goodchild, Sun et Yang 1992). Une application analogue dans le domaine socio-économique serait une recherche sur un marché dans le cadre de laquelle un modèle de type gravitationnel (p. ex. le modèle de Huff) pourrait être utilisé pour prédire des limites bien nettes alors qu'en réalité ces limites sont beaucoup moins nettement définies. Une approche probabiliste comme celle suggérée ajoute une mesure distincte de l'incertitude au produit de SIG tout en le rendant davantage réaliste dans un monde où les relations sont floues. Des travaux connexes ont été menés en rapport avec les limites vectorielles de cartes superficie-classe où un processus de généralisation exigerait essentiellement une vue continue de l'espace et où la surface représente la probabilité d'appartenance à une classe particulière (Mark et Csillag 1989).

Tel que mentionné dans l'introduction, des successions d'opérations sur SIG mènent à la propagation d'erreurs dans le processus de l'analyse spatiale. L'une des conséquences de ce fait est que des produits libres d'erreurs

peuvent être contaminés lorsque combinés à des données renfermant des erreurs. Arbia et Haining (1992) ont mis au point un modèle matnématique théorique du processus de la propagation des erreurs. Une approche plus pratique consiste à mettre au point des indices de la précision des ensembles de données et à les utiliser pour documenter la qualité des cartes dérivées produites en analysant la généalogie des produits dérivés (Lanter et Veregin 1990). Ces travaux de poursuite des erreurs ont des conséquences importantes dans deux domaines: l'analyse des risques associés à la prise de décisions basées sur des produits de SIG d'une précision restreinte et la mise au point de normes de précision pour les organismes (Amrhein et Shut 1990).

Lorsque sont mieux compris la nature et les effets des erreurs dans les bases de données spatiales, l'étape suivante consiste à mettre au point des moyens et des méthodes permettant de réduire ou gérer les erreurs inhérentes à l'exploitation des SIG. Les modèles de représentation de surfaces constituent un exemple de la manière dont différentes structures de données permettent d'améliorer la précision du stockage d'objets spatiaux dans une base de données de SIG. Ces modèles ont été mis au point pour la représentation du terrain, mais ils sont également utilisés pour la modélisation de phénomènes socio-économiques comme les surfaces de densité de la population ou les surfaces de probabilité utilisées en modélisation criminologique. Une bonne compréhension des avantages relatifs de ces structures de données peut permettre de réduire les erreurs introduites par l'utilisation d'une mauvaise spécification du modèle (Kumler et Goodchild 1991).

Les questions reliées à l'échelle spatiale et à l'agrégation spatiale constituent l'un des domaines prioritaires de recherche en analyse spatiale (Amrhein et Flowerdew 1989; Rogerson 1990). Tel que précédemment mentionné, l'une des sources majeures d'erreur en modélisation à l'aide des SIG est le fait que les résultats fournis par les modèles dépendent de l'échelle (Fotheringham 1989). Plusieurs projets de recherche menés dans le cadre de l'Initiative 1 traitaient par conséquent de la mise au point de méthodes d'analyse indépendantes de l'échelle (Tobler 1989) et de l'évaluation des effets de l'agrégation spatiale.

Nous avons indiqué plus haut que l'intégration d'ensembles de données de provenances hétérogènes constitue l'une des raisons pour lesquelles la question de la précision est davantage sensible dans les bases de données spatiales que dans le cas de l'analyse classique. Des ensembles de données nécessaires pour une analyse sont souvent stockés suivant des formats incompatibles. Par exemple, un échantillon ponctuel peut être utilisé avec la superficie couverte par un polygone. Afin de pouvoir utiliser les données dans l'analyse, il faut souvent effectuer une interpolation. Cela signifie que pour l'analyse spatiale au moyen de SIG, de robustes sous-programmes d'interpolation spatiale sont nécessaires pour le traitement de configurations incompatibles de données spatiales (Flowerdew et Green 1989; Deichmann, Goodchild et Anselin 1990).

Dans cette partie nous avons présenté un bref résumé des recherches complétées ou en cours dans le cadre de l'Initiative 1 du NCGIA. Une information plus complète peut être trouvée ailleurs (Goodchild et Gopla 1989; NCGIA 1990; Goodchild 1990). Dans les parties suivantes nous présenterons des réflexions concernant des problèmes spécifiques de précision spatiale rencontrés dans l'analyse spatiale de données socio-économiques.

3. IMPLICATIONS POUR LES ANALYSES SPATIALES FAISANT INTERVENIR DES DONNÉES SOCIO-ÉCONOMIQUES

Tout utilisateur de données socio-économiques connaît les problèmes que pose l'obtention du degré d'uniformité nécessaire au niveau des données à utiliser dans l'analyse et la modélisation spatiales. Ces problèmes ont été bien résumés par Wassily Leontief (1986, p. 424):

> *"L'absence de coordination efficace dans le domaine général de la formulation et de la mise en oeuvre des politiques n'a d'égale que l'absence d'un plan général en matière de collecte, de structuration et de présentation des faits et des données dont dépend de manière si critique la prise de décisions tant dans le secteur public que dans le secteur privé".*

Dans la pratique, la classification générale suivante des problèmes au niveau des données s'impose: incompatibilités sectorielles/comptables; problèmes temporels; et erreurs spatiales.

Bien que ces trois ensembles de problèmes soient ennuyeux en modélisation régionale, les méthodes mises au point pour les régler sont plus ou moins efficaces. Les problèmes sectoriels et comptables se manifestent dans le processus de la collecte de données en raison d'un manque de coordination entre des organismes (souvent en concurrence les uns avec les autres). Le problème majeur est que les définitions des données socio-économiques changent souvent d'un organisme à un autre. Par exemple, aux États-Unis les définitions de la production industrielle utilisées dans les relevés comptables du Bureau of Economic Analysis diffèrent de celles utilisées par le Bureau of Census même si ces deux organismes font partie du ministère du commerce. Ces problèmes ne peuvent souvent être solutionnés que par des méthodes heuristiques ou intuitives.

Les questions d'ordre temporel ont à l'opposé fait l'objet d'une grande attention, principalement en économétrie et en analyse générale de successions chronologiques. Les mises à jour à intervalles irréguliers des données, les délais entre la collecte et la diffusion des données (p. ex. les relevés comptables nationaux des entrées et des sorties aux É.-U.) et les valeurs manquantes dans les successions chronologiques font partie de ce type de problèmes. Il existe quantité de travaux concernant la prévision des données socio-économiques ainsi que le problème des valeurs manquantes dans les successions chronologiques (voir Vandaele 1983; Granger 1986). De la même façon, le problème des valeurs manquantes dans des ensembles de données spatiales en coupe instantanée a récemment été abordé par Griffith, Bennett et Haining (1989). Cependant, une rationalisation des efforts consentis par les organismes responsables de la collecte des données permettrait d'améliorer considérablement une diffusion opportune et complète d'une grand nombre d'ensembles de données.

Les questions de précision spatiale sont reliées à la nature des erreurs spatiales qui peuvent être attribuables à des erreurs de spécification, à des erreurs de mesure ou à une combinaison des deux (Anselin 1989). Dans un contexte spatial, l'on désigne par erreur de spécification l'utilisation d'un modèle dont l'analyse ne tient pas compte de phénomènes spécifiques à l'emplacement comme la dépendance spatiale ou l'hétérogénéité spatiale (voir l'examen dans Anselin et Griffith 1988). D'une importance spécifique dans ce contexte est la question d'erreurs spatiale d'auto corrélation (à savoir si les erreurs sur une surface spatiale sont uniformément distribuées ou agglomérées). Dans les parties suivantes du présent article nous traiterons principalement des erreurs de mesure. Sous sa forme la plus simple l'erreur de mesure est présente lorsqu'une variable est repérée par rapport au mauvais emplacement géographique. La plupart du temps, cela est attribuable à des erreurs lors du processus de collecte. De toute évidence, des erreurs de position auront une incidence sur les opérations spatiales telles que interpolations, agrégations ou actions tampon. Elles auront de plus une incidence sur les résultats de tests d'effets spatiaux au stade de la modélisation. En plus des erreurs de position, il y a aussi fréquemment des erreurs associées aux attributs stockés avec les objets spatiaux. Ces erreurs d'attributs peuvent être attribuables à des erreurs de classification ou simplement à des erreurs d'introduction des données.

Les erreurs fréquemment dites conceptuelles (Chrisman 1989) constituent un autre type d'erreur de mesure. Elles ont trait au processus du transfert des entités du monde réel aux objets des bases de sonnées spatiales, ou en d'autres termes, à la manière dont les données sont enregistrées sous formes d'unités spatiales d'observation. Les données socio-économiques sont habituellement recueillies suivant des unités administratives principalement arbitraires qui ne tiennent pas compte des propriétés réelles des distributions des données. Des ensembles inconséquents ou inappropriés d'objets spatiaux utilisés pour la collecte et la structuration des données nuisent fréquemment à une analyse spatiale sensée. Ce problème se manifeste lorsque différents organismes recueillent des données pour différents ensembles de régions, ou lorsque des données sont recueillies à une échelle spatiale, mais nécessaires à une autre échelle, à un niveau moindre d'agrégation. Par exemple, des données peuvent n'être disponibles qu'au niveau national, mais nécessaires au niveau du comté ou du district. En modélisation économique régionale c'est par exemple le cas pour la production régionale brute au niveau du substrat ou pour les comptes d'entrée-sortie publiés.

Ce deuxième type d'erreur de mesure est relié à des questions fondamentales en analyse spatiale qui ont fait l'objet d'une grande attention, mais qui n'ont pas encore de solution satisfaisante. Les données socio-économiques régionales sont habituellement produites en faisant l'agrégation des caractéristiques d'agents économiques individuels (personnes ou entreprises) dans une portion de l'espace (Arbia 1989), ou en compilant des échantillons nationaux par unités superficie. L'agrégation repose sur des méthodes d'échantillonnage par lesquelles des conclusions pour toute une région sont tirées d'un certain nombre d'observations. Puisque les possibilités d'agrégation d'observations individuelles en groupes d'observations sont considérables, il est possible de soutenir que les conventions spécifiques utilisées pour l'agrégation auront une incidence sur les résultats de

l'analyse. Ce problème est appelé *problème de l'unité spatiale modifiable*, et découle, d'après Anselin (1988), du fait que les données sont recueillies pour des unités spatiales ayant des limites arbitraires et irrégulières, ce qui va à l'encontre de la notion générale d'espace homogène impliquant que l'agrégation n'est valide que si les caractéristiques clés sont uniformément réparties dans l'espace. Non seulement le morcellement de l'espace choisi pour l'agrégation, mais aussi le degré d'agrégation influencent les résultats de l'analyse spatiale. Ce problème est appelé *problème de l'erreur écologique* (voir Openshaw et Taylor 1979). D'après Arbia (1989), si les unités sont regroupées en en faisant la somme pour obtenir des unités plus grandes, la moyenne, la covariance et la corrélation varient. Tobler (1989) soutient cependant que les problèmes avec des unités spatiales modifiables et les problèmes dépendant de l'échelle ne devraient pas être attribués à la configuration des données spatiales, mais plutôt à une mauvaise méthode d'analyse.

En pratique, l'analyste n'a que rarement la possibilité d'effectuer un nouveau morcellement de sa région d'étude ou une agrégation à une échelle permettant d'éliminer les phénomènes parasites dans les ensembles de données; il doit donc travailler avec les données disponibles. Il est par conséquent important d'élaborer une notion théorique du problème qui aidera à évaluer l'incertitude rattachée aux résultats à différents degrés d'agrégation. Cela pourrait mener à la mise au point de méthodes d'analyse spatiale qui soient vraiment indépendantes de la configuration spatiale des données ou, comme le dit Tobler, "à une analyse spatiale indépendante du cadre" (Tobler 1989). Afin de se faire une idée des problèmes entravant l'analyse de l'impact socio-économique, il est utile d'étudier un exemple. Dans la partie suivante nous présentons par conséquent un cas d'effort de modélisation multirégionale dans lequel se pose le problème de l'unité spatiale modifiable.

4. UN EXEMPLE: MODÉLISATION RÉGIONALE INTÉGRÉE EN CALIFORNIE

Une part importante des travaux en cognition régionale consiste à mettre au point des modèles régionaux et multirégionaux intégrés pour l'analyse de l'impact socio-économique (p. ex. Isard 1986). Un exemple d'un tel effort de modélisation multirégionale relié aux politiques visant les ressources en eau en Californie est décrit dans Anselin, Rey et Deichmann (1990). Les quatre régions à la base de cette étude sont des agrégats des 58 comtés de Californie représentant une région urbaine et une région rurale, respectivement dans les parties septentrionale et méridionale (figure 1). L'objectif général du modèle est l'évaluation des effets de la dynamique des populations sur les changements de la demande finale des secteurs industriels. Ces changements entraînent à leur tour des changements de l'utilisation de l'eau dans différentes régions et ainsi des modifications de la configuration de l'écoulement de l'eau d'une région à l'autre. Dans le cadre du projet, l'emphase a été placée sur la détermination de l'incidence de différentes stratégies de modélisation sur les résultats du modèle. De manière plus spécifique, l'étude était concentrée sur le choix d'une échelle spatiale faisant intervenir une seule région ou plusieurs régions, sur la sélection d'une approche de mise en relation plutôt que d'une approche d'inclusion des divers modules et sur le problème de l'agrégation spatiale de données zonales incompatibles (pour les particularités voir Anselin, Rey et Deichmann 1990).

Pendant la collecte des données et la mise au point du modèle, un certain nombre des problèmes mentionnés plus haut se sont posés. Les données par comté pour les sept variables majeures suivantes étaient requises pour l'état de la Californie: emploi; paye, salaires, gains et revenus; production; valeur ajoutée; population; transports et mouvement des marchandises; et approvisionnement en eau ainsi que utilisation et transfert de l'eau. Un examen détaillé de la disponibilité des données et des problèmes qui leur sont associés a déjà été publié (Rey 1988). Nous nous concentrerons ici sur les problèmes spatiaux en cause.

Tel que précédemment mentionné, l'un des problèmes majeurs que pose le travail avec des données à référence spatiale publiées est que différents organismes publient des données concernant des régions ou des zones différentes. Si l'analyse est effectuée à une échelle de très grande agrégation et que les régions s'emboitent les unes dans les autres, le problème peut être réglé par simple agrégation. Si l'analyse est cependant assez détaillée, les données zonales davantage regroupées doivent être décomposées, ce qui exige une méthode d'estimation, p. ex. pour passer de données au niveau national à des données au niveau de l'état. À titre d'exemple mentionnons l'estimation du produit régional brut d'après des variantes de la méthode de Kendrick-Jaycox (p. ex. Weber 1979) dans lesquelles les rapports nationaux de la production sur ses composantes et les données régionales concernant les composantes sont utilisés afin d'estimer la production régionale brute.

l'analyse. Ce problème est appelé *problème de l'unité spatiale modifiable*, et découle, d'après Anselin (1988), du fait que les données sont recueillies pour des unités spatiales ayant des limites arbitraires et irrégulières, ce qui va à l'encontre de la notion générale d'espace homogène impliquant que l'agrégation n'est valide que si les caractéristiques clés sont uniformément réparties dans l'espace. Non seulement le morcellement de l'espace choisi pour l'agrégation, mais aussi le degré d'agrégation influencent les résultats de l'analyse spatiale. Ce problème est appelé *problème de l'erreur écologique* (voir Openshaw et Taylor 1979). D'après Arbia (1989), si les unités sont regroupées en en faisant la somme pour obtenir des unités plus grandes, la moyenne, la covariance et la corrélation varient. Tobler (1989) soutient cependant que les problèmes avec des unités spatiales modifiables et les problèmes dépendant de l'échelle ne devraient pas être attribués à la configuration des données spatiales, mais plutôt à une mauvaise méthode d'analyse.

En pratique, l'analyste n'a que rarement la possibilité d'effectuer un nouveau morcellement de sa région d'étude ou une agrégation à une échelle permettant d'éliminer les phénomènes parasites dans les ensembles de données; il doit donc travailler avec les données disponibles. Il est par conséquent important d'élaborer une notion théorique du problème qui aidera à évaluer l'incertitude rattachée aux résultats à différents degrés d'agrégation. Cela pourrait mener à la mise au point de méthodes d'analyse spatiale qui soient vraiment indépendantes de la configuration spatiale des données ou, comme le dit Tobler, "à une analyse spatiale indépendante du cadre" (Tobler 1989). Afin de se faire une idée des problèmes entravant l'analyse de l'impact socio-économique, il est utile d'étudier un exemple. Dans la partie suivante nous présentons par conséquent un cas d'effort de modélisation multirégionale dans lequel se pose le problème de l'unité spatiale modifiable.

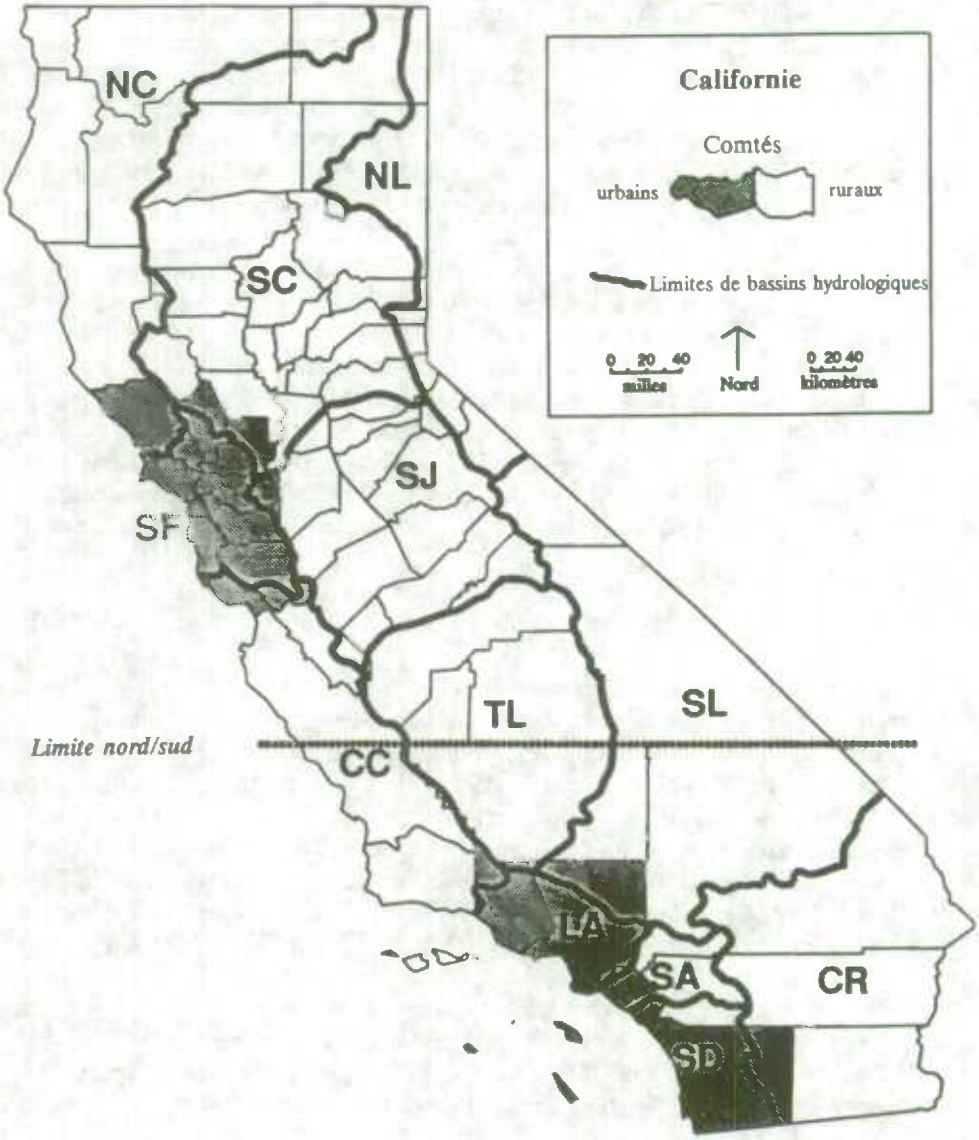
4. UN EXEMPLE: MODÉLISATION RÉGIONALE INTÉGRÉE EN CALIFORNIE

Une part importante des travaux en cognition régionale consiste à mettre au point des modèles régionaux et multirégionaux intégrés pour l'analyse de l'impact socio-économique (p. ex. Isard 1986). Un exemple d'un tel effort de modélisation multirégionale relié aux politiques visant les ressources en eau en Californie est décrit dans Anselin, Rey et Deichmann (1990). Les quatre régions à la base de cette étude sont des agrégats des 58 comtés de Californie représentant une région urbaine et une région rurale, respectivement dans les parties septentrionale et méridionale (figure 1). L'objectif général du modèle est l'évaluation des effets de la dynamique des populations sur les changements de la demande finale des secteurs industriels. Ces changements entraînent à leur tour des changements de l'utilisation de l'eau dans différentes régions et ainsi des modifications de la configuration de l'écoulement de l'eau d'une région à l'autre. Dans le cadre du projet, l'emphase a été placée sur la détermination de l'incidence de différentes stratégies de modélisation sur les résultats du modèle. De manière plus spécifique, l'étude était concentrée sur le choix d'une échelle spatiale faisant intervenir une seule région ou plusieurs régions, sur la sélection d'une approche de mise en relation plutôt que d'une approche d'inclusion des divers modules et sur le problème de l'agrégation spatiale de données zonales imcompatibles (pour les particularités voir Anselin, Rey et Deichmann 1990).

Pendant la collecte des données et la mise au point du modèle, un certain nombre des problèmes mentionnés plus haut se sont posés. Les données par comté pour les sept variables majeures suivantes étaient requises pour l'état de la Californie: emploi; paye, salaires, gains et revenus; production; valeur ajoutée; population; transports et mouvement des marchandises; et approvisionnement en eau ainsi que utilisation et transfert de l'eau. Un examen détaillé de la disponibilité des données et des problèmes qui leur sont associés a déjà été publié (Rey 1988). Nous nous concentrerons ici sur les problèmes spatiaux en cause.

Tel que précédemment mentionné, l'un des problèmes majeurs que pose le travail avec des données à référence spatiale publiées est que différents organismes publient des données concernant des régions ou des zones différentes. Si l'analyse est effectuée à une échelle de très grande agrégation et que les régions s'emboitent les unes dans les autres, le problème peut être réglé par simple agrégation. Si l'analyse est cependant assez détaillée, les données zonales davantage regroupées doivent être décomposées, ce qui exige une méthode d'estimation, p. ex. pour passer de données au niveau national à des données au niveau de l'état. À titre d'exemple mentionnons l'estimation du produit régional brut d'après des variantes de la méthode de Kendrick-Jaycox (p. ex. Weber 1979) dans lesquelles les rapports nationaux de la production sur ses composantes et les données régionales concernant les composantes sont utilisés afin d'estimer la production régionale brute.

Figure 1. Régions économiques et régions hydrologiques d'étude



Une variante plutôt davantage complexe du problème se manifeste lorsque les systèmes zonaux sont totalement ou partiellement incompatibles. Cela peut se produire lorsque les données proviennent de sources hétérogènes ou lorsque les limites de districts ont changé dans le temps. Puisque les données pour le modèle de l'eau en Californie provenaient d'une multitude de sources, ce problème s'est posé plusieurs fois. Par exemple, l'intégration des données sur l'approvisionnement en eau aux données sur le transfert de l'eau a été compliquée par le fait que pour ces rubriques les données sont publiées pour douze bassins hydrologiques dont les limites sont en grande partie incompatibles avec celles des comtés. D'autre part, les données sur l'utilisation de l'eau par l'industrie sont disponibles par comté. Dans le cadre final d'analyse des activités, il a donc fallu introduire un ensemble de facteurs de pondération traduisant les approvisionnements en eau par bassin hydrologique en demande d'eau par région économique.

Deux méthodes d'interpolation différentes ont été appliquées. La première est une interpolation directe de superficies (Goodchild et Lam 1980), dans laquelle deux ensembles de polygones sont superposés et la superficie des chevauchements pour chaque paire de région source et région cible est calculée. Disposés sous forme de matrice et normalisés, les dénombrements pour les zones sources peuvent ensuite être redistribués sur les zones

cibles proportionnellement aux superficies des chevauchements. De toute évidence, l'hypothèse sous-jacente pour cette méthode est que la distribution des variables est homogène à l'intérieur des zones sources et la méthode sera principalement appliquée dans les cas pour lesquels aucune information concernant une variable auxiliaire n'est disponible. Afin d'obtenir une indication de l'erreur potentielle introduite par l'utilisation de pondérations en fonction de la superficie dans une application socio-économique, on a obtenu une deuxième estimation basée sur une carte assez détaillée des centres de gravité de 5757 secteurs de dénombrement avec les populations totales qui leur étaient associées en 1980. Une nouvelle agrégation de ces points en régions économiques et hydrologiques fournit la part de la population de chaque bassin qui habite dans chacune des régions économiques (voir Anselin, Rey et Deichmann 1990 pour les détails). Les deux matrices de pondération sont présentées au tableau 1.

Tableau 1. Matrices de pondération pour l'interpolation spatiale

	Facteurs de pondération en fonction de la superficie				Facteurs de pondération en fonction de la population			
	Urbaine sud	Rurale sud	Urbaine nord	Rurale nord	Urbaine sud	Rurale sud	Urbaine nord	Rurale nord
NC	0.0000	0.0000	0.0687	0.9313	0.0000	0.0000	0.4936	0.5064
SF	0.0000	0.0000	0.9998	0.0002	0.0000	0.0000	1.0000	0.0000
CC	0.0222	0.5173	0.0774	0.3831	0.0000	0.4520	0.2424	0.3046
LA	0.9981	0.0019	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
SA	0.2037	0.7963	0.0000	0.0000	0.6003	0.3997	0.0000	0.0000
SD	0.8362	0.1638	0.0000	0.0000	0.9882	0.0118	0.0000	0.0000
SC	0.0000	0.0000	0.0344	0.9656	0.0000	0.0000	0.0489	0.9511
SJ	0.0000	0.0000	0.0225	0.9775	0.0000	0.0000	0.0541	0.9459
TL	0.0018	0.3417	0.0000	0.6565	0.0000	0.3052	0.0000	0.6948
NL	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000
SL	0.0476	0.4875	0.0000	0.4649	0.3175	0.6064	0.0000	0.0761
CR	0.0649	0.9351	0.0000	0.0000	0.0069	0.9931	0.0000	0.0000

Pour les applications multirégionales multisectorielles ni l'une ni l'autre des deux approches ne devrait être généralement supérieure. Dans les secteurs urbains (comme ceux de la fabrication ou des services) la population fournit une approximation plus judicieuse, alors que des pondérations pour les superficies pourraient sans doute fournir une meilleure approximation pour l'activité agricole. De toute évidence, les configurations d'écoulement réelles prévues avec un cadre d'analyse basé sur la programmation linéaire et l'activité sont très sensibles aux différentes stratégies d'interpolation (voir Anselin, Rey et Deichmann 1990). Dans un contexte de politiques, le choix d'une méthode régionale d'attribution aurait une grande incidence sur les résultats du modèle et ainsi sur des décisions de gestion qui pourraient être d'une grande portée. Étant donné la fréquence des problèmes de dispositions zonales incompatibles, une beaucoup plus grande attention devrait être consacrée à la mise au point de robustes schémas d'interpolation et d'analyses de la sensibilité des résultats de diverses stratégies d'interpolation et de modélisation.

Dans les cas pour lesquels des totaux de contrôle pour une variable auxiliaire ne sont pas disponibles, la dérivation d'estimations pour des zones incompatibles doit reposer sur l'approximation cartographique ou mathématique. Plusieurs autres méthodes ont été proposées en plus de la méthode de la pondération en fonction de la superficie. Dans l'interpolation pycnophylactique (conservant la masse) (Tobler 1979), on suppose une distribution uniforme dans laquelle des emplacements voisins ont des valeurs similaires de la densité. Par cette méthode une approximation de la surface est obtenue en drapant un fin réseau de points sur la région d'étude. La valeur totale pour la variable dans une région est ensuite itérativement distribuée sur les points de manière à ce que le total pour chaque région reste constant et que des points voisins aient des valeurs similaires. Une autre stratégie suggérée par Flowerdew et Green (1989) consiste en une méthode d'estimation statistique par laquelle l'on tient compte de l'information additionnelle disponible concernant la zone.

Les problèmes de systèmes zonaux incompatibles avec le modèle pour la Californie décrit ci-haut ont mené à une généralisation du modèle d'interpolation en fonction de la superficie. Les résultats initiaux ont été présentés par Deichmann, Goodchild et Anselin (1990). Si la distribution dans les zones cibles peut être supposée constante et que le nombre des zones cibles est plus petit que le nombre des zones sources, les densités dans

les zones cibles peuvent être estimées sous forme de coefficients d'une courbe de régression passant par l'origine avec les populations des zones sources comme variables dépendantes et les superficies des chevauchements entre les zones sources et les zones cibles comme variables indépendantes. Une ordinaire régression simple des moindres carrés ne guarantit toutefois pas que les coefficients (densités) estimés seront positifs et que la population estimée totale correspond à la population totale connue. Les recherches en cours ont par conséquent été concentrées sur l'utilisation de méthodes de régression avec contraintes (p. ex. Judge et Yancey 1986).

Une extension de cette méthode englobe les cas dans lesquels l'on dispose d'un troisième ensemble de zones, des zones de contrôle, présentant des densités constantes. Si le nombre des zones de contrôle est inférieur au nombre de zones sources, les densités des zones de contrôle peuvent être estimées de la manière précédemment décrite. La densité des zones cibles peut ensuite être déduite en intégrant la surface de densité de la population représentée par les densités des zones de contrôle sur la superficie de chaque zone cible. Les résultats préliminaires montrent qu'il n'est pas nécessaire que les zones de contrôle soient définies avec très grande précision pour obtenir une amélioration considérable de l'estimation des populations des zones cibles. Dans le cas de l'application californienne, quatre zones de contrôle pour lesquelles on a supposé des densités constantes ont été interactivement introduites dans une couche de SIG "par examen empirique", ceci pourrait donc être considéré comme l'opinion d'un expert.

5. CONCLUSION: L'IMPORTANCE DE L'ANALYSE DE SENSIBILITÉ

Parmi les problèmes que posent les efforts de modélisation spatiale, celui de dispositions incompatibles des sources de données est l'un des plus remarquables. Dans les parties qui précèdent, des méthodes de solution de ce problème ont été décrites; certaines de ces méthodes ont été mises au point dans le cadre de l'Initiative 1 du NCGIA sur la précision des bases de données spatiales. Puisque les bases de données spatiales deviennent plus grandes à mesure que progresse la technologie dans le domaine et en raison d'une tendance croissante à la combinaison de bases de données différentes, il devient de plus en plus nécessaire d'intégrer des données de provenances hétérogènes. D'autre part, la technologie des SIG offre la possibilité d'un cadre de mise en oeuvre de méthodes permettant de traiter ces tâches d'intégration. Les questions traitées dans le cadre de l'Inititive 1 du NCGIA sur la précision des bases de données spatiales sont en grande partie recoupées par les sujets de recherche menées dans le cadre de l'Initiative 14 sur l'analyse géographique et les SIG qui doit être lancée au printemps de 1992. L'amélioration des possibilités d'analyse au moyen de SIG exige qu'il existe des outils de manipulation et d'exploration de données permettant à l'utilisateur de concentrer ses efforts sur l'analyse de données confirmatives (voir Anselin et Getis 1992; Goodchild, Haining et Wise 1992).

L'un des atouts majeurs des SIG est le fait que les données y soient stockées de manière à assurer une grande souplesse dans l'adoption d'un cadre spatial d'analyse. Une application particulière découlant de ce caractère fonctionnel est possible lorsque de l'information auxiliaire peut être utilisée pour améliorer des estimations statistiques, de manière analogue à la tendance actuelle à l'utilisation de couches d'un SIG dans une base de données afin de faciliter la classification d'images obtenues par télédétection (Davis et Simonett 1991). L'utilisation d'images classées de l'appareil de cartographie thématique du Landsat pour faciliter une estimation croisée de la population d'une région (Langford, Maguire et Unwin 1990) constitue un pas dans la bonne direction. Ainsi, des zones de contrôle réelles ou subjectives stockées dans un SIG pourraient facilement être utilisées dans la méthode générale décrite par Deichmann, Goodchild et Anselin (1990) et l'interpolation pycnophylactique de Tobler pourrait être modifiée de manière à permettre de tenir compte d'étendues de la région d'étude dont on sait qu'elles ne présentent pas une distribution uniforme. En termes de mise en oeuvre réelle, un SIG pourrait servir de support pour une batterie de méthodes d'interpolation dans laquelle l'analyste choisirait celle qui est la plus compatible avec la nature des données et les hypothèses concernant leur distribution.

Il reste encore beaucoup à faire avant que soit mis au point un système générique de manipulation de données qui permette de travailler avec des données socio-économiques à référence spatiale sans avoir à se préoccuper de l'échelle et de la disposition des unités spatiales. L'on soutient par conséquent que dans l'analyse spatiale une emphase beaucoup plus grande doit être placée sur l'analyse de sensibilité qui devrait être utilisée à tous les stades du processus de la modélisation. Cela implique une analyse menée avec soin des propriétés statistiques des méthodes utilisées, et la mise à l'épreuve d'un certain nombre de configurations spatiales

possibles et de méthodes d'analyse, telles que mesures de corrélation croisée ou méthodes d'interpolation. L'objectif devrait être de fournir aux décideurs, qui sont les utilisateurs des résultats de l'analyse spatiale, une forme ou une autre de limites de confiance constituant une indication claire de l'incertitude associée au produit de l'analyse. De toute évidence jusqu'à maintenant l'expérience technique et l'effort de calcul considérables nécessaires en modélisation spatiale ont nui à la mise au point de telles méthodes. Les progrès accomplis en technologie des SIG et les efforts visant la mise au point de bases de données spatiales et de systèmes d'analyse véritablement intégrés devraient vraisemblablement permettre d'accéder à la souplesse qu'exigent des formes robustes d'analyse spatiale.

REMERCIEMENTS

Le présent article traite de travaux de recherche actuellement menés au National Center for Geographic Information and Analysis (NCGIA). Nous sommes reconnaissants de l'aide fournie par la National Science Foundation des États-Unis (subvention SES-88-10917).

BIBLIOGRAPHIE

Amrhein, C.G., et Flowerdew, R. (1989). The effect of data aggregation on a Poisson model of Canadian migration, dans *The Accuracy of Spatial Databases*, M. Goodchild et S. Gopal, Éds., Taylor et Francis, London, 229-238.

Amrhein, C.G., et Schut, P. (1990). Data quality standards and geographic information systems, *Actes, SIG pour la prochaine décennie*, Ottawa: Association canadienne des sciences géodesiques et cartographiques, 918-930.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, Dordrecht: Kluwer Academic Publishers.

Anselin, L. (1989). What is special about spatial data? Alternative perspectives on spatial data analysis, Technical Paper 89-4, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA.

Anselin, L., et Getis, A. (1992). Spatial statistical analysis and geographic information systems, *Annals of Regional Science*, 26 (sous presse).

Anselin, L., et Griffith, D.A. (1988). Do spatial effects really matter in regression analysis, *Papers of the Regional Science Association*, 65, 11-34.

Anselin, L., Rey, S., et Deichmann, U. (1990). The implementation of integrated models in a multi-regional system, dans *New Directions in Regional Analysis: Multi-Regional Approaches*, Éds., L. Anselin et M. Madden, London, Belhaven, 146-170.

Arbia, G. (1989). *Spatial data configuration in statistical analysis of regional economic and related problems*, Dordrecht: Kluwer Academic.

Arbia, G., et Haining, R.P. (1992). Error propagation through map operations, *Technometrics*, 34 (sous presse).

Chrisman, N.R. (1989). Modeling error in overlaid categorical maps, dans *The Accuracy of Spatial Databases*, M. Goodchild et S. Gopal, Éds., Taylor et Francis, London, 21-34.

Davis, F.W., et Simonett, D.S. (1991). GIS and remote sensing, dans *Geographical Information Systems: Principles and Applications*, D.J. Maguire, Éds., M.F. Goodchild et D.W. Rhind, New York: Wiley and Sons, 191-214.

Deichmann, U., Goodchild, M.F., et Anselin, L. (1990). A general framework for the spatial interpolation of socio-economic data, *Proceedings, Advanced Computing in the Social Sciences Conference*, Williamsburg, VA.

Flowerdew, R., et Green, M. (1989). Statistical methods for inference between incompatible zonal systems, dans *The Accuracy of Spatial Databases*, M. Goodchild et S. Gopal, Éds., Taylor et Francis, London, 239-248.

Fotheringham, A.S. (1989). Scale-independent spatial analysis, dans *The Accuracy of Spatial Databases*, M. Goodchild et S. Gopal, Éds., Taylor et Francis, London, 221-228.

Goodchild, M.F. (1990). Modeling error in spatial databases, *Proceedings, GIS/LIS 90*, Anaheim, 1, 154-162.

Goodchild, M.F., et Gopal, S. (Éds.) (1989). *The Accuracy of Spatial Databases*, London: Taylor and Francis.

Goodchild, M.F., et Lam, N.S. (1980). Areal interpolation: A variant of the traditional spatial problem, *Geoprocessing*, 1, 297-312.

Goodchild, M.F., Sun, G., et Yang, S. (1992). Development and test of an error model for categorical data, *International Journal of Geographical Information Systems*, 6 (sous presse).

Goodchild, M.F., Haining, R., et Wise, S. (1992). Integrating GIS and spatial data analysis, *International Journal of Geographical Information Systems*, 6 (sous presse).

Granger, C.W.J. (1986). *Forecasting Economic Time Series*, Boston: Academic Press.

Griffith, D.A., Bennett, R.J., et Haining, R.P. (1989). Statistical analysis of spatial data in the presence of missing observations: A methodological guide and an application to urban census data, *Environment and Planning A*, 21, 1511-1523.

Isard, W. (1986). Reflections on the relevance of integrated models for policy analysis, *Regional Science and Urban Economics*, 16, 165-180.

Judge, G.G., et Yancey, T.A. (1986). *Improved Methods of Inference in Econometrics*, Amsterdam: North Holland.

Kumler, M.P., et Goodchild, M.F. (1991). A new technique for selecting the vertices of a TIN, and a comparison of TINs and DEMs over a variety of surfaces, *Technical Papers, 1991 ACSM-ASPRS Annual Convention*, Baltimore, 2, 179.

Langford, M., Maguire, D.J., et Unwin, D.J. (1990). Cross area population estimation using remote sensing and GIS, *Proceedings, 4th International Symposium on Spatial Data Handling*, Zürich, 1, 541-550.

Lanter, D.P., et Veregin, H. (1990). A lineage meta-database program for propagation of error in geographic information systems, *Proceedings, GIS/LIS 90*, Anaheim, 1, 144-153.

Leontief, W. (1986). *Input-Output Economics*, 2e édition, New York: Oxford University Press.

Mark, D.M., et Csillag, F. (1989). The nature of boundaries on area-class maps, *Cartographica*, 21, 65-78.

NCGIA (1990). NCGIA 18 Month Report, Rapport technique 90-7, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA.

Openshaw, S., et Taylor, P. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem, dans *Statistical Applications in the Spatial Sciences*, N. Wrigley et R.J. Bennett, Éds., London: Pion, 127-144.

Rey, S. (1988). Data availability and data problems for an integrated multiregional model of California water resources, Rapport W-700-88.2, Community and Organization Research Institute, University of California, Santa Barbara, CA.

Rogerson, P.A. (1990). Migration analysis using data with time intervals of differing widths, *Papers of the Regional Science Association*, 68, 97-106.

Tobler, W.R. (1989). Frame independent analysis, dans *The Accuracy of Spatial Databases*, Éds., M. Goodchild et S. Gopal, London: Taylor et Francis, 115-122.

Tobler, W.R. (1979). Smooth pycnophylactic interpolation for geographical regions, *Journal of the American Statistical Association*, 74, 367, 519-530.

Vandaele, W. (1983). *Applied Time Series and Box-Jenkins Models*, New York: Academic Press.

Veregin, H. (1989). A taxonomy of error in spatial databases, Technical Paper 89-12, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA.

Weber, R.E. (1979). A synthesis of methods proposed for estimating gross state product, *Journal of Regional Science*, 19, 217-230.

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

PRÉCISION DE L'INDICE DES PRIX FRANÇAIS ET OPTIMISATION DES ÉCHANTILLONS

P. Ardilly[1] et F.Guglielmetti[2]

RÉSUMÉ

L'objet de cette étude est, d'une part de calculer des estimations de précision pour les indices à différents niveaux géographiques et à différents niveaux de la nomenclature des produits, et d'autre part d'optimiser les tailles d'échantillons d'agglomérations et de relevés, compte tenu des informations disponibles. Dans le cadre de la rénovation de l'indice des prix de détail, on espère ainsi économiser des relevés en maintenant une qualité égale à la qualité actuelle.

MOTS CLÉS: Indice de prix; estimateurs de variance; optimisation d'échantillons.

1. INTRODUCTION

L'indice des prix calculé en France par l'INSEE se définit comme un "indice de prix à la consommation des ménages urbains dont le chef est ouvrier ou employé". Le champ de l'indice est l'ensemble des biens et services consommés par les ménages de référence.

Ce champ est représenté au travers d'une stratification des biens et services. Chaque strate est appelé "Poste", et il y a environ 300 postes. Dans chaque poste, on choisit des représentants appelés "Variétés", en fonction de la part de consommation qu'elles représentent dans le poste, ainsi que d'une connaissance à priori de la représentativité de l'évolution de leurs prix dans la poste. Il y a environ 1 000 variétés. Les postes peuvent être regroupés en secteurs, au nombre de 4 (Alimentaire - Habillement - Autres biens manufacturés - Services).

L'échantillonnage des relevés de prix s'effectue à deux degrés de manière stratifiée. On constitue des strates géographiques en croisant des grandes régions appelées ZEAT et des catégories d'agglomérations notées *CC* qui se définissent selon le nombre d'habitants recensés dans l'agglomération. Il y a 8 ZEAT et 5 *CC*, qui sont:

CC 2: Agglomérations de 2 000 à 10 000 habitants
CC 4: Agglomérations de 10 000 à 100 000 habitants
CC 6: Agglomérations de 100 000 à 200 000 habitants
CC 8: Agglomérations de plus de 200 000 habitants
CC 9: Agglomération de Paris.

Pour les *CC* 2, 4 et 6, dans chaque croisement *CC*-ZEAT, on assimile le tirage à un sondage aléatoire simple d'agglomérations. Pour les *CC* 8 et 9, le tirage d'agglomérations est exhaustif.

Dans chaque agglomération, on échantillonne des relevés de prix. Là encore, le tirage est assimilé à un sondage aléatoire simple. Il n'existe évidemment pas de base de sondage à ce niveau, et on laisse l'enquêteur réaliser le tirage, sous quelques contraintes cependant (dont l'une des plus importantes est de respecter la structure de consommation dans les types de point de vente: commerce particulier, supermarché, hypermarché, etc.).

---

[1] P. Ardilly: Division Méthodes Statistiques et Sondages, INSEE, Paris, France.

[2] F. Guglielmetti: Division Prix de Détail, INSEE, Paris, France.

On peut alors constituer des indices élémentaires variété-agglomération. On distingue deux types de variétés: les variétés dites "homogènes", où on peut ajouter les prix des produits la constituant (exemple: la baguette de pain), et les variétés dites "hétérogènes", où une telle opération n'aurait pas de sens (exemple: la poupée d'enfant).

Si on note $p^t(j,i,v)$, le j ième relevé de prix concernant la variété v dans l'agglomération i à la date t, et si $n(i,v)$ est la taille de l'échantillon, alors on définit les estimations des véritables indices alimentaires inconnues par:

$$\hat{I}(i,v) = \frac{\bar{p}^t(i,v)}{\bar{p}^{to}(i,v)},$$

où

$$\bar{p}^t(i,v) = \frac{1}{n(i,v)} \sum_{j=1}^{n_t(i,v)} p^t(j,i,v)$$

si la variété est homogène, et par:

$$\hat{I}(i,v) = \frac{1}{n(i,v)} \sum_{j=1}^{n(i,v)} \frac{p^t(j,i,v)}{p^{to}(j,i,v)}$$

si la variété est hétérogène.

Des indices agrégés sont alors définis à différents niveaux géographiques et pour différentes nomenclatures de produits. On s'intéresse essentiellement aux indices annuels par *CC* et France entière en glissement de décembre à décembre, aux niveaux variété, puis postes, puis secteurs et enfin ensemble des produits.

2. PRÉCISION DES INDICES

2.1 Méthode

Le premier niveau considéré est celui de l'indice de variété par *CC*.

On utilise:

$$\hat{I}(cc,v) = \frac{1}{m(cc,v)} \sum_{i=1}^{m(cc,v)} \hat{I}(i,v),$$

où $m(cc,v)$ est le nombre total d'agglomérations dans la *CC* où la variété v est relevée. Ce nombre est presque toujours très inférieur à la taille de l'échantillon d'agglomérations. En effet, et ceci est l'un des aspects complexes du problème, on part du nombre de relevés à effectuer par poste (calculé grâce aux estimations de consommation des enquêtes Budget de Famille, et en s'appuyant sur les données de la Comptabilité Nationale), et on répartit ce nombre entre les variétés proportionnellement à la part de consommation (dit "poids") de chaque variété. Souvent, le poids de la variété est suffisamment faible pour qu'on n'ait pas la possibilité d'effectuer des relevés dans chaque agglomération. D'autre part, le nombre total de relevés étant d'environ 160 000 chaque mois, à répartir sur 1 000 variétés et un peu plus de 100 agglomérations, on peut facilement constater que le nombre de relevés espéré en moyenne par variété-agglomération est inférieur à 2. Dans ces conditions, pour peu que la variété soit concentrée dans un faible nombre de points de vente, voire ne soit pas disponible dans l'agglomération, il est possible qu'aucun relevé n'y soit effectué (produit épuisé, point de vente fermé, etc.).

La précision estimée prend la forme classique:

$$\hat{V}\,[\hat{I}\,(cc,v)] = \underset{INTER}{\hat{V}}\,(cc,v) + \underset{INTRA}{\hat{V}}\,(cc,v),$$

où

$$\underset{INTER}{\hat{V}}\,(cc,v) = \frac{1}{m\,(cc,v)}\left(1 - \frac{m\,(cc,v)}{M\,(cc)}\right)\left[s^2\,(cc,v) - \frac{1}{m\,(cc,v)}\sum_{i=1}^{m\,(cc,v)}\frac{s^2\,(i,v)}{n\,(i,v)}\right]$$

et

$$\underset{INTRA}{\hat{V}}\,(cc,v) = \frac{1}{[m\,(cc,v)]^2}\cdot\sum_{i=1}^{m\,(cc,v)}\frac{s^2\,(i,v)}{n\,(i,v)},$$

avec:

$$s^2\,(cc,v) = \frac{1}{m\,(cc,v)-1}\sum_{i=1}^{m\,(cc,v)}\left[\hat{I}\,(i,v) - \hat{I}\,(cc,v)\right]^2$$

et

$$s^2\,(i,v) = \frac{1}{[\bar{p}^{\,\omega}\,(i,v)]^2}\cdot\frac{1}{n\,(i,v)-1}\cdot\sum_{j=1}^{n\,(c,v)}\left(p^{\,t}\,(j,i,v) - \hat{I}\,(i,v)\;p^{\,\omega}\,(j,i,v)\right)^2$$

si la variété v est homogène,

$$s^2\,(i,v) = \frac{1}{n\,(i,v)-1}\sum_{j=1}^{n\,(i,v)}\left(\frac{p^{\,t}\,(j,i,v)}{p^{\,\omega}\,(j,i,v)} - \hat{I}\,(i,v)\right)^2$$

si la variété v est hétérogène.

$M\,(CC)$ est le nombre d'agglomérations dans la CC.
En CC 8 et 9, la variance inter est nulle.

Les changements de strate qui affectent les agglomérations au cours du temps ne sont pas pris en compte, compte tenu de faible nombre d'agglomérations concernées.

Pour obtenir la précision au niveau national, il est nécessaire de pondérer les variances des indices par CC-variété. Si on note $W\,(CC/v)$ le poids "économique", c'est-à-dire, en fait, la part de la consommation attribuée à la CC à variété v donnée, on a:

$$\hat{I}\,(v) = \sum_{cc} W\,(cc/v)\;.\;\hat{I}\;\;(cc,v),$$

soit:

$$\hat{v}\,(\hat{I}\,(v)) = \sum_{cc} W^2\,(cc/v)\;.\;\hat{V}\,[\hat{I}\,(cc,v)].$$

Si la variété v n'est pas relevée dans la CC, la pondération $W\,(CC/v)$ utilisée pour le calcul de l'indice est nulle. On considère cette situation comme anormale (puisque toute variété est nécessairement consommée dans toute CC), auquel cas la variance n'est pas calculée au niveau variété: par contre, on la prend en compte au niveau poste par imputation.

Si on souhaite obtenir une estimation de la précision par poste, on pondère les précisions des indices de variété par le carré des poids des variétés dans le poste. Il existe cependant un problème de covariance entre indices: si on peut considérer a priori que les indices élémentaires de deux variétés d'un même poste dans une agglomération donnée sont indépendants (les relevés l'étant), il n'en est plus de même pour les indices

synthétiques d'agglomérations de deux variétés d'une même poste. Les calculs menés sur le terme de covariance nous incitent cependant a négliger numériquement celle-ci devant les autres termes.

Ainsi, il n'apparait pas clairement qu'une agglomération plus inflationniste que la moyenne pour une variété d'un poste le soit également pour les autres variétés du même poste.

Dans ces conditions, si le poste comprend des variétés dont la précision n'était pas calculable au niveau variété, on décide, soit d'imputer la précision moyenne des variétés du poste où le calcul était possible, soit de s'en tenir à ces dernières, mais en les repondérant de façon à ce que la somme de leurs poids soit contrainte à valoir un. Cela donne deux estimations de précision, fondées sur deux traitements différents. Par pondération, on passe ensuite aux secteurs et à la précision d'ensemble.

2.2 Résultats principaux

Les calculs menés montrent que:

- La précision des indices de variété est mauvaise, voire très mauvaise. Cela s'explique par le grand nombre des variétés qui ont un poids faible, c'est-à-dire peu de relevés (souvent moins de 100). Les variétés dont les écarts types calculés sont supérieurs à 2 représentent 11% des variétés, mais seulement 2% de la pondération. L'écart-type médian est 1,1.

- La précision des indices de postes est un peu meilleure, mais reste globalement médiocre: 50% des indices de poste ont un écart-type calculé supérieur à 0,7. Ce résultat revêt une importance beaucoup plus grande que pour les variétés, parce que les indices de poste sont publiés. Les postes pour lesquels les estimateurs sont supérieurs à 1 représentent 25% des postes, mais seulement 8% de la pondération.

- Les résultats des années 1987 à 1990 sont très semblables à partir du niveau poste, et confèrent à l'indice général annuel estimé un écart-type voisin de 0,05.

Le tableau ci-dessous fournit la précision nationale par secteur en 1988, hors produits frais:

	Écart-type	Pondération (%)	Nombre de relevés (x 1 000)
Alimentation	0,061	21	43
Habillement	0,140	10	23
Autres Manufacturés	0,070	35	45
Services	0,082	34	19
Ensemble	0,042	100	130

3. OPTIMISATION DES ÉCHANTILLONS D'AGGLOMÉRATIONS

Elle s'effectue sur l'indice général en tenant compte de l'ensemble des variétés et des pondérations retenues.

Un échantillon d'agglomérations étant tiré une fois pour toutes, on cherche à minimiser un coût sous contrainte de précision en ne faisant pas intervenir les nombres de relevés qui, eux, pourront évoluer d'année en année. La fonction de coût est simplement égale au nombre total d'agglomérations tirées car il n'était pas possible d'attribuer un coût d'approche par agglomération.

On s'intéresse, dans cette résolution, au nombre d'agglomérations $m\ (CC,Z)$ à tirer par croisement grande région - CC (la grande région se note Z).

On résout donc:

$$\text{MIN} \sum_{cc,z} m\,(CC,Z)$$

sous contraintes:

$$\begin{cases} \sum_{cc,z,v} W^2\,(CC,Z,V)\ (1 - \dfrac{m\,(CC,Z)}{M\,(CC,Z)} \times \dfrac{S^2\,(CC,Z,V)}{m\,(cc,z)} = V_{ref} \\ 0 \le m\ (CC,Z) \le M\,(CC,Z). \end{cases}$$

V_{ref} est la variance calculée avec l'échantillon actuel et sur le même modèle. Les dispersions S^2 sont estimées, ce qui fait que le critère comprend la variance inter-agglomérations plus une "partie" de la variance intra-agglomérations. Si les valeurs sont manquantes parce que le nombre total de relevés pour la variété v dans le croisement CC-Z est insuffisant, on impute une dispersion égale à la moyenne des dispersions des indices calculée sur l'ensemble des grandes régions à variété - CC fixées.

Si la pondération CC-Z-Variété est nulle, on n'effectue aucune correction, ce qui équivaut à repondérer les variétés pour lesquelles l'information existe.

A cette occasion, on a replacé les agglomérations dans leurs CC actuelle (plus exactement celle du recensement 1982) pour se rapprocher des conditions qui prévaudront lors de la mise en service effective du nouvel échantillon.

Du point de vue algorithmique, on traite la contrainte d'inégalité en n'en tenant pas compte dans un premier temps. Si la taille demandée est supérieure au nombre total d'agglomérations disponibles en CC-Z, alors on sature l'inégalité et on recommence sans prendre en compte la CC-Z concernée. Cela survient seulement pour les grosses agglomérations de la région du midi-méditerranéen.

Les calculs menés sur les années 1981 à 1990 montrent une grande stabilité numérique:

Répartition optimale de l'échantillon d'agglomérations par catégorie d'agglomération sous la contrainte du maintien de la précision de l'indice de l'année

Catégories d'agglomération	1981	1982	1983	1984	1985	1986	1987
2 000 - 10 000 hab.	14	12	14	10	14	14	11
10 000 - 1 000 000 hab.	30	26	30	29	29	29	33
1 000 000 - 200 000 hab.	13	14	13	14	15	16	13
+200 000 hab.	24	22	24	23	23	23	22
Paris	1	1	1	1	1	1	1
Total	82	75	82	77	82	83	80

Le détail, pour l'année 1989, dans chaque *CC*-Z est donné par les 2 tableaux suivants, le premier étant l'échantillon actuel et le second l'optimum:

Échantillon actuel (1989):

		GRANDES RÉGIONS								
		1	2	3	4	5	7	8	9	TOTAL
	2	1	6	-	4	5	3	3	3	25
	4	2	10	3	5	4	6	5	4	39
CC	6	-	2	4	2	4	1	2	2	17
	8	1	4	4	3	2	2	4	4	24
	9									
TOTAL		4	22	11	14	15	12	14	13	105

Échantillon optimum:

		GRANDES RÉGIONS								
		1	2	3	4	5	7	8	9	TOTAL
	2	1	3	1	2	2	2	1	2	14
	4	2	7	2	3	3	5	4	3	29
CC	6	-	2	2	2	3	1	1	2	13
	8	1	3	4	3	2	2	4	5	24
	9									
TOTAL		4	15	1	10	10	10	10	12	80

Échantillon optimum: moyenne des répartitions optimales obtenues pour les indices de 1981 à 1987

GRANDES RÉGIONS

1 Région parisienne
2 Bassin parisien
3 Nord
4 Nord-Est
5 Ouest
7 Sud-Ouest
8 Centre-Est
9 Midi méditerranéen

CC: catégorie d'agglomérations

2 Agglomérations de 2 000 à 10 000 hab.
4 Agglomérations de 10 000 à 100 000 hab.
6 Agglomérations de 100 000 à 200 000 hab.
8 Agglomérations de plus de 200 000 hab.
9 Agglomération de Paris

4. OPTIMISATION DES ÉCHANTILLONS DE RELEVÉS

L'optimisation est effectuée à échantillon d'agglomérations fixé. Dans un premier temps, on cherche à minimiser la variance intra agglomérations sous contrainte de coût. Les variables principales sont les nombres de relevés $n(CC,v)$ par catégorie d'agglomération CC pour la variété v. On introduit une grille de coûts unitaires de collecte $c(CC,v)$, ainsi qu'un budget global C.

On résout:

$$\text{Min} \sum_{cc,v} W^2(CC,v) \frac{s^2(CC,v)}{n(CC,v)}$$

sous constraintes: (4.1)

$$\sum n(CC,v)\, c(CC,v) = C.$$

Lorsque l'estimation de dispersion $s^2(CC,v)$ est non renseignée, mais que le poids $W(CC,v)$ est non nul, on impute la moyenne des dispersions des différentes CC à variété v donnée. Si le poids est nul, on impose deux relevés seulement.

Il existe par ailleurs des traitements particuliers à certaines variétés. Cette optique d'ensemble ne permet malheureusement pas de contrôler les précisions par poste.

Pour y remédier, dans un deuxième temps, on peut minimiser en coût en imposant certaines limites à l'écart-type des postes.

On dresse alors la liste P des postes dont l'écart-type est supérieur à une borne $\sigma(p)$ donnée, puis on résout:

$$\text{Min} \sum_{p \in P} \sum_{\substack{cc \\ v \in p}} n(cc,v) \cdot c(cc,v)$$

sous constraintes: (4.2)

$$\forall p \in P: \sum_{\substack{cc \\ v \in p}} \frac{w(cc,v)}{w(p)} \cdot \frac{s^2(cc,v)}{n(cc,v)} \leq [\sigma(p)]^2.$$

On détermine le coût attaché aux postes de la liste P, (soit la valeur de la fonction objectif à l'optimum), et on le soustrait au budget global C. On résout enfin un programme du type (4.1) sur les autres postes, c'est-à-dire ceux qui ne font pas partie de la liste P. Cette pratique a pour contre-coup de détériorer les écart-types des "bons" postes, mais il s'agit du choix qui a été réalisé. En toute rigueur, il faudrait itérer le processus car il est possible que de (rares) postes, suite à l'utilisation de (4.2) puis de (4.1), se voient attribuer un écart-type qui dépasse la limite $\sigma(p)$ imposée.

Puis éviter les fluctuations des tailles à un niveau aussi fin que le croisement CC-Variété, on effectue un lissage en calculant la moyenne sur les trois dernières années des $n(CC,v)$ optimum.

Avec la grille de coût suivante:

Secteur	*CC*				
	2	4	6	8	9
Alimentation	3	2	2	1	1
Habillement	9	5	5	3	3
Produits manufacturés	5	3	3	2	2
Services	7	5	4	3	3

On obtient, pour un programme de type (4.1)

	Catégorie d'agglomération						Écart-type
	2	4	6	8	9	T	
RÉPARTITION ACTUELLE							
Alimentation	5	10	5	10	13	43	0.066
Habillement	2	6	3	5	7	23	0.149
Autres manufacturés	3	15	7	11	11	43	0.065
Services	2	5	3	4	5	19	0.100
Ensemble	12	32	18	30	36	128	0.044
RÉPARTITION OPTIMUM							
Alimentation	4	10	6	12	18	50	0.045
Habillement	1	5	2	6	7	21	0.104
Autres manufacturés	2	9	5	11	17	44	0.038
Services	2	4	3	6	11	26	0.047
Ensemble	9	28	16	16	35	53	0.025

La répartition optimum assure un coût (à peu près) égal au coût de collecte actuel. On rappelle que les écart-types sont des écart-types intra seulement.

Lorsqu'on utilise les programmes de type (4.2) puis (4.1), sous contrainte du nombre de relevés total constant, on obtient, pour 4 valeurs de $\sigma(p)$:

	Absence de contrainte (rappel)	Écart-type intra (poste) inférieur à:			
		0.9	0.8	0.7	0.6
ÉCART-TYPE PAR SECTEUR					
Alimentation	0.052	0.055	0.057	0.060	0.087
Habillement	0.095	0.110	0.111	0.115	0.114
Autres manufacturés	0.039	0.047	0.046	0.051	0.062
Services	0.043	0.055	0.054	0.055	0.077
Ensemble	0.025	0.030	0.030	0.040	
STRUCTURE DE L'ÉCHANTILLON (milliers de relevés)					
Alimentation	35	36	35	30	0.087
Habillement	24	24	26	37	0.114
Autres manufacturés	40	40	41	45	0.062
Service	29	28	26	16	0.077
Ensemble	128	128	128	128	

Finalement, quelle que soit l'option, on cherche à répartir les $n(CC,v)$ relevés entre un nombre maximum d'agglomérations dans la CC, mais en tenant compte de l'existant à chaque fois que c'est possible.

5. CONCLUSIONS

Considérant les tailles d'échantillons mises en jeu, et les hypothèses faites tout au long du processus du calcul, il semblerait que l'on puisse accorder une assez grande confiance aux précisions des indices de secteur et à l'indice d'ensemble, ainsi qu'à l'effectif optimum de relevés par secteur - catégorie d'agglomérations.

Dans le cadre de la rénovation de l'indice des prix à la consommation, ces résultats ont finalement deux aspects: d'une part la connaissance nouvelle d'une information touchant ce point sensible qu'est la qualité de l'indice, d'autre part un aspect opérationnel. En effet, l'optimisation des échantillons, couplée avec les estimations de précision, permet, outre la mise à jour annuelle des effectifs, la remise en cause de la nomenclature des poste et du choix des variétés sur des bases statistiques.

Cependant, on ne peut prendre en compte les précisions des postes pour l'optimisation que si l'on dispose de plusieurs années.

La seule opération qui a pu être menée sur des données antérieures à 1987 est une estimation par bootstrap sur les indices estimés d'agglomérations pour l'ensemble des produits entre 1981 et 1987. Cette estimation est trop grossière pour fournir des résultats comparables à ceux de la méthode analytique, mais suffisante pour montrer que la précision de l'indice d'ensemble ne dépend pas du niveau de l'inflation. Ce résultat, intéressant en soi, permet d'espérer une certaine stabilité dans le temps des résultats de l'optimisation.

BIBLIOGRAPHIE

Cochran, W.G. (1977). *Sampling techniques*, New York: Wiley.

INSEE, (1987). *Pour comprendre l'indice des prix*, (2e édition).

INSEE. *Bulletins mensuels de statistique.*

U.S. Bureau of labor statistics. *Item-outlet Sample redesign for the 1987 US consumer price Index revision.*

Divers travaux présentés au Séminaires sur les statistiques des prix à la consommation, GENEVE, (juin 1986).

SESSION 7

Géographie médicale

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

ÉTUDE ET CRITIQUE D'ATLAS DE MALADIES PRODUITS À TRAVERS LE MONDE

S.D. Walter[1], S.E. Birnie[1] et L.D. Marrett[2]

RÉSUMÉ

Quarante-neuf atlas de maladies ont été étudiés afin de caractériser la méthodologie employée pour produire les cartes qu'ils renferment par rapport aux populations visées, aux maladies représentées, aux techniques cartographiques et aux méthodes statistiques employées. On a trouvé peu de cohérence pour ce qui est du choix des fonctions de données à cartographier, des occurrences minimales requises, de la méthode utilisée pour standardiser l'âge ou des systèmes employés pour colorer les cartes. De nombreux atlas ne comprennent pas de renseignements descriptifs de base, on y met l'accent sur la signification statistique plutôt que sur les taux et on s'y concentre sur les risques élevés plutôt que sur les risques faibles. Peu d'atlas comprennent des données d'ordre environnemental ou une interprétation étiologique. À cause des différences d'ordre méthodologique, il est difficile de faire des comparaisons entre les atlas. Nous proposons un ensemble de lignes directrices qui pourraient être appliquées lors de la production des prochains atlas.

MOTS CLÉS: Cartographie des maladies; mortalité; morbidité; variation géographique; méthodologie.

1. INTRODUCTION

1.1 Éléments de base

On a connu récemment un renouveau d'intérêt pour la répartition géographique des maladies, ce qui a amené une prolifération d'atlas de maladies. Cette situation est probablement due, en partie, au fait que l'on a reconnu le rôle possible de l'épidémiologie descriptive pour expliquer les causes de la maladie, pour surveiller les effets des changements dans les expositions et pour planifier les services de santé et cela est imputable, en partie, au fait que tant les données que les logiciels nécessaires pour préparer les cartes et analyser les données géographiques sont facilement disponibles.

Au cours d'une étude de l'agrégation spatiale de l'incidence du cancer en Ontario, un certain nombre d'atlas de maladies ont été examinés et l'on a remarqué des variations importantes dans ce qui était cartographié et dans la façon dont ce travail avait été effectué. On a donc décidé d'essayer de repérer tous les atlas de maladies publiés récemment, de les examiner de façon systématique et de résumer leur contenu et les méthodes utilisées pour présenter les données qu'ils renferment. Nous croyons que tous les atlas nationaux et internationaux publiés au cours des quelque 15 années qui ont précédé la fermeture de l'enquête (début 1990) ont été examinés. Les atlas publiés plus de 15 ans auparavant n'ont pas toujours été inclus et, en général, on ne les a pas recherchés de façon active. De même, l'étude n'a pas porté sur les atlas qui n'ont été publiés que récemment.

---

[1] S.D. Walter et S.E. Birnie, Département d'épidémiologie clinique et de biostatistique, Université McMaster, Hamilton, (Ontario), Canada L8N 3Z5.

[2] L.D. Marrett, Unité de recherche en épidémiologie de la Fondation ontarienne pour la recherche en cancérologie et le traitement du cancer, Département de médecine préventive et de biostatistique, Université de Toronto, Toronto (Ontario), Canada M5S 1A8.

La présente communication renferme un résumé des résultats de cette enquête. Des données plus détaillées sur les méthodes utilisées et sur les résultats obtenus ont été publiées récemment (Walter et Birnie 1991).

1.2 Objectifs

L'enquête visait trois objectifs, nommément:

1. Décrire les techniques cartographiques utilisées actuellement;

2. Évaluer les possibilités qu'offrent les atlas pour fournir aux épidémiologistes des renseignements interprétables et utiles et

3. Aider à l'élaboration des atlas qui seront produits éventuellement et, pour ce faire, proposer un ensemble de lignes directrices d'ordre méthodologique.

2. MÉTHODES

Les atlas ont été repérés de diverses façons: recherches bibliographiques et contacts personnels; de plus, nous avons fait connaître, de façon informelle, notre intérêt. Les données extraites des atlas comprenaient des renseignements descriptifs de base sur les régions et les maladies cartographiées (p. ex., la taille de la population, la superficie de la région, les maladies cartographiées et le nombre de cas); les techniques cartographiques et la méthodologie statistique employées et des données et des analyses additionnelles étaient incluses. Deux d'entre nous (SW et SB) ont examiné, de façon indépendante, tous les atlas et les divergences d'opinion ont été réglées par une discussion et un réexamen. Dans la mesure du possible, un atlas a été étudié en entier. Cependant, il n'a parfois été possible d'examiner que des photocopies de certaines pages.

Quand des données nécessaires, comme le nombre de cas, la taille de la population ou la superficie de la région n'étaient pas mentionnées directement dans l'atlas, ces renseignements ont été soit déduits à partir d'autres données figurant dans l'atlas, soit obtenus d'autres sources, dans la mesure du possible.

3. RÉSULTATS

3.1 Données descriptives de base

Quarante-neuf atlas étaient inclus dans l'enquête (voir la liste complète à l'annexe A). Le tableau 1 donne la liste de tous les pays ou régions visés par ces atlas. L'Amérique du Nord, l'Europe et d'autres régions développées du monde comme le Japon, l'Australie et la Nouvelle-Zélande, sont généralement bien représentées. Certaines régions figurent même dans plus d'un atlas, le plus souvent pour des maladies ou des périodes différentes. Il existe toutefois des lacunes remarquables pour des pays d'Afrique, du Moyen-Orient, de l'Amérique Centrale, de l'Amérique du Sud et de l'Asie; ce sont des régions où il y a beaucoup à apprendre, mais où les données du genre requis peuvent ne pas être facilement disponibles.

Comme le montre nettement le tableau 2, des 49 atlas étudiés, 38 sont consacrés au cancer, un nombre plus réduit porte sur toutes les causes de décès, alors qu'un nombre encore plus réduit fait état d'autres causes de décès que le cancer. De plus, la majorité des atlas sont basés sur les décès dus à la maladie plutôt que sur son incidence.

Tableau 1: Régions qui figurent dans les atlas examinés

Atlas nationaux et internationaux		
Australie	Finlande	Portugal
Autriche	France	Écosse (2)[1]
Belgique	Italie	Espagne
Brésil	Japon	Suisse (3)[1]
Canada (3)[1]	Pays-Bas	Taïwan
Chine	Nouvelle-Zélande	Royaume-Uni (2)[1]
Danemark	Pays nordiques	États-Unis (4)[1]
Angleterre et Pays de Galles (3)[1]	Norvège	Uruguay
Communauté économique européenne	Pologne (2)[1]	Allemagne de l'Ouest (2)[1]
Intra-nationaux		
Florence (Italie)	Navarre (Espagne)	Slovakie (Tchécoslovaquie)
Frioul-Vénétie Julienne (Italie)	Québec (Canada)	Uppsala (Suède)
Isère (France)	Saskatchewan (Canada)	Victoria (Australie)

[1] Les chiffres entre parenthèses indiquent le nombre d'atlas qui comprennent cette région.

Tableau 2: Données cartographiées dans les atlas étudiés

Maladies cartographiées	Nombre d'atlas
Cancer	38
Toutes les causes	8
Autre	3
Mesure de la maladie cartographiée	
Incidence	13
Mortalité	35
Incidence et mortalité	1

Tableau 3: Caractéristiques des populations et des maladies cartographiées dans les atlas étudiés

	Minimum	Médiane	Maximum
N^bre de groupes de maladies[1]	3	18	59
Population (millions)	0.5	16	825
N^bre de cas (milliers)	7	235	6 406

[1] Ne comprend pas les "totaux" ou les "autres" causes. Chaque groupe de causes n'est compté qu'une fois même si l'on retrouve deux cartes (p. ex., des cartes établies selon le sexe).

Le nombre de groupes de maladies cartographiés (tableau 3) varie considérablement, même pour les atlas du cancer (4 à 36, pas illustrés). Il y a aussi une variation considérable dans la taille des populations qui figurent sur les cartes (de 500 000 à 825 millions de personnes) et dans le nombre de cas de maladies qui se sont produits dans les régions cartographiées (de 7 000 à 6.4 millions).

3.2 Fourniture des données de base et des caractéristiques des régions cartographiées

Il ressort clairement de l'étude du tableau 4, que plusieurs variables descriptives relativement de base ne sont pas toujours fournies. Par exemple, seulement 51% des atlas indiquent la taille de la population et seulement 29% la superficie, alors qu'entre 70% et 80% montrent le nombre de cas de maladies et le nombre de cas cartographiés.

Tableau 4: Fréquence de la mention de variables choisies dans les atlas étudiés

Variable	Précisée	Non précisée
Taille de la population	25 (51%)	24 (49%)
N[bre] de cas de maladies dans la population	39 (80%)	10 (20%)
N[bre] de cas cartographiés	35 (71%)	14 (29%)
Superficie	14 (29%)	35 (71%)

Tableau 5: Caractérisation des régions cartographiques employées dans les atlas étudiés

Variable	Minimum	Médiane	Maximum
N[bre] de régions cartographiées	4	75	3 072
Population par région (milliers)	10	240	23 800
Superficie par région (milliers de kilomètres carrés)	0.1	3	1 700
N[bre] de cas par région	235	2,804	136 307
N[bre] de cas par région pour l'événement le moins fréquent cartographié	0.1	4.0	56

Le tableau 5 décrit les divisions géographiques utilisées sur les cartes. Dans certains atlas un très petit nombre de régions sont cartographiées, alors que d'autres en renferment un grand nombre. L'atlas de la Chine (A8, Annexe A), par exemple, se trouve à l'extrémité supérieure avec 2 392 régions, alors qu'un bon nombre des atlas intra-nationaux (A41-A49, Annexe A) se trouvent à l'extrémité inférieure. Il est peut-être surprenant de constater que la plus grande population régionale moyenne est plus de 2 000 fois plus grande que la plus petite (de 23 millions dans l'atlas du Brésil à 10 000 dans celui de la Finlande (A4 et A14, Annexe A, respectivement)), et l'on relève même un écart plus grand dans la superficie moyenne par région. Pour le plus petit groupe de maladies cartographié, il y a un écart considérable dans le nombre moyen de maladies par région. La différence dans les définitions des régions et des groupes de maladies employées dans les atlas entraîne des coefficients de variation très différents pour les valeurs régionales qui figurent sur les cartes.

Les questions du nombre de régions à cartographier et de la superficie que chacune d'entre elles devrait occuper sont évidemment reliées et ne sont pas sans importance: les régions doivent être suffisamment grandes pour produire des estimations de taux stables, tout en étant assez petites pour fournir des renseignements significatifs et, quand cela est possible, pour représenter une population relativement homogène. Bien que l'on puisse s'attendre à ce qu'il existe une superficie optimale recommandée pour une région, il n'en existe pas. Il se peut que la variation considérable observée soit due à des questions sur lesquelles le cartographe n'a aucun contrôle.

Par exemple, des raisons politiques peuvent expliquer comment des régions ont été déterminées, les données utilisées au numérateur ou au dénominateur peuvent être codées ou disponibles seulement d'une certaine façon ou la qualité des données peut imposer le niveau de regroupement. En fait, on a observé dans les atlas étudiés une tendance à ce que les superficies renfermant des populations plus considérables comprennent plus de régions (ce qui n'est pas surprenant), mais aussi que la superficie et la population de ces régions soient plus petites en moyenne, que celles des régions utilisées dans les atlas portant sur des populations plus petites.

Peu d'atlas mentionnent les critères utilisés pour choisir les groupes de maladies ou les régions à cartographier. Un seul atlas mentionne les critères utilisés pour ces deux éléments, 16 atlas précisent les critères appliqués pour l'un ou pour l'autre et 45 ne précisent aucun critère. (Les atlas ont été comptés plus d'une fois s'ils incluaient plus d'un ensemble de cartes.) Quand il est mentionné, le critère utilisé pour cartographier un groupe de maladies est habituellement basé sur sa fréquence, alors que la cartographie d'une région est déterminée en fonction de sa population.

3.3 Méthodes cartographiques et statistiques

Il y a aussi une variation considérable dans le choix de la fonction ou des fonctions de données qu'il faut illustrer sur les cartes (Tableau 6). (Des atlas peuvent être comptés plus d'une fois dans ce tableau s'ils renferment des cartes de fonctions de données différentes selon, par exemple, la rareté d'une maladie.) Dans la majorité des atlas (65%), on a choisi de cartographier des taux relatifs, soit seuls, soit combinés avec d'autres fonctions, alors que pour un nombre important d'entre eux (27%), on mentionne des taux absolus. Dans quelques atlas (13%), on ne montre que des niveaux de confiance ou des fréquences relatives.

Tableau 6: Fonctions de données tracées figurant dans les atlas étudiés

Fonction des données	Nombre d'atlas[1]
Taux absolus (T)	12
Taux relatifs (T)	20
Signification (S)	6
Fréquences relatives (F)	2
T et S combinés sur la même carte	2
TR et S combinés sur la même carte	15
T et TR combinés sur la même carte	1
TR et S sur des cartes distinctes	2
T et TR sur des cartes distinctes	1
T, TR et F sur des cartes distinctes	1
	Total: 62

[1] Des atlas peuvent être comptés plus d'une fois.

La majorité des atlas présentent leurs données à l'aide de cartes en plages, c.-à-d. au moyen de hachures ou en faisant varier la couleur des régions, ce qui exige la réduction de la fonction de données en une seule dimension dans l'échelle des couleurs/des hachures. Pour obtenir ce résultat on a, le plus souvent, divisé la fonction de données soit en percentiles, soit en catégories de risque relatif sur une échelle multiplicative. La méthode des percentiles assure que l'on a un nombre prédéterminé de régions dans chacun des groupes de percentiles tels qu'utilisés, par exemple, dans l'atlas du cancer pour l'Écosse (A26, Annexe A), où l'on a cartographié les premiers 5%, les 10%, 20%, 30%, 20% et 10% suivants ainsi que les 5% les plus élevés des valeurs de la fonction. L'échelle de risque multiplicative peut mettre l'accent autant sur les écarts positifs que sur les écarts négatifs par rapport à la «moyenne», mais on peut retrouver peu de régions dans certains groupes et un bon nombre dans d'autres. Cette dernière méthode de classement est employée, par exemple, dans l'atlas du cancer pour la Finlande (A14, Annexe A). Il est intéressant de remarquer que, lorsque la signification ainsi que les taux sont illustrés, de nombreux atlas donnent la préséance à la signification statistique plutôt qu'aux valeurs des taux

et aux accroissements du risque par rapport à ses diminutions. Dans certains atlas on a aussi employé des méthodes de lissage pour éviter les problèmes associés aux petits nombres, comme cela a été fait pour la Finlande (A14, Annexe A), mais il devient alors difficile de reconnaître les entités régionales.

Dans un peu plus de la moitié des atlas étudiés, on a recours à la couleur. Bien qu'il existe une théorie de la couleur qui propose certaines classifications naturelles des couleurs, peu d'atlas suivent ces classifications, ce qui entraîne l'utilisation d'une grande variété de combinaisons de couleurs. Toutefois, certaines tendances communes au niveau de l'utilisation sont évidentes (Tableau 7). Par exemple, le plus souvent on utilise le rouge ou l'orange pour indiquer un risque élevé, alors que le vert est généralement employé pour représenter un risque faible et le blanc pour désigner une tendance neutre. Toutefois, il est intéressant de remarquer que les mêmes couleurs ont été utilisées, dans différents atlas, pour représenter des catégories de risques différentes. Par exemple, le bleu a été employé pour représenter la catégorie à faible risque dans 7 atlas, la catégorie à risque élevé dans 4 atlas et la catégorie neutre dans deux autres.

Tableau 7: Couleurs utilisées pour la cartographie dans les atlas étudiés[1]

Couleur	Nombre d'utilisations pour désigner la:		
	Catégorie à risque le plus élevé	Catégorie neutre	Catégorie à faible risque
Rouge	26	0	0
Orange	6	1	1
Brun	2	2	1
Bleu	4	2	7
Vert	1	4	21
Jaune	0	8	6
Blanc	0	10	5

[1] Seuls les atlas dans lesquels on utilise la couleur figurent dans ce tableau. Les atlas peuvent être comptés plus d'une fois si l'on y emploie diverses combinaisons de couleurs.

Dans la majorité des atlas, on utilise la standardisation indirecte de l'âge avec une population-type interne (Tableau 8). En général, les atlas dans lesquels on utilise la standardisation indirecte portent sur des populations totales plus petites, bien que la population moyenne et le nombre de cas par région soient, en fait, plus élevés que dans les régions où l'on a recours à la standardisation directe. Cela est plutôt surprenant, compte tenu du fait que les petites fréquences régionales constituent la principale raison pour laquelle on a recours à la standardisation indirecte plutôt qu'à la standardisation directe.

Tableau 8: Méthode de standardisation de l'âge et population-type utilisées dans les atlas étudiés[1]

Genre de population-type	Méthode de standardisation de l'âge	
	Directe	Indirecte
Interne	10	24
Externe	9	3
Non précisée	1	0

[1] Certains atlas ne sont pas inclus parce que la méthode de standardisation employée n'est pas précisée.

3.4 Fourniture de données et d'analyses additionnelles

Dans la majorité des atlas on trouve des fréquences régionales et (ou) des taux sous forme soit de tableaux imprimés, soit de microfiches (Tableau 9). Dans un certain nombre d'atlas, on estime aussi le nombre de personnes et de régions selon le niveau de risque. C'est seulement dans la minorité des atlas que l'on trouve des données régionales sur des variables qui peuvent être liées à la santé comme la densité de la population, les facteurs climatiques et le statut socio-économique. Alors que le but le plus souvent déclaré pour la production d'un atlas est relié à l'identification d'agents étiologiques (habituellement par élaboration d'hypothèses), les données sont rarement fournies pour faciliter cette tâche au lecteur ou même pour la lui rendre possible, sans que ce dernier ait à recourir à des données supplémentaires tirées d'autres sources. Moins de 50% des atlas étudiés renferment une interprétation quelconque des cartes et seulement trois comprennent une analyse spatiale en bonne et due forme.

Tableau 9: Tableaux ou cartes supplémentaires inclus dans les atlas étudiés

Tableaux	**Nombre d'atlas (% sur 49)**
Fréquence des cas	39 (80%)
Taux régionaux	45 (92%)
Nombre de régions ou de personnes selon le niveau de risque	16 (33%)
Cartes	
Densité de la population	12 (24%)
Climat	5 (10%)
Caractéristiques physiques	5 (10%)
Statut socio-économique	4 (8%)
Origine ethnique	3 (6%)
Géologie	3 (6%)
Autre	5 (10%)

4. CONCLUSIONS ET RECOMMANDATIONS

On admet généralement que les atlas peuvent être une source de renseignements interprétables et utiles pour les épidémiologistes. Cette opinion est renforcée par le fait que l'on a réalisé des études exhaustives d'hypothèses étiologiques fondées sur des renseignements contenus dans des atlas. Toutefois, les atlas n'ont pas encore atteint leurs possibilités maximales, partiellement à cause du manque de normes ainsi que de méthodes et de présentations de haute qualité et, partiellement, parce qu'on n'a généralement pas fourni suffisamment de renseignements et d'analyses avec les cartes des maladies pour permettre et encourager l'interprétation. Comme étape en vue d'accroître la comparabilité et l'utilité des atlas de maladies pour les épidémiologistes, nous proposons, au tableau 10, un ensemble de lignes directrices.

Bien que cette liste soit plutôt longue, les quatre derniers éléments («Renseignements ou analyses supplémentaires») sont moins importants que les autres. De plus, nous admettons que tous les atlas ne peuvent pas fournir le volume de renseignements qui figurent dans la section «Données»: ces éléments pourraient être considérés comme souhaitables, mais non essentiels. L'acceptation générale de lignes directrices normalisées et la fourniture d'un nombre accru d'analyses et de renseignements connexes pourraient augmenter considérablement l'utilité des atlas de maladies pour les épidémiologistes.

Tableau 10: Liste proposée d'éléments à inclure dans les atlas de maladies

A. Renseignements de base:

1. Un énoncé clair de l'objet de l'atlas;
2. Un commentaire sur la qualité des données, y compris leur variation régionale;
3. La justification du choix des unités régionales cartographiées, et des renseignements sur la stabilité des données régionales;
4. La précision des critères utilisés pour cartographier des valeurs régionales (p. ex., fréquences minimales);
5. La précision des critères utilisés pour choisir et grouper des maladies à des fins de cartographie;
6. La précision de la méthode de standardisation de l'âge (et du groupement des âges) employée. Nous recommandons l'emploi de la méthode de standardisation directe avec une population-type interne, afin de faciliter les comparaisons entre les régions dans un même atlas. Nous recommandons aussi l'adoption de groupes d'âge normalisés pour tous les atlas.

B. Données:

7. Les fréquences, selon la région, des maladies et des populations cartographiées;
8. Les valeurs régionales pour la fonction de données cartographiée ainsi que sa précision;
9. Les fréquences et les populations selon l'âge par région (pour permettre des calculs qui faciliteraient les comparaisons entre les atlas);

C. Cartes:

10. Une carte-index pour permettre de trouver et d'identifier des régions particulières;
11. Des cartes de taux d'événements plutôt que de fréquences. Le fait de cartographier des niveaux significatifs ne devrait pas exclure la représentation des taux;
12. Une distinction claire entre les taux et le niveau significatif quand une fonction de données combinée est utilisée;
13. Une combinaison logique des couleurs;
14. Des groupes de traçage symétriques, où l'accent est mis autant sur le risque élevé que sur le risque faible et sur le niveau significatif dans l'une ou l'autre des deux directions;

D. Renseignements ou analyses supplémentaires:

15. Une analyse des tendances chronologiques;
16. Une analyse en bonne et due forme de la structure spatiale;
17. Des cartes des éléments environnementaux et du style de vie liés à la santé, de préférence en utilisant les mêmes régions que pour les cartes de maladies;
18. Un commentaire analytique sur les résultats, même si ce n'est que pour postuler des hypothèses afin d'expliquer la variation régionale observée.

REMERCIEMENTS

SDW reçoit une bourse de chercheur national en santé de Santé et Bien-être social Canada. Ce projet a aussi été appuyé, en partie, par une subvention du ministère de la Santé de l'Ontario.

Les auteurs voudraient remercier les personnes mentionnées ci-après qui les ont aidés à trouver plusieurs des atlas ainsi que pour la traduction: M. Tom Broz, Fondation ontarienne du cancer; Docteur Jacques Estève, CIRC, Lyon; Docteur R. Frentzel-Beyme, Deutches Krebsforschungszentrum, Heidelberg; Docteur Yang Mao, Santé et Bien-être social Canada, Ottawa; Fumihisa Matsumoto, Chiba, Japon; Docteur Chrisoph Minder, Université de Berne, Suisse; Docteur C.S. Muir, CIRC, Lyon; Docteur R. Semenciw, Santé et Bien-être social Canada.

BIBLIOGRAPHIE

Walter SD, Birnie SE (1991). Mapping mortality and morbidity patterns: An international comparison, *Int J Epidemiol*, 20, 678-689.

ANNEXE A. Liste des atlas étudiés

A1. Giles, G.G., Armstrong, B.K., Smith, L.R. (Eds) (1987). Cancer in Australia 1982. National Cancer Statistics Clearing House, Scientific Publication No. 1.

A2. Austrian Central Statistical Office (1989). Osterreichischer Todesursachenatlas 1978/1984. Vienna: Austrian Central Statistical Office.

A3. Ryckeboer, R., Janssens, G., Thiers, G.L. (1983). Atlas de la mortalité par cancer en Belgique, 1969-1976. Bruxelles: Institut d'hygiène et d'épidémiologie, Ministère de la Santé publique et de l'Environnement.

A4. Brumini, R. (ed) (1982). Cancer no Brasil: dados Histopatológicos 1976-80. Rio de Janiero: Campanha Nacional de Combate as Câncer, Ministerio de Saude.

A5. Santé et Bien-être social Canada (1980). Répartition géographique de la mortalité au Canada. Volume 1: cancer. Hull: Centre d'édition du gouvernement du Canada.

A6. Santé et Bien-être social Canada (1980). Répartition géographique de la mortatlié au Canada. Volume 2: mortalité générale. Hull: Centre d'édition du gouvernement du Canada.

A7. Santé et Bien-être social Canada (1984). Répartition géographique de la mortalité au Canada. Volume 3: mortalité en milieu urbain. Hull: Centre d'édition du gouvernement du Canada.

A8. The editorial committee for the Atlas of Cancer Mortality in the People's Republic of China (1979). Atlas of cancer mortality in the People's Republic of China. Beijing: China Map Press.

A9. Carstensen, B., Møller J.O. (1986). Atlas over Kroeftforekomst i Denmark 1970-79. Danish Cancer Registry, Danish Cancer Society, Environmental Protection Agency.

A10. Gardner, M.J., Winter, P.D., Taylor, C.P., Acheson, E.D. (1983). Atlas of cancer mortality in England and Wales 1968-78. Chichester: John Wiley and Sons.

A11. Gardner, M.J., Winter, P.D., Barker, D.J.P. (1984). Atlas of mortality from selected disease in England and Wales 1968-78. Chichester: John Wiley and Sons.

A12. Department of Health and Social Security (1988). Outcome indicators - avoidable deaths. In *On the State of the Public Health*, the annual report of the Chief Medical Officer of the Department of Health and Social Security. 74-82. London: Her Majesty's Stationery Office.

A13. Holland, W.W. (Ed.) (1988). European Community Atlas of "Avoidable Death". Oxford: Oxford University Press.

A14. Pukkala, E., Gustavsson, N., Teppo, L. (1987). Atlas of cancer incidence in Finland 1953-1982. Helsinki: Cancer Society of Finland Publication No. 37.

A15. Rezvani, A., Doyon, F., Flamant, R. (1985). Atlas de la mortalité par cancer en France (1971-1978). Paris: Les éditions Inserm.

A16. Cislaghi, C., DeCarli, A., LaVecchia, C., Laverda, N., Mezzanotte, G., Smans, M. (1986). Data, statistics and maps on cancer mortality, Italia, 1975/1977. Bologna: Pitagora Editrice.

A17. Segi, M. (1977). Atlas of cancer mortality for Japan by cities and counties 1969-71. Tokyo: DAIWA Health Foundation.

A18. Netherlands Central Bureau of Statistics (1980). Atlas of cancer mortality in the Netherlands 1969-1978. The Hague: Staatsuitgeverij.

A19. Borman, B. (1982). A cancer mortality atlas of New Zealand. Wellington: National Health Statistics Centre, Department of Health, Special Report No. 63.

A20. Møller J.O., Carstensen, B., Glattre, E., Malker, B., Pukkala, E., Tulinius, H. (1988). Atlas of cancer incidence in the Nordic Countries. Helsinki: Nordic Cancer Union.

A21. Glattre, E., Finne, T.E., Olesen, O., Langmark, F. (1986). Atlas over Kreftinsidens I Norge. 1970-79. Oslo: Norwegian Cancer Society.

A22. Staszewski, J. (1976). Epidemiology of cancer of selected sites in Poland and Polish Migrants. Ballinger Publishing Company, Cambridge, Mass.

A23. Zatonski, W., Becker, N. (1988). Atlas of cancer mortality in Poland 1975-1979. Paris: Springer-Verlag.

A24. DaMotta, L.C., Falcao J.M. (1987). Atlas Do Cancro Em Portugal 1980-1982. Ministerio Da Saude, Departamento de Estudos e Planeamento da Saude, Lisboa.

A25. Lloyd, O.L., Williams, F.L.R., Berry W.G., du V. Florey, C. (1987). An atlas of mortality in Scotland. Including the geography of selected socio-economic characteristics. Croom Helm, London.

A26. Centre international de recherche sur le cancer (1985). Atlas of cancer in Scotland; 1975-1980. Incidence and Epidemiological Perspective. Lyon: IARC Scientific Publication No. 72.

A27. López-Abente, G., Escolar, A., Errezola, M. (1984). Atlas del cáncer en espana. Vitoria-Gasteiz: Gráficas Santamaria.

A28. Brooke, E. (1976). Géographie de la mortalité due au cancer en Suisse 1969-71. Berne: Institut universitaire de médecine sociale et préventive.

A29. Office Fédéral de la Statistique (1987). La Distribution géographique de la mortalité cancéreuse en Suisse. 1979/81. Berne: Office Fédéral de la Statistique.

A30. Bisig, B., Paccaud, F. (1987). Répartition géographique des principales causes de décès en Suisse 1969/1972, 1979/1982. Berne: Office Fédéral de la Statistique.

A31. Chen, K-P., Wu, H-Y., Yeh, C-C., Cheng, Y-J. (1979). Color atlas of cancer mortality by administrative and other classified districts in Taiwan area: 1968-1976. Taipei: National Science Council, Taiwan Republic of China.

A32. Howe, M. (1963). National atlas of disease mortality in the United Kingdom. London: Nelson.

A33. Howe, M. (1970). National atlas of disease mortality in the United Kingdom, 95-189. London: Nelson.

A34. Mason, T.J., McKay, F.W., Hoover, R., Blot, W.T., Fraumeni, J.F. Jr. (1976). Atlas of cancer mortality for U.S. counties: 1950-1969. Washington, D.C.: DHEW Publication (NIH) 75-780, U.S. Government Printing Office.

A35. Mason, T.J., Fraumeni, J.F. Jr., Hoover, R., Blot, W.J. (1981). An atlas of mortality from selected diseases. Washington D.C. DHHS Publication (NIH) 81-2397, U.S. Government Printing Office.

A36. Riggan, W.B., Creason, J.P., Nelson, W.C. et coll. (1987). U.S. cancer mortality rates and trends, 1950-1979. Volume IV: Maps. Washington, D.C.: United States Environmental Protection Agency.

A37. Pickle, L.J., Mason, T.J., Howard, N., Hoover, R., Fraumeni, J.F. Jr.(1987). Atlas of U.S. cancer mortality among whites, 1950-1980. Washington, D.C.: DHHS Publication (NIH) 87-2900, U.S. Government Printing Office.

A38. Vassallo, J.A. (1989). Registro Nacional de cancer del Uruguay. Cancer en el Uruguay. Montevideo, Dirección del Registro Nacional de Cáncer.

A39. Frentzel-Beyme, R., Leutner, R., Wagner, G., Wiebelt, H. (1979). Cancer atlas of the Federal Republic of Germany. Berlin: Springer-Verlag.

A40. Becker, N., Frentzel-Beyme, R., Wagner, G. (1984). Atlas of cancer mortality in the Federal Republic of Germany. Berlin: Springer-Verlag.

A41. Geddes, M., Vigotti, M.A., Biggeri, A., Cervellini, D., Salvadori, P. (1985). Atlante della mortalità per tumori nella provincia di Firenze: 1971-1979. Firenze: Notiziariodella Sezione Fiorentinadella Lega Italiana per la Lotta contro i Tumori.

A42. Franceschi, S., Meneghel, G., Mezzanotte, G. et coll. (1986). Atlas of cancer mortality in the Friulia-Venezia-Giulia Region, 1975-1977. Aviano: Centro di riferimento oncologico.

A43. Menegoz, F., Colonna, M., Lutz, J.M., Schaerer, R. (1989). Atlas du cancer dans le départment de l'Isère. Registre du Cancer de l'Isère, Grenoble.

A44. Vicente, J.A., Aranzadi, A.A., Elizaga, N.A. (1987). Cáncer en Navarra 1973-1982. Pamplona: Servico Regional de Salud.

A45. Ghadirian, P., Thouez, J.P., PetitClerc, C., Rannou, A., Beaudoin, Y. (1989). L'incidence des cancers: Atlas de la province de Québec 1982-1983. Montréal, Société de recherche sur le cancer Inc.

A46. Saskatchewan Cancer Foundation (1988). Saskatchewan cancer atlas 1970-1987. Saskatchewan Cancer Foundation, Cancer Registry Report.

A47. Plesko, I., Dimitrova, E., Somogyi, J. et al. (1989). Atlas vyskytu Zhubných Nádoror V SSR. Bratislava: Veda Vydavatel'stvo Slovenskej Akadémie Vied.

A48. Isaksson, H-O., Hesselius, I. (1989). Cancerutvecklingen i Uppsala/Orebroregionen. Uppsala: Regionalt Onkologiskt Centrum.

A49. Giles, G., Jolley, D., Lecatsas, S., Handsjuk, H. (1988). Atlas of cancer in Victoria. Incidence 1982-1983, Mortality 1979-1983. Melbourne, Victoria: Victorian Cancer Registry.

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

SURVOL DES MÉTHODES ANALYTIQUES ET DES TECHNIQUES DE PRÉSENTATION EN GÉOGRAPHIE MÉDICALE

G.J. Sherman[1]

RÉSUMÉ

Les premiers résultats positifs en épidémiologie ont souvent résultés de l'étude de la distribution spatiale de la maladie. Une des premières analyses, qui est souvent citée, a été celle effectuée par John Snow qui a tracé sur une carte les cas de choléra à Londres au XIX^{e} siècle et qui a réussi à identifier un puits contaminé comme source de l'épidémie. Bien que les techniques analytiques aient fait des progrès énormes depuis ce temps, une bonne partie du travail effectué aujourd'hui en géographie médicale par les épidémiologistes est essentiellement de même nature. Les cartes de points établies pour les cas et les cartes de régions dressées pour les taux sont encore très populaires, en partie à cause de l'attrait visuel de toute image comparativement à un tableau, mais devrait-on encore considérer ces cartes suffisantes? L'aptitude du cerveau humain est-elle surfaite quand il s'agit de résoudre visuellement un problème d'identification d'une structure? Quel a été le succès des cartes de maladie pour ce qui est de "fournir des indices en matière d'étiologie"?

Divers outils et techniques cartographiques sont mentionnés, ainsi qu'un certain nombre de stratégies auxiliaires et de suivi.

MOTS CLÉS: Distributions spatiales des maladies; cartogramme; espaces relatifs; analyse d'agrégation.

1. INTRODUCTION

Certains succès initiaux de l'épidémiologie peuvent être attribués à l'étude des distributions spatiales des maladies. Citons à titre d'exemple l'analyse de John Snow qui, en portant sur une carte les cas de choléra de Londres au XIX^{e} siècle, a pu isoler un puits contaminé comme source de la maladie. Bien que les techniques d'analyse cartographique aient beaucoup changé depuis, le principe demeure le même: la proximité spatiale de cas indique l'existence d'un facteur de risque commun. L'observation d'une telle proximité peut contribuer à la compréhension de l'étiologie de la maladie, ou à tout le moins fournir des pistes pour la poursuite des études. L'étude des distributions spatiales est relativement directe, voire facile, et constitue un bon point de départ pour expliquer le lien exposition-résultat.

Dans la présente communication, je mets l'accent sur les techniques plutôt que sur les résultats. Qui plus est, en raison des contraintes manifestes de temps et comme la communication est orale, il sera à peine ou pas du tout question de nombreuses techniques spécifiques. Les techniques mentionnées le seront en raison de leur commodité, de leur utilité ou de leur acceptation, et il ne faut y voir aucune recommandation quant à leur utilisation dans une situation donnée: les techniques sont certes importantes, mais ne sont tout de même qu'un élément du processus d'étude.

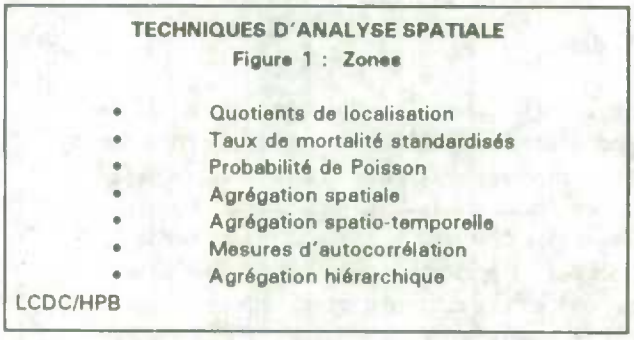
Un nombre restreint mais croissant d'études épidémiologiques sont axées sur la distribution géographique de la maladie. La plupart se fondent sur des cartes géopolitiques qui déterminent les zones de manifestation élevée

---

[1] G.J. Sherman, Santé et bien-être social Canada, division des maladies de nourrissons et d'enfants, Parc Tunney, LCDC, édifice #6, pièce 28C, Ottawa (Ontario), Canada K1A 0L2.

et basse, au moyen de taux ajustés en fonction de la race, du sexe, de facteurs spécifiques et parfois de l'âge. Habituellement, l'interprétation épidémiologique ou l'analyse statistique s'arrête à ce point. Il ne faut toutefois pas en conclure qu'on ne dispose pas d'un impressionnant ensemble de techniques. Gesler, dans une communication de 1986 qui m'a beaucoup aidé à mettre de l'ordre dans mes idées, répartit les techniques d'analyse spatiale en six catégories, qui portent respectivement sur des points, des lignes, des zones, des surfaces, la comparaison de cartes et des espaces relatifs. Je m'inspirerai abondamment de cette communication et lui emprunterai son plan.

Zones



Il est possible d'établir des cartes des zones de la maladie de bien des façons, et ce type de représentation est un des plus communs. Les plus simples sont basées sur les discontinuités naturelles des distributions discrètes ou encore sur la moyenne et l'écart type de distributions, et sont essentiellement descriptives. Les méthodes comme les quotients de localisation et les taux standardisés de mortalité (ou morbidité) sont des formes d'analyse simples courantes et peuvent être davantage utiles pour l'étude de l'évolution de la maladie. Dans plusieurs cas, on s'est servi de la distribution de Poisson pour déterminer des sous-zones où les taux de morbidité étaient significativement élevés ou bas. Un exemple bien connu est l'Atlas de la mortalité du Canada, paru en quatre volumes et produit en collaboration par Santé et bien-être et Statistique Canada.

Un facteur limitatif majeur du report de taux de morbidité sur des cartes géopolitiques est le fait que les sous-unités géographiques comme les comtés ne sont pas définis par rapport à la population à risque à l'étude. De grandes zones faiblement peuplées tendent à dominer visuellement la carte, alors que l'attention devrait être concentrée sur les zones fortement peuplées. L'analyse statistique rigoureuse des cartes fondées sur des limites géopolitiques est complexe pour la même raison, c'est-à-dire que les sous-unités géographiques renferment souvent des populations à risque considérablement différentes. L'analyse statistique des cartes géopolitiques consiste habituellement à appliquer un test de signification afin d'identifer les taux élevés vraisemblablement attribuables à des influences autres que de simples fluctuations aléatoires. Une autre façon d'éliminer les fluctuations aléatoires associées aux petites populations à risque est de fondre les zones faiblement peuplées en grandes unités géographiques, mais cette méthode a pour effet de diminuer la spécificité de l'analyse géographique.

Les données idéales pour l'analyse géographique devraient renfermer la localisation des cas de la maladie à l'étude. Mais les lieux précis sont rarement indiqués, la meilleure indication étant le code postal complet. Naturellement, les dénominateurs démographiques par âge et par sexe doivent être définis pour les mêmes unités géographiques.

Selvin et coll. ont décrit une technique de cartographie informatisée qui transforme la carte géopolitique originale (ils se sont servis des secteurs de recensement de la ville et du comté de San Francisco) en cartogramme, c.-à-d. en une carte qui déforme le lieu et la zone géographique de sorte que soit égalisée la densité d'une quelconque autre grandeur. Des méthodes statistiquement rigoureuses permettent ensuite de repérer les agrégats spatiaux non aléatoires. Dans leur exemple, le cartogramme a été défini de façon à égaliser la densité de la population à risque. Sur une carte géopolitique ordinaire, la plupart des inégalités de risque entre les différentes zones géographiques découlent des inégalités de la densité de population. La transformation égalise donc la densité de population et facilite ainsi le repérage et l'analyse des autres facteurs influant sur la distribution de la maladie.

La technique du cartogramme a deux avantages sur les méthodes classiques: 1) l'analyse préserve entièrement le détail géographique et élimine la nécessité de combiner arbitrairement des zones comptant de faibles populations à risque; et 2) la transformation en cartogramme préserve les relations d'adjacence, de sorte que les renseignements provenant de zones adjacentes peuvent être visuellement ou mathématiquement intégrés dans le modèle d'interprétation.

Il existe un certain nombre d'algorithmes publiés et disponibles pouvant servir à la transformation cartographique. Cependant, peu sont conviviaux.

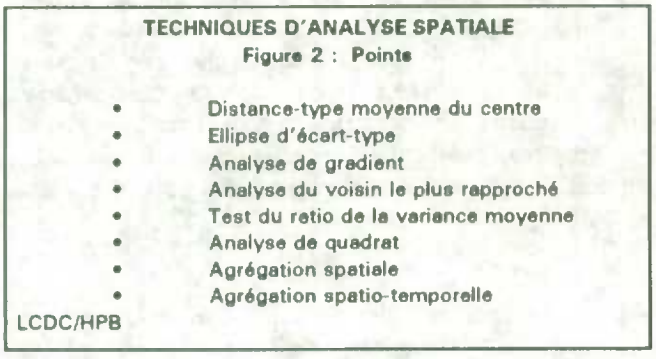
Les études de la distribution spatiale des cas d'une maladie sont généralement faciles d'exécution, sont relativement peu coûteuses et constituent un bon point de départ dans la compréhension de l'étiologie de la maladie, particulièrement quand on soupçonne que des agents environnementaux sont des facteurs de risque. Des facteurs de confusion possibles comme la consommation de tabac, la condition socio-économique et l'accès à des soins médicaux, entre autres, peuvent toutefois être encore à la source d'évolutions non aléatoires observées. La carte à densité égalisée n'élimine en effet qu'un seul facteur de confusion, soit la distribution de la population.

Des tests d'agrégation spatiale de la maladie et des méthodes d'agrégation spatio-temporelle ont été mis au point pour l'étude des données spatiales. Pour surmonter les problèmes de latence et de mobilité dans l'étude de l'issue des maladies chroniques comme la maladie de Hodgkin, des chercheurs se sont servis de méthodes cas-témoins, c'est-à-dire qu'ils ont établi des ratios d'incidence approchés dont ils ont ensuite testé la signification de manière standard, pour chacune des zones à l'étude.

Diverses mesures d'autocorrélation spatiale, notamment la statistique I de Moran, ont servi à l'étude de l'évolution des maladies. Glick a établi plusieurs méthodes permettant d'appliquer la statistique d'autocorrélation de Moran pour données d'intervalle à l'examen de l'évolution spatiale de la maladie et à la recherche de facteurs biologiques, chimiques, physiques, culturels et ethniques susceptibles d'être associés à cette évolution. Les facteurs de pondération utilisés pour calculer la statistique I de Moran peuvent être fondés sur l'adjacence simple d'unités géographiques, sur les proportions de limites communes, sur le fait qu'une quelconque variable de catégorisation présente ou non une valeur identique dans deux unités spatiales, ou sur la distance entre les centres des unités. Des corrélogrammes spatiaux peuvent être établis pour mesurer l'autocorrélation à divers écarts spatiaux. Glick a été jusqu'à étudier les tendances de la fonction d'autocorrélation sur des transects linéaires et à examiner les résidus de modèles de tendance au moyen de cette méthode.

Grimson et ses collaborateurs ont mis au point une méthode de détermination des agrégats hiérarchiques de zones à risque élevé par examen des adjacences entre paires de taux ordonnés. Le nombre observé d'adjacences est comparé aux résultats de simulations de Monte Carlo qui établissent les probabilités de la présence de joints pour un nombre approprié d'unités. Grimson privilégie la simulation de Monte Carlo dans les cas où les données spatiales ne sont pas indépendantes et où les unités spatiales sont de forme irrégulière, deux conditions extrêmement fréquentes.

Points



Les cartes ponctuelles ont probablement été utilisées aussi fréquemment que les cartes spatiales et précèdent certainement ces dernières dans l'étude épidémiologique de la maladie, à preuve John Snow que j'ai déjà mentionné. De toutes les techniques analytiques élaborées pour les cartes ponctuelles (distance-type, ellipses d'écart-type, voisin le plus rapproché, quadrat, etc.), les épidémiologistes ont privilégié l'examen de la distribution des points à la recherche d'agrégats possibles. Plusieurs méthodes d'agrégation différentes ont été élaborées, mais celles qui prennent en compte le passage du temps ont reçu une attention considérable. On attribue généralement à Knox le crédit de la notion d'agrégation spatio-temporelle. C'est lui en effet qui a énoncé que la détection de l'épidémicité dans un ensemble de données dépend d'une distribution temporelle, d'une distribution spatiale ainsi que des interactions entre le temps et l'espace. Pour l'étude de ces interactions, on examine la proximité dans l'espace de paires de cas relativement voisins dans le temps. Les paires sont classées en fonction des deux critères puis servent à construire une table de contingence. Les fréquences des paires observées peuvent alors être comparées à des valeurs prévues au moyen d'une formule de distribution sur intervalles de temps. Comme je l'ai mentionné, l'agrégation spatio-temporelle peut également être appliquée aux données spatiales.

Lignes

TECHNIQUES D'ANALYSE SPATIALE
Figure 3 : Lignes

- Marche aléatoire
- Vecteurs
- Théorie des graphes
  - Nodalité
  - Connectivité
  - Dispersion
  - Hiérarchies nodales
  - Analyses des déplacements

LCDC/HPB

L'ensemble de techniques analytiques que Gesler qualifie de «unidimensionnelles» ou «linéaires» a très peu été utilisé pour l'étude de la maladie. Je n'en connais pas la raison, si ce n'est que ces techniques sont davantage purement mathématiques plutôt que statistiques et qu'il existe très peu de logiciels conviviaux permettant de les exécuter.

La notion de marche aléatoire a servi à analyser le déplacement du «front clinique» de la maladie. L'idée est de comparer le sens réel du déplacement de la maladie à un déplacement aléatoire. Les écarts peuvent être indicateurs de certaines contraintes non aléatoires ou de paramètres environnementaux présents en certains endroits.

Les géographes médicaux se sont aussi servis dans une certaine mesure de la théorie des graphes et de l'analyse des réseaux dans des recherches sur la maladie et la prestation de soins médicaux. Pour ce qui est des recherches sur la maladie, on a élaboré des réseaux dans les études de diffusion qui indiquent les divers types de relations (appelées joints) entre les unités spatiales étudiées. Les réseaux de telles relations constituent une façon commode d'illustrer certains processus et ne commandent pas nécessairement une analyse de paramètres tels la nodalité ou la connectivité. Ces travaux me sont toutefois si étrangers que je dérogerai à ma présentation en citant simplement ici, à titre d'exemple, les travaux de Haggett qui a construit sept graphes différents pour représenter sept modèles de diffusion possibles de la rougeole dans le sud-ouest de l'Angleterre: diffusion régionale, diffusion urbaine-rurale, contagion locale, contagion par vagues, contagion sur le trajet au travail, diffusion en fonction de la taille de la population et diffusion en fonction de la densité de population.

La théorie des graphes présente des applications manifestes dans le cas des systèmes de diffusion, et il ne surprend donc pas qu'elle a été le plus utilisée dans les études sur les déplacements dans les hôpitaux, la progression de maladies, la modélisation de localisation/allocation, etc.

Surfaces

TECHNIQUES D'ANALYSE SPATIALE
Figure 4: Surfaces

- Isolignes
- Établissement de surfaces de dérive
  - Polynômes en série de puissances
  - Série de Fourier

LCDC/HPB

Il est possible de construire une surface ou un champ scalaire en trois dimensions par recours à une variable "z" ou "hauteur". On peut réduire la même variable en coordonnées bidimensionnelles et tracer des isolignes. L'établissement de surfaces de dérive est une technique bien connue, mais elle n'a pas été beaucoup utilisée et l'a été surtout pour l'analyse des processus de diffusion de la maladie, c.-à-d. essentiellement dans les mêmes conditions que la théorie des graphes unidimensionnelle, en offrant toutefois l'avantage (pour la plupart des personnes) d'une présentation visuelle davantage intuitive.

Comparaison de cartes

TECHNIQUES D'ANALYSE SPATIALE
Figure 5: Comparaison de cartes

- Courbes de Lorenz
- Coefficient de correspondance spatiale
- Coefficient de corrélation
- Cartes de différences

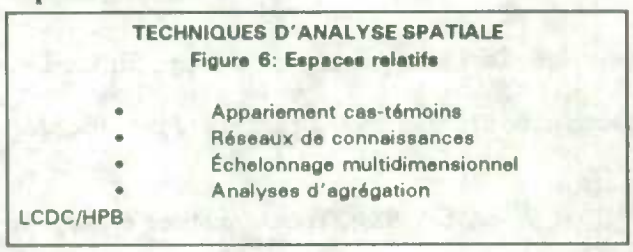
LCDC/HPB

La comparaison de cartes est une technique qui a gagné en popularité et en accessibilité ces récentes années grâce à l'apparition dans le commerce de puissants micro-ordinateurs et logiciels. Bien sûr, la comparaison de cartes existait avant les micro-ordinateurs. Les variables dépendantes et indépendantes étaient simplement portées sur des cartes différentes à la même échelle, puis ces dernières étaient comparées visuellement. Il s'agit là d'ailleurs, à mon avis, essentiellement de ce que nous faisons pour la plupart quand on nous présente des cartes spatiales de taux de morbidité.

La méthode statistique de comparaison de cartes probablement la plus utilisée est l'analyse de corrélation ou la «corrélation écologique» fondée sur une technique de corrélation paramétrique ou basée sur les statistiques d'ordre (non paramétrique). Cela semble être l'approche de choix (outre les calques et l'ombrage thématique) de la plupart des logiciels existants de cartographie interactive tournant sur micro-ordinateurs. De fait, cette approche permet un haut niveau d'élaboration statistique (agrégation hiérarchique, analyse factorielle pour expliquer les relations sous-jacentes entre variables et pour réduire le nombre de variables, etc.) dans la préparation des mesures ou des indices qui sont ensuite comparés.

Un autre type de technique de comparaison de cartes utilisé dans une moindre mesure par les géographes médicaux est basé sur le coefficient de correspondance d'aires, qui est le ratio de l'aire sur laquelle sont localisés simultanément deux phénomènes et de l'aire totale couverte par les deux phénomènes (soit le ratio de l'intersection de deux phénomènes et de leur union).

Espaces relatifs



Gatrell désigne les analyses dimensionnelles dont il a été question jusqu'à maintenant sous le vocable d'étude d'arrangements spatiaux, car ces analyses sont fondées sur les propriétés métriques de la distance. Il laisse entendre qu'il ne s'agit là que du début de l'analyse spatiale et qu'il nous faut maintenant nous pencher sur les espaces relatifs et les relations non métriques entre des ensembles d'objets.

L'échelonnage multidimensionnel (EMD) est un des outils de l'arsenal des «espaces relatifs». L'EMD est en réalité une classe de techniques pour lesquelles les proximités entre objets constituent les données d'entrée. Par proximité, il faut entendre tout nombre qui indique le degré réel ou perçu de ressemblance ou de dissemblance entre deux objets, ou toute mesure du genre. Le principal extrant est une représentation spatiale consistant en une configuration géométrique de points, comme sur une carte. Cette configuration reflète la «structure cachée» des données et souvent, après mûre réflexion, en facilite la compréhension.

L'analyse d'agrégation (qui n'est pas la même chose que l'agrégation de points ou que l'agrégation spatio-temporelle) est une technique de catégorisation basée sur la «distance taxonomique» entre les données. Les techniques de l'analyse d'agrégation, nombreuses, servent à explorer les données produites par la mesure d'un nombre de caractéristiques pour chacun des individus ou des objets d'un ensemble assorti. Le but de l'exploration est de déterminer si les objets peuvent être subdivisés en groupes relativement distincts ou apparentés. Ce but est manifestement différent de celui de l'analyse discriminante ou de techniques d'attribution semblables qui servent à attribuer des objets à des groupes connus. L'analyse d'agrégation a un objet plus difficile et intrinsèquement plus intéressant, celui de trouver les groupes en premier lieu.

2. SOMMAIRE et CONCLUSIONS

S'il est vrai qu'une image vaut mille mots, il est aussi vrai que la plupart des gens surestiment grossièrement leur capacité d'analyse visuelle. Certains peuvent jouer simultanément une douzaine de parties d'échecs à l'aveugle. Peut-être certaines personnes sont capables de calculer des autocorrélations spatiales de cinquième ordre ou d'effectuer de l'analyse en composantes principales sans calepin. Je n'ai jamais rencontré de telles personnes. Mon point est que les cartes spatiales des taux de morbidité, ce que j'appelle les «données cartographiées», ont perdu leur utilité. Elles sont jolies, multicolores, mais nous apprennent rarement quoi que ce soit. Par contre, lorsque les analyses spatiales appropriées ont été effectuées et que les résultats sont présentés sur une carte bien conçue, il est probable que cela soit mieux que d'examiner, disons, de grandes colonnes de chiffres. Mais il ne s'agit pas là d'un travail simple, puisqu'il faut posséder à la fois des talents en épidémiologie, en statistique, en cartographie et en dessin. D'autres conférenciers auront traité ces sujets qui sont d'importance fondamentale.

J'estime que le temps est venu d'adopter un traitement plus perfectionné et plus informatif des issues de la maladie et des données covariables dans nos publications de type atlas. Ces publications sont consultées par un public beaucoup plus large que celui des revues spécialisées auxquelles elles ont été restreintes. Notre défi est

d'extraire de l'information de la mine de données dont nous disposons et de présenter cette information de façon compréhensible.

BIBLIOGRAPHIE

Gatrell, A.C. (1983). *Distance and Space: A Geographical Perspective*. Clarendon Press, Oxford.

Gesler, W. (1986). The uses of spatial analysis in medical geography: A review, *Soc Sci Med*, 23, 10, 963-973.

Glick, B. (1979). The spatial autocorrelation of cancer mortality, *Soc Sci Med*, 13, 123-179.

Grimson, R.C. (1981). Searching for hierarchical clustering of disease: spatial patterns of sudden infant death syndrome, *Soc Sci Med*, 15, 287-293.

Haggett, P. (1984). Hybridizing alternative models of an epidemic diffusion process, *Econ Geogr*, 52, 136-146.

Knox, G. (1963). Detection of low intensity epidemicity: application to cleft lip and palate, *Br J prev Soc Med*, 17, 121-127.

Selvin, S., Merrill, D., Schulman, J., Sacks, S., Bedell, L., et Wong, L. (1988). Transformations of maps to investigate clusters of disease, *Soc Sci Med*, 26, 2, 215-221.

Selvin, S., Shaw, G., Schulman, J., et Merrill, D.W. (1987). Spatial distribution of disease: three case studies, *JNCI*, 79, 3, 417-423.

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

LE BON ET LE MAUVAIS USAGE DES DONNÉES ET DES ENQUÊTES FÉDÉRALES - RECHERCHE EN GÉOGRAPHIE MÉDICALE AU CANADA

M.W. Rosenberg et A.M. James[1]

RÉSUMÉ

On peut diviser la géographie médicale en études portant sur la distribution spatiale et temporelle des maladies et études sur l'accès aux services de santé. Les deux types d'études amènent les géographes oeuvrant dans le domaine médical à employer les résultats d'enquêtes fédérales et des données de source fédérale dans l'élaboration de modèles statistiques et de statistiques spatiales. De plus, un nombre croissant de chercheurs tentent de lier des donnés environnementales à des questions touchant la santé. Ces efforts soulèvent des questions à propos de la pertinence de l'échelle et des techniques. Pour illustrer ces thèmes, nous examinons un choix de statistiques et de modèles spatiaux afin de présenter certains des problèmes rencontrés dans la recherche en géographie médicale au Canada.

MOTS CLÉS: Distribution spaciale; distribution temporelle; ondes épidémiologiques; modèles autorégressif spatiaux.

1. INTRODUCTION

Bien que la géographie médicale soit l'une des plus petites sous-disciplines de la géographie moderne, les géographes qui la pratiquent sont extrêmement actifs tant sur le plan national (p. ex. voir Rosenberg 1990) que sur le plan international (p. ex. voir Earickson 1988, 1990). Au Canada cependant, la géographie médicale reste dans une grande mesure *terra incognita*. Malgré toute notre activité, nous n'en connaissons encore que relativement peu sur les liens entre la santé et l'environnement, la diffusion des maladies dans l'espace, le comportement en matière de santé et l'accès aux services de santé ou encore les raisons de l'organisation spatiale des services de santé.

Ce paradoxe peut en partie s'expliquer de manière chiffrée; il n'y a que très peu de géographes oeuvrant dans ce domaine et le nombre de sujets qui peuvent être définis comme étant du ressort de la géographie médicale canadienne est considérable. Nous soutiendrons cependant qu'un deuxième facteur d'explication se dégage d'un examen des genres de questions médicales auxquelles les géographes aimeraient trouver des réponses, des statistiques spatiales que les géographes ont mises au point et des données nécessaires pour répondre à ces questions.

Le raisonnement est structuré comme suit. Dans la partie suivante, les intérêts des géographes oeuvrant dans le domaine médical sont brièvement exposés. Ensuite certains exemples de l'utilisation des statistiques en géographie médicale sont décrits. Dans la troisième partie, le rôle des enquêtes et des données du fédéral en géographie médicale au Canada est discuté. Enfin, une conclusion tournée vers l'avenir cherche à entrevoir quelle orientation il faut donner à l'utilisation des statistiques spatiales et des enquêtes et données du gouvernement fédéral dans l'étude de la géographie médicale au Canada.

---

[1] M.W. Rosenberg et A.M. James, Département de Géographie, Université Queen's, Kingston (Ontario), Canada K7L 3N6.

2. L'ÉTUDE DE LA GÉOGRAPHIE MÉDICALE

Historiquement, la géographie médicale s'est principalement intéressée à deux grands domaines: l'étude de la distribution spatiale et temporelle des maladies et l'étude de la prestation des soins de santé (voir Akhtar 1991; Meade et coll. 1988; Jones et Moon 1987). Ces thèmes ont été reliés du point de vue conceptuel par un intérêt plus général de la part des géographes pour les relations entre l'espèce humaine et le milieu dans lequel vous vivons. En géographie médicale, cela pourrait se traduire par l'étude des liens entre la santé et l'environnement.

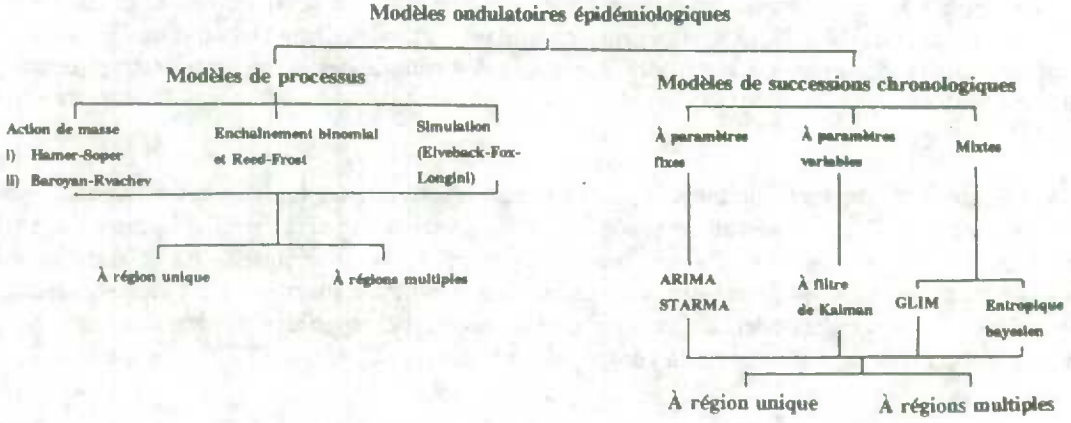
La première étape de presque toute recherche en géographie médicale consiste à cartographier les taux de morbidité et de mortalité de l'emplacement, des populations à risque, des médecins, des services de santé et des utilisateurs éventuels de services médicaux. C'est la visualisation des configurations spatiales des maladies, des taux de mortalité ou des points de prestation de soins de santé qui a stimulé la mise au point de statistiques spatiales par les géographes. Ces statistiques permettent aux géographes de mettre en relation les paramètres de la santé, de la vie et même de la mort avec les caractéristiques socio-économiques des populations touchées ainsi qu'avec le milieu qu'elles habitent de façon à éclairer les liens complexes entre les configurations spatiales et les processus naturels, culturels, économiques, politiques et sociaux.

3. STATISTIQUES SPATIALES

En géographie médicale, trois catégories de modélisation statistique spatiale ont été particulièrement importantes: les modèles ondulatoires épidémiologiques, les modèles autorégressifs spatiaux et les modèles écologiques spatiaux. Le temps et l'espace disponibles ne permettent pas une discussion détaillée de ces modèles et c'est pourquoi chacun ne sera que brièvement commenté.

Cliff et Haggett (1986) ont présenté une utile revue des modèles ondulatoires épidémiologiques. Les modèles ondulatoires épidémiologiques sont regroupés en modèles de processus et en modèles de successions chronologiques à la figure 1. Les modèles de processus «génèrent ou simulent les processus biologiques par lesquels des réceptifs contractent une maladie infectieuse» alors que les modèles de successions chronologiques incorporent le dossier historique d'une maladie afin d'identifier «la forme des forces génératrices qui la produisent» (Cliff et Haggett 1986, p. 85). Qu'il soit décidé de modéliser la diffusion d'une maladie sous forme de processus ou sous forme de succession chronologique, les préoccupations fondamentales restent la modélisation véritable des ondes, du seuil d'onde et de la forme d'onde.

Figure 1: Modèles ondulatoires épidémiologiques



Source: Cliff et Haggett (1986 : 87).

En utilisant le modèle Hamer-Soper comme exemple, Cliff et Haggett (1986, pp. 86 à 89) illustrent l'une des manières dont a progressé la modélisation des ondes épidémiologiques dans le temps. À l'équation un, le nombre d'infectants, I, au temps $t + 1$ est égal au nombre d'infectants au temps t plus un coefficient de diffusion, b, multiplié par le nombre de réceptifs, S, multiplié par le nombre d'infectants au temps t moins un coefficient de guérison, c, multiplié par le nombre d'infectants au temps t.

$$I_{t+1} = I_t + bSI_t - cI_t \tag{1}$$

Afin de donner davantage de réalisme au modèle, il faut tenir compte des naissances lors du calcul du nombre des réceptifs. À l'équation deux, le nombre des réceptifs au temps $t + 1$ est égal au nombre des réceptifs au temps t moins le coefficient de diffusion multiplié par le nombre de réceptifs multiplié par le nombre des infectants au temps t plus le nombre de naissances, a.

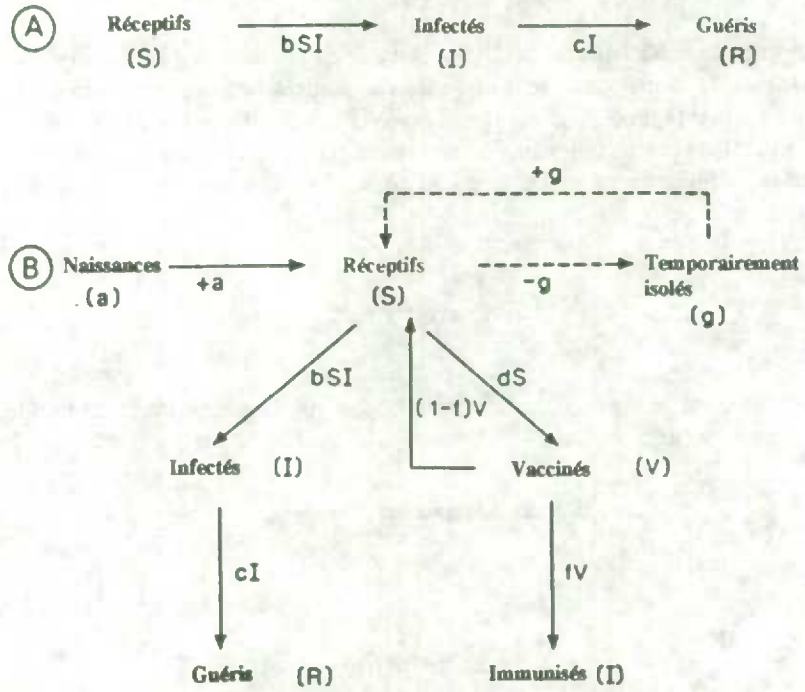
$$S_{t+1} = S_t - bSI_t + a \tag{2}$$

L'équation trois représente une troisième modification de l'équation un. Des expressions définissant le nombre de personnes vaccinées, V, le coefficient du taux de vaccination, d, le nombre de personnes devenues immunisées, I, et le taux d'immunisation réussie, f.

$$S_{t+1} = S_t - bSI_t + a - dS_t + (1 - f)V_t \tag{3}$$

Le développement des équations un à trois est résumé à la figure 2.

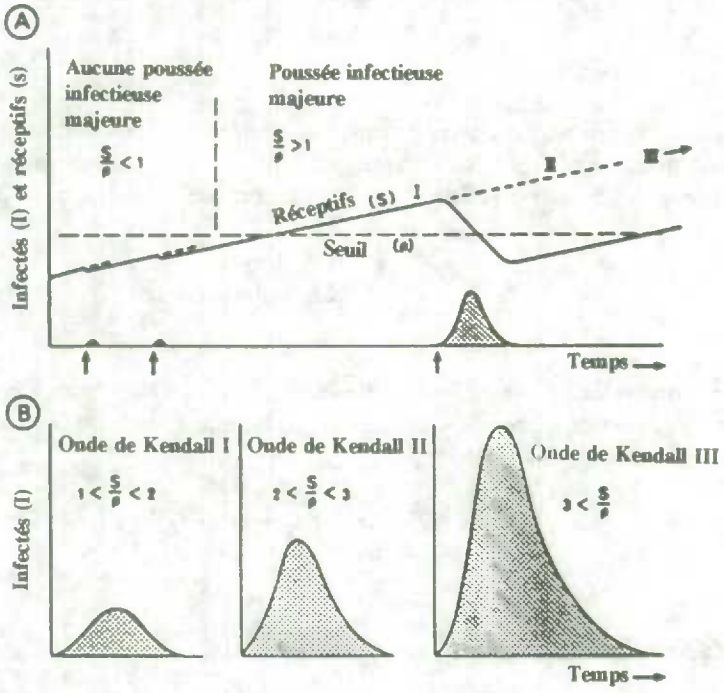
Figure 2: Modèles produisant des ondes



Source: Cliff et Haggett (1986 : 88).

Le seuil d'onde peut être mesuré au moyen d'un rapport simple, $p=c/b$, où p est le seuil, c le coefficient de guérison et b le coefficient de diffusion (Cliff et Haggett 1986, p.98). Les mesures de seuil peuvent être reliées aux modèles ondulatoires tel qu'illustré à la figure 3.

Figure 3: Relation entre la pente des ondes épidémiologiques et le seuil d'onde



Source: Cliff et Haggett (1986 : 99).

Élaborées à partir des mesures de modélisation et de seuils d'ondes, les mesures de formes d'ondes cherchent à caractériser les divers aspects de la propagation d'une maladie d'après des mesures comme le retard moyen, la durée de l'onde, la vitesse de l'onde, l'aplatissement de l'onde et des coefficients de diffusion (Cliff et Haggett 1986, pp. 103 à 106). À l'équation quatre une simple mesure du retard moyen dans les ondes d'une épidémie est définie; T est la durée de l'épidémie en mois, x_t est le nombre de cas signalés pendant le mois t et n le nombre total de cas signalés.

$$\bar{t} = \frac{1}{n} \sum_{t=1}^{T} t x_t \qquad (4)$$

Les équations cinq et six définissent la durée d'onde normalisée, s, et l'aplatissement de l'onde, b_2, est défini d'après ces équations aux équations sept et huit.

$$\text{durée d'onde normalisée, } s = \sqrt{m_2} \qquad (5)$$

où:

$$m_2 = \frac{1}{n} \sum_{t=1}^{T} (t - \bar{t})^2 x_t \qquad (6)$$

et

$$\text{aplatissement de l'onde, } b_2 = m_4 / m_2^2 \qquad (7)$$

où:

m_2 est défini comme ci-haut;

$$m_4 = \frac{1}{n} \sum_{t=1}^{T} (t - \bar{t})^4 x_t \tag{8}$$

Le coefficient de diffusion, b, est défini aux équations neuf et dix sous forme d'un modèle logistique où P_t est la proportion cumulée de cas jusqu'au temps t, e la base des logarithmes naturels et a et b des paramètres à estimer de sorte qu'à l'équation dix, b représente le taux de croissance moyen.

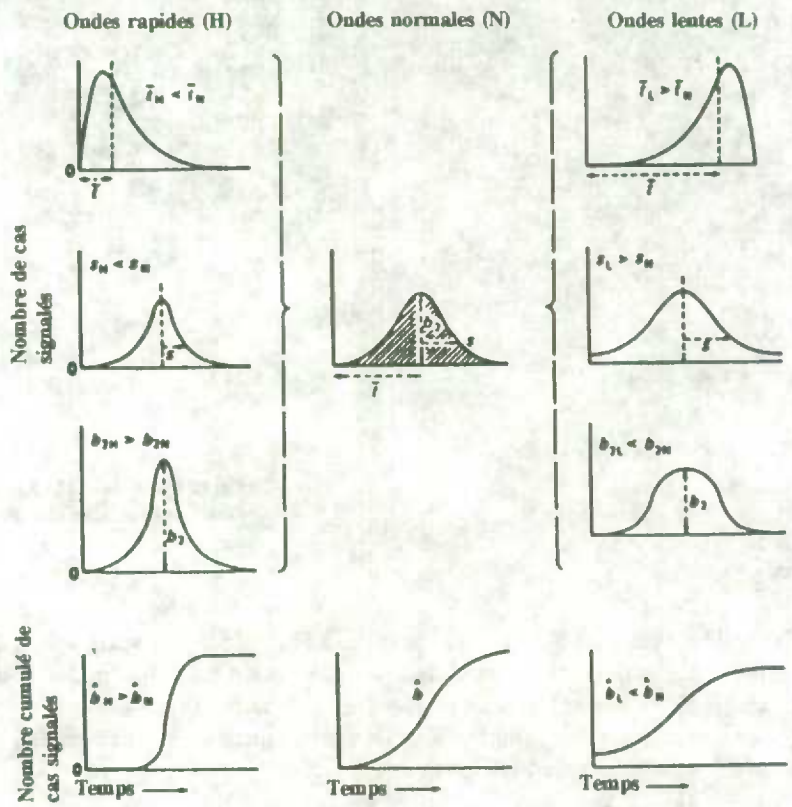
$$P_t = 1/(1 + e^{a-bt}) \tag{9}$$

où:

$$ln\left(\frac{1}{P_t} - 1\right) = a - bt \tag{10}$$

Les relations entre retard moyen, durée normalisée de l'onde et coefficients réel et prévu de diffusion sont illustrées à la figure 4. Il est à remarquer que des données à référence géographique uniforme dans le temps et dans l'espace constituent la clé d'une modélisation spatiale de ce genre.

Figure 4: Définitions géographiques de formes et de mesures d'ondes



Source: Cliff et Haggett (1986 : 104).

Des modèles autorégressifs spatiaux ont été utilisés pour identifier l'importance des effets régionaux et de voisinage dans la distribution géographique des données sur la mortalité lorsqu'une dépendance spatiale est

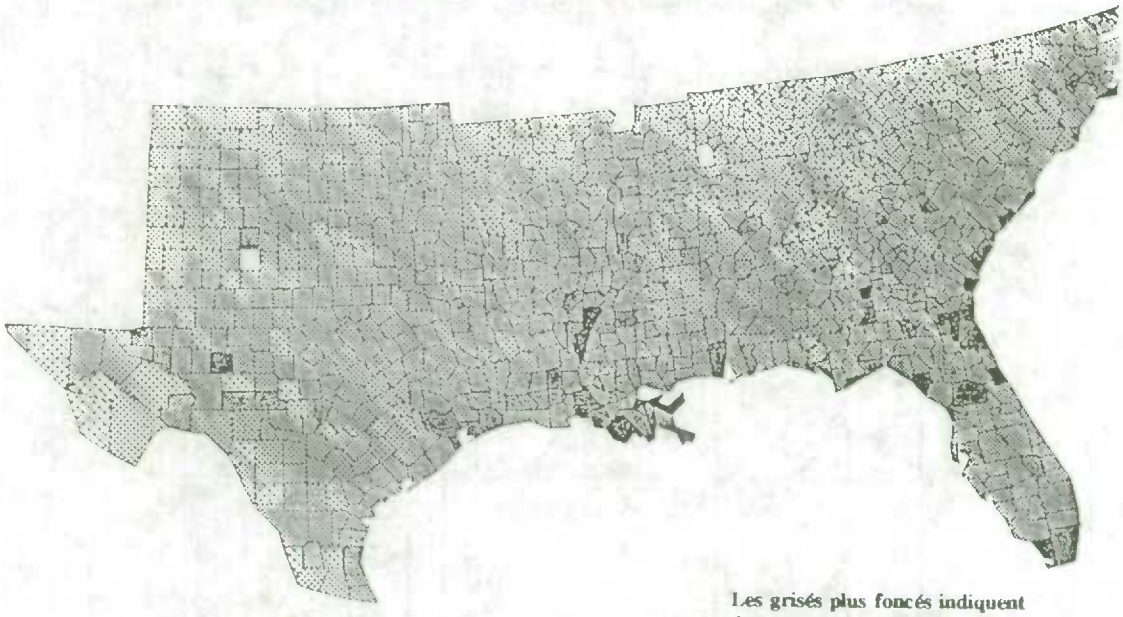
commune. Kennedy (1988) fournit un exemple de la manière dont il serait possible de définir un tel modèle. L'équation onze (modifiée d'après Kennedy (1988)) définit un modèle autorégressif spatial par lequel les effets régionaux et de voisinage sont simultanément estimés.

$$Y_i = \underbrace{B_1 h_i + B_2 p_i}_{\text{composante de la tendance régionale}} \underbrace{+ B_3 Z_{3i} + \ldots + V_7 Z_{7i} + e_i}_{\text{composante de l'effet de voisinage}} \qquad (11)$$

Y_i est le taux de mortalité normalisé selon l'âge dans le comté $_i$. h_i est $Cos(p_i)*h_i$ où h_i est la coordonnée en longitude du centre de gravité du comté $_i$ et p_i la latitude du centre de gravité du comté $_i$. Z_{ki} est égal à $W_{ij}W_j$ où W est la matrice de pondération pour le j^e comté voisin du comté $_i$ et V la matrice des taux de mortalité des comtés voisins $_j$ du comté $_i$ et k l'ordre du comté voisin. e_i est le terme d'erreur.

La distribution réelle du taux de mortalité par cancer du poumon chez la population mâle par comté dans les états du golfe du Mexique et du sud-est des États-Unis est présentée à la figure 5 tandis que la figure 6 présente la distribution prévue du taux de mortalité par cancer du poumon chez la population mâle par comté obtenue au moyen du modèle d'autocorrélation spatiale (Kennedy 1988).

Figure 5: Cancer du poumon chez les hommes par comté dans les états du golfe du Mexique et du sud-est de la côte de l'Atlantique



Les grisés plus foncés indiquent des taux de mortalité plus élevés

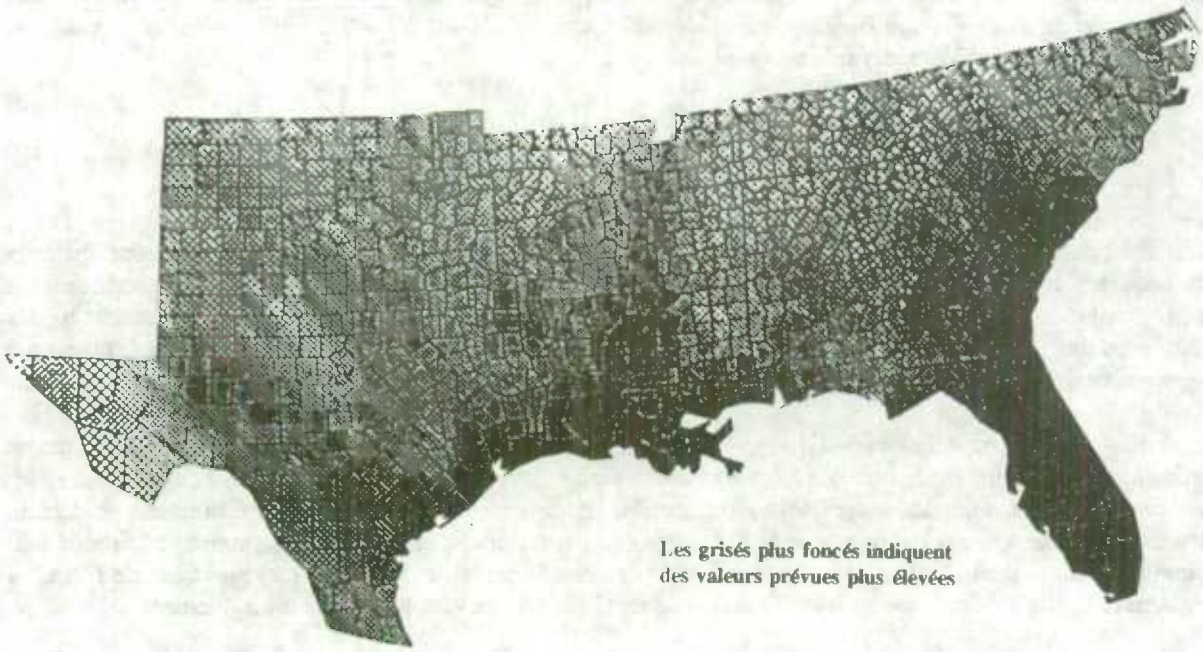
Source: Kennedy (1988 : 121).

Paraphrasant Susan Kennedy (1988, p. 120), on peut dire que l'échelle de la variation spatiale joue un rôle important dans l'approche géographique de la modélisation des maladies. Une maladie peut présenter une importante variation à une échelle d'ensemble mais ne montrer qu'une petite variation à une autre échelle. Par conséquent, il est important d'analyser les données à l'échelle à laquelle elles présentent une variation dans l'espace. Nous reviendrons à cette question plus loin.

Les modèles écologiques spatiaux sont de loin les plus couramment présentés en géographie médicale. Ils prennent de manière caractéristique la forme d'une équation de régression multiple dans laquelle la variable dépendante est un nombre ou un taux de mortalité par unité spatiale ou une mesure de l'accès aux soins de santé et les variables indépendantes sont des mesures de conditions socio-économiques ou physiques dans les unités spatiales correspondantes. On doit soupçonner que leur popularité est, à tout le moins en partie, fonction d'un

manque de données concernant les personnes qu'il est possible de se procurer à une échelle géographique moindre que l'échelle provinciale.

Figure 6 : Mortalité prévue par cancer du poumon chez les hommes



Source: Kennedy (1988 : 125).

4. BON ET MAUVAIS USAGE DES ENQUÊTES ET DES DONNÉES DU FÉDÉRAL

Où les spécialistes de la géographie médicale au Canada recherchent-ils leurs données sur la santé lorsqu'ils ne les recueillent pas eux-mêmes? Ils les trouvent principalement dans les *Statistiques démographiques* et les *Rapports médicaux* et dans une moindre mesure dans l'*Enquête sociale générale* et l'*Enquête sur la santé et les limitations des activités*.

Pour presque tous les exemples publiés de recherches en géographie médicale des variables sont tirées des *Statistiques démographiques* ou des *Rapports médicaux* et sont utilisées comme variables dépendantes; des données tirées du recensement ou de bases de données environnementales et adaptées à des modèles écologiques spatiaux sont utilisées comme variables indépendantes. Par exemple, Foster et Norrie (1988) ont utilisé des données sur le cancer tirées de la *Répartition géographique de la mortalité au Canada* et mettent en corrélation des taux de mortalité pour des cancers spécifiques au niveau de la division de recensement avec l'information sur la qualité de l'eau tirée d'une base de données créée par l'Université d'Ottawa et Santé et Bien-être social Canada. Godon et coll. (1989) ont utilisé le *Registre du cancer du Québec* et des données sur l'agriculture ainsi que des données socio-économiques tirées du *recensement* pour illustrer les liens écologiques entre certains cancers et l'utilisation de pesticides dans les régions rurales au Québec. Cependant une de leurs conclusions est que la démonstration d'une causalité exigerait une recherche au niveau individuel.

Qu'en est-il des données au niveau individuel? Nous ne connaissons aucun exemple publié d'utilisation de l'*Enquête sociale générale* par des spécialistes de la géographie médicale. Moore et coll. (1990) et Moore et Rosenberg (1991) ont abondamment utilisé l'*Enquête sur la santé et les limitations des activités*, mais cette recherche sur la démographie fondamentale des personnes atteintes d'invalidité en Ontario a été principalement descriptive.

Il ne faut pas en conclure que les spécialistes de la géographie médicale n'analysent pas de données au niveau individuel. Il existe un grand nombre d'exemples de recherches publiées pour lesquelles des données ont été

recueillies par des géographes dans le cadre de leurs propres enquêtes (p. ex. Shannon et coll. 1988) ou pour lesquelles des données individuelles ont été combinées à des données socio-économiques tirées du *recensement* (p. ex. Liaw et coll. 1989). Les méthodologies vont de la corrélation simple aux modèles de logit.

D'après un relevé de la documentation récemment publiée, il est également frappant de constater qu'il n'existe aucun exemple de modélisation ondulatoire épidémiologique ou de modélisation autorégressive spatiale de maladies au Canada par des géographes canadiens.

5. L'UTILISATION FUTURE DES STATISTIQUES SPATIALES ET DES ENQUÊTES DU FÉDÉRAL DANS L'ÉTUDE DE LA GÉOGRAPHIE MÉDICALE AU CANADA

Quelles conclusions pouvons-nous tirer des parties qui précèdent quant à l'utilisation des statistiques spatiales et des enquêtes du gouvernement fédéral dans l'étude de la géographie médicale au Canada? Les spécialistes de la géographie médicale forment au Canada un petit groupe dont les recherches sont principalement basées sur les modèles écologiques spatiaux et dont les données sont principalement tirées des *Statistiques démographiques*, du *recensement*, de certaines bases de données environnementales et de leurs propres enquêtes.

Les spécialistes en géographie médicale sont-ils incapables d'utiliser certaines des méthodes plus évoluées de modélisation ondulatoire épidémiologique et les modèles autorégressifs spatiaux? En toute honnêteté, certains le sont probablement, mais les autres sont en mesure d'appliquer ces méthodes. Ce qui nous est également apparu évident, mais qui ressort moins de la documentation examinée, c'est que les géographes spécialisés dans le domaine médical souhaitent trouver des explications reliant la santé des personnes à leurs lieux de travail et de résidence, à leur mode de vie et à leur comportement. Ils se trouvent ainsi face à un dilemme.

Quiconque a utilisé l'*Enquête sociale générale* et l'*Enquête sur la santé et les limitations des activités* est en mesure de critiquer la qualité des données ou l'envergure de l'information présentée. Il est ironique de constater qu'elles ne renferment presqu'aucune information géographique qui permettrait aux géographes ou en fait aux autres scientifiques des domaines médical ou social de déterminer *si la géographie importe* pour reprendre une expression rendue célèbre par Doreen Massey (1984).

Pour en revenir au commentaire de Susan Kennedy concernant l'importance de la variation spatiale de l'échelle, il n'y aura vraisemblablement pas d'autres exemples d'application des modèles autorégressifs spatiaux ou des modèles ondulatoires épidémiologiques à l'étude de la géographie médicale au Canada à moins que les données soient diffusées au niveau de la subdivision de recensement ou même à des échelles géographiques plus petites. La conséquence directe est que l'évolution de l'utilisation des statistiques spatiales pour comprendre la géographie médicale canadienne sera ralentie au Canada, parce que les données nécessaires pour la mise à l'épreuve des modèles sont en pratique inaccessibles aux géographes ou n'existent pas du tout.

Il serait facile de conclure à ce stade que si Statistique Canada fournissait simplement davantage de données à des échelles géographiques plus petites tous les autres problèmes se régleraient d'eux-mêmes. Nous n'en croyons rien. Nous aimerions plutôt formuler les suggestions suivantes. Une collaboration plus étroite est nécessaire entre les géographes et les statisticiens et démographes de Statistique Canada qui s'intéressent plus particulièrement à la santé et aux soins de la santé. Cette collaboration pourrait prendre la forme de travaux conjoints de modélisation des genres décrits menés à Statistique Canada avec des données à de plus petites échelles. De tels travaux pourraient permettre de solutionner certains des problèmes d'accessibilité reliés à des questions de confidentialité ou de coûts. Deuxièmement, dans le cas des enquêtes à façon, les spécialistes de la géographie médicale devraient être consultés pour la formulation de questions concernant la géographie démographique de manière à permettre d'effectuer des analyses vérifiant les liens entre la santé, le mode de vie, le lieu de travail, le lieu de résidence et le milieu physique. Troisièmement, un des sujets que nous n'avons pas abordé est celui de la forme d'une éventuelle collaboration entre les géographes et les démographes et statisticiens de Statistique Canada en vue de la mise au point et de la mise à l'épreuve de méthodes permettant l'estimation de taux à des échelles géographiques plus grandes qui pourraient être utilisés pour effectuer des estimations plus fiables à des échelles géographiques plus petites.

Ainsi, en quoi consiste la mauvaise utilisation des enquêtes et des données du fédéral dans l'étude de la géographie médicale au Canada? Ces données sont mal utilisées parce que trop peu de recherches sur les liens entre la santé et l'environnement sont effectuées à une époque où le grand public est de plus en plus préoccupé par cette question. Notre hypothèse voulant que les statistiques spatiales puissent constituer une contribution dans ce domaine ne sera cependant vérifiée que lorsque les spécialistes canadiens de la géographie médicale commenceront à utiliser les méthodes discutées dans le présent article ainsi que dans d'autres articles du présent numéro spécial et qu'accessoirement les enquêtes et données du fédéral fourniront les types d'information nécessaires pour qu'il vaille la peine d'utiliser les statistiques spatiales dans l'étude de la géographie médicale au Canada.

BIBLIOGRAPHIE

Akhtar, R. (1991). *Environment and Health*. New Delhi: Ashish Publishing House.

Cliff, A.D., et Haggett, P. (1986). Disease Diffusion, *Medical Geography: Progress and Prospect*, London: Croom Helm, 84-125.

Earickson, R., ed. (1988). Medical Geography - Selected Papers from the 1986 Rutgers Symposium, *Social Science and Medicine*, 26, 1.

Earickson, R., ed. (1990) Medical geography - Selected Papers from the 1988 Kingston Symposium, *Social Science and Medicine*, 30, 1.

Foster, H. et Norrie, I. (1988). Water quality and cancer of the digestive tract: the Canadian experience, *Proceedings of the Third International Symposium of Medical Geography*, rédigé par M. Anderson, M. Rosenberg et R. Tinline, Kingston, Canada: Université Queen, Département de géographie, 91-101.

Godon, D., Thouez, J.-P., et Lajoie, P. (1989). Analyse géographique de l'incidence des cancers au Québec en fonction de l'utilisation des pesticides en agriculture, 1982-1983, *The Canadian Geographer*, 33, 3, 204-217.

Jones, K., et Moon, G. (1987). *Health, Disease and Society: An Introduction to Medical Geography*, London: Routledge and Kegan Paul.

Kennedy, S. (1988). A geographic regression model for medical statistics, *Social Science and Medicine*, 26, 1, 119-129.

Liaw, K-L., Wort, S.A., et Hayes, M.V. (1989). Intraurban mortality variation and income disparity: a case study of Hamilton-Wentworth, *The Canadian Geographer*, 33, 2, 131-145.

Massey, D. (1984). Introduction: geography matters, *Geography Matters!* rédigé par D. Massey et J. Allen, Cambridge: Cambridge University Press, 1-11.

Meade, M., Florin, J.W., et Gesler, W.M. (1988) *Medical Geography*, New York: Guilford Press.

Moore, E.G., et Rosenberg, M.W. (1991). Disability in the adult population living in Ontario Institutions. Rapport préparé par the Ontario Ministry of Community and Social Services. Kingston, Canada: Université Queen, Département de géographie.

Moore, E.G., Burke, S.O., et Rosenberg, M.W. (1990). The disabled adult residential population of Ontario. Rapport préparé par the Ontario Ministry of Community and Social Services, Kingston, Canada: Université Queen, Département de géographie.

Rosenberg, M.W., ed. (1990). Focus: achieving health for all - the geographer's role, *The Canadian Geographer*, 34, 4, 331-346.

Shannon, H.S., Hertzman, C., Julian, J.A., Hayes, M.V., Henry, N., Charters, J., Cunningham, I., Gibson, E.S., et Sackett, D.L. (1988). Lung cancer and air pollution in an industrial city - a geographical analysis, *Canadian Journal of Public Health*, 79, 255-259.

Statistique Canada (1985). *General Social Survey - Health and Social Support*. Ottawa: Supply and Services Canada.

Statistique Canada (1986). *Health and Activity Limitation Survey*. Ottawa: Supply and Services Canada.

Statistique Canada. *Vital Statistics: Causes of Death*. Ottawa: Supply and Services Canada. Comes in various years and volumes.

Statistique Canada. *Health Reports*. Ottawa: Canadian Centre for Health Information. Comes in various years and volumes.

SESSION 8

Analyse spatiale des données d'enquêtes

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

INÉGALITÉS EN MATIÈRE DE SANTÉ SELON LES CARACTÉRISTIQUES DES QUARTIERS

Russell Wilkins[1]

RÉSUMÉ

L'utilisation du fichier de conversion des codes postaux de Statistique Canada et de logiciels pour trouver les codes postaux permet de coder des fichiers de données sur la santé pour lesquels des données relatives aux adresses sont disponibles selon le secteur de recensement, le secteur de dénombrement ou le côté d'îlot avec un nombre raisonnable d'interventions manuelles. Quand on établit un rapport entre les données sur la santé ainsi codées et les données du recensement ou d'autres données sur les caractéristiques du quartier pour les mêmes régions, on peut effectuer divers types d'analyses. De cette façon, on peut s'attaquer, au moins indirectement, à des questions d'équité, d'environnement et d'efficacité administrative, sans avoir à recueillir d'autres données. Des exemples d'un tel travail comprennent des études des tendances relatives à la mortalité infantile et à l'insuffisance pondérale selon le revenu du quartier dans les régions urbaines du Canada. Pour la population, excluant les pensionnaires d'établissements institutionnels, vivant dans des régions urbaines, on peut aussi effectuer des analyses plus simples en n'utilisant que les trois premiers caractères du code postal, même quand on ne dispose pas de l'adresse complète.

MOTS-CLÉS: Codes postaux; données régionales; statut socio-économique.

1. INTRODUCTION

On a besoin de données sur la santé selon la région à des fins de planification, de surveillance, d'évaluation et de recherche. Les données sur les naissances, les décès et l'hospitalisation sont codées couramment au niveau municipal, mais dans les grandes villes, où la majorité de la population habite, on a besoin d'un codage géographique plus précis pour répartir les indicateurs de santé entre les districts de services. On peut aussi utiliser les données sur la santé codées en fonction du secteur de recensement ou d'unités plus petites pour étudier les inégalités socio-économiques et les risques environnementaux, basés sur les caractéristiques du quartier.

Dans le passé, le codage géographique régional devait être effectué à la main à l'aide de répertoires de rues et en consultant des cartes. Récemment, toutefois, l'utilisation du fichier de conversion des codes postaux (FCCP) de Statistique Canada et de logiciels pour trouver les codes postaux a permis d'automatiser la plus grande partie de ce travail. De plus en plus, il sera possible de produire et d'utiliser des données régionales de façon courante.

Après avoir expliqué brièvement comment ce codage peut être effectué et à partir de quels fichiers, je présenterai quelques exemples du genre de conclusions qu'on peut tirer de telles études, en me basant sur le travail auquel j'ai participé au cours des dernières années.

---

[1] Russell Wilkins, Centre canadien d'information sur la santé, Statistique Canada, 18-N, édifice R.H. Coats, Ottawa (Ontario) Canada K1A 0T6.

2. MÉTHODES

2.1 Sources de données

Les données sur les indicateurs de santé qui peuvent être analysées par régions locales comprennent les statistiques de l'état civil sur les naissances vivantes, les décès et les mortinaissances, les statistiques sur la morbidité hospitalière tirées des dossiers d'admission - radiation, les données tirées des registres des maladies (comme pour les cas de cancer ou d'insuffisance rénale), les données du recensement sur l'invalidité provenant de l'échantillon de 20%, les données sur les personnes qui reçoivent des soins prolongés et des collections de données spéciales ou recueillies à des fins particulières sur divers sujets.

Il pourra arriver que les dénominateurs utilisés pour le calcul des taux relatifs à ces indicateurs de santé soient disponibles dans le même fichier (comme dans le cas des taux d'insuffisance pondérale calculés à partir du fichier des naissances vivantes), bien que les chiffres de population du recensement selon la région locale constituent la source habituelle de données pour les dénominateurs. Les données du recensement sur la population par région locale peuvent être obtenues pour l'une quelconque des unités géographiques normalisées du recensement ou pour des agrégats de ces unités ou, dans le cas de la population, excluant les pensionnaires d'établissements institutionnels, vivant dans des régions urbaines, pour des "régions de tri d'acheminement" non normalisées (correspondant aux trois premiers caractères du code postal). D'autres sources de données pour les dénominateurs comprennent les renseignements du genre registre de population produits à partir de listes compilées spécialement de bénéficiaires de soins de santé au niveau provincial (comme au Québec et en Saskatchewan) ou de données sur les déclarants compilées par Statistique Canada pour les régions de tri d'acheminement urbaines et les régions de services correspondant à des codes postaux ruraux.

2.2 Genres d'analyses basées sur les données codées

Les données sur la santé codées en fonction des régions locales peuvent être utilisées pour décrire l'état de santé général selon la région géographique (en fonction des plus petites unités pour lesquelles des données statistiquement fiables peuvent être produites); elles peuvent être groupées pour des secteurs de services administratifs locaux ou pour des secteurs de services de programmes (comme les secteurs, unités, districts et régions de services de santé et de services sociaux) ou groupées selon les caractéristiques du quartier comme le pourcentage de la population qui vit sous le seuil de pauvreté. Dans ce dernier cas, la région de résidence est utilisée comme un indicateur du statut socio-économique. À un niveau un peu moins précis d'approximation, on peut effectuer des analyses géographiques et des analyses socio-économiques en n'utilisant que les trois premiers caractères du code postal (RTA) plutôt que les régions plus précises que sont les secteurs de recensement ou les secteurs de dénombrement basés sur le code postal complet à six caractères.

2.3 Utilisation du fichier de conversion des codes postaux (FCCP)

Le FCCP (Statistique Canada 1991) est utilisé pour établir un rapport entre les codes postaux et la géographie de la région locale. Dans les régions urbaines du Canada, le même code postal sert généralement pour trente personnes et ces dernières résident habituellement sur un même côté d'îlot et dans un même secteur de dénombrement. Dans de tels cas, le FCCP montrera que le code postal est lié à un centroïde de carte unique qui, à son tour, est relié à toute la hiérarchie des unités géographiques du recensement: du côté d'îlot et (ou) du secteur de dénombrement au secteur de recensement, à la région métropolitaine de recensement ou à l'agglomération de recensement; ou du côté d'îlot et (ou) du secteur de dénombrement à la subdivision de recensement, à la division de recensement, à la province et à la région.

À l'extérieur des régions urbaines, toutefois, chaque code postal rural s'applique à toute la région desservie par un bureau de poste rural, région dans laquelle peuvent se trouver plusieurs milliers de personnes. Dans le FCCP, pour presque tous les codes postaux ruraux, il existe un rapport entre le code postal et plus d'un secteur de dénombrement et souvent plus d'une subdivision de recensement (ou d'une municipalité). Au Canada, il est facile d'identifier les codes postaux ruraux parce que le deuxième caractère du code postal est toujours "0".

Une description détaillée des méthodes utilisées pour coder les adresses postales en fonction des régions a été présentée dans un atelier de Statistique Canada qui a suivi le symposium et cette description sera publiée dans

un numéro à venir de *Rapports sur la santé* (publication n° 82-003 au catalogue de Statistique Canada). On peut obtenir, sur demande, des sous-programmes SAS pour apparier des codes postaux à des emplacements géographiques précis à l'aide du FCCP, en s'adressant à l'auteur au Centre Canadian d'Information sur la Santé.

2.4 Analyses basées sur les trois premiers caractères du code postal

Quand les données sur l'adresse ne sont pas disponibles pour vérifier les codes postaux douteux ou quand on désire une méthode d'analyse plus simple et plus rapide, il est souvent possible d'effectuer une analyse un peu moins précise mais néanmoins significative en n'utilisant que les trois premiers caractères du code postal, que Postes Canada désigne par l'expression "région de tri d'acheminement" (RTA). Dans les régions urbaines, une RTA typique peut renfermer de 20 000 à 40 000 personnes, elle contient donc de 5 à 10 fois plus de personnes qu'un secteur de recensement typique qui en comprend 4 000. Il faut toutefois remarquer que la taille des RTA est basée sur le volume de courrier reçu, de sorte qu'une RTA dans une région du centre-ville qui comprend de nombreuses entreprises aura une plus petite population qu'une RTA dans une zone de banlieue où il y a peu d'entreprises. Bien qu'elles soient socialement moins homogènes que les secteurs de recensement, les RTA constituent une façon commode de diviser de grandes régions urbaines en unités infra-municipales relativement petites qui peuvent être utilisées pour cartographier les taux de fréquence d'apparition des indicateurs de santé, les groupant en des approximations des districts de services de santé et de services sociaux ou les regroupant selon des caractéristiques socio-économiques.

Statistique Canada peut fournir des données sur les profils du recensement pour la population, excluant les pensionnaires d'établissements institutionnels, vivant dans les RTA urbaines (par l'intermédiaire de CAMSIM ou des bureaux régionaux). Les RTA rurales sont exclues car elles chevauchent les RTA urbaines et les personnes placées en établissement institutionnel ne sont pas comprises parce que les données relatives au code postal sur les questionnaires du recensement n'ont été saisies que pour les ménages faisant partie de l'échantillon de 20%, ce qui exclut les personnes placées en établissement institutionnel. Puisque les taux de placement en établissement institutionnel sont élevés chez les personnes âgées de 75 ans et plus (le groupe d'âge le plus élevé qui figure dans les profils pour les RTA), je recommande de limiter les analyses de données basées sur les RTA aux personnes âgées de moins de 75 ans, pour lesquelles les taux de placement en établissement institutionnel sont très faibles.

3. RÉSULTATS

3.1 Codage des données en fonction du secteur de recensement

Basé sur des données préparés pour une étude d'issues de grossesse défavorables et de mortalité infantile selon le revenu du quartier (Wilkins, Sherman et Best 1991), le tableau 1 fournit un exemple de données sur les régions locales codées en fonction du secteur de recensement. Pour chaque secteur de recensement, le nombre total de naissances vivantes est indiqué ainsi que le pourcentage de naissances selon l'âge, l'état matrimonial et le lieu de naissance de la mère ainsi que le pourcentage des naissances où il y avait insuffisance pondérale, naissance prématurée ou retard de croissance. Des taux plus stables auraient pu être obtenus si l'on avait groupé les données pour plusieurs années. Les données sur la santé codées en fonction du secteur de recensement peuvent être utilisées pour déterminer des populations cibles et (ou) groupées pour correspondre aux régions desservies par les districts locaux et régionaux de santé et de services sociaux.

3.2 Groupement de secteurs de recensement selon des caractéristiques socio-économiques

Une autre façon d'utiliser les données régionales consiste tout d'abord à grouper les secteurs de recensement selon des caractéristiques socio-économiques puis à calculer les taux d'apparition des indicateurs de santé. La carte 1 montre comment, dans la région métropolitaine de recensement (RMR) d'Ottawa - Hull, on a groupé les secteurs de recensement en quintiles basés sur le pourcentage de la population de moins de dix-huit ans vivant dans des familles dont le revenu était inférieur au seuil de faible revenu de Statistique Canada (d'après des totalisations spéciales du recensement de 1986 produites pour le Centre canadien d'information sur la santé). Dans ce cas, les points de découpage ont été choisis de façon à attribuer un cinquième des naissances totales à chaque quintile. Cela ne donne pas le même résultat que si l'on attribuait un cinquième des secteurs de

recensement à chaque groupe puisque, pour des raisons historiques, les secteurs de recensement de la région centrale ont tendance à être à la fois plus petits et plus pauvres, alors que ceux de la banlieue ont tendance à être plus grands et moins pauvres.

TABLEAU 1: NAISSANCES SELON LE SECTEUR DE RECENSEMENT (SR): CARACTÉRISTIQUES DES MÈRES ET DES NOUVEAU-NÉS PAR RÉGION MÉTROPOLITAINE DE RECENSEMENT (RMR), CANADA URBAIN, 1986 (POURCENTAGES PAR LIGNE)

CMA	CT	BIRTHS	----------AGE----------			UNMAR	FOREIG	LBW		SGA
RMR	SR	NAISS	<20	20-34	35+	CÉLIB	ÉTRANG	PPN	PRE	RDC
OTTAWA-HULL										
505	2.02	43	2.3	74.4	23.3	7.0	16.3	4.7	2.3	7.0
505	2.03	154	1.9	87.7	10.4	11.7	22.1	5.8	7.1	4.5
505	3.00	58	3.4	82.8	13.8	10.3	34.5	3.4	6.9	8.6
505	4.00	43	4.7	93.0	2.3	25.6	16.3	2.3	7.0	9.3
505	5.00	70	5.7	71.4	22.9	12.9	24.3	4.3	7.1	5.7
505	6.00	26	--	76.9	23.1	3.8*	46.2	3.8*	--	15.4
505	7.01	17	17.6	70.6	11.8	17.6	11.8	11.8	11.8	--
505	7.02	68	2.9	92.6	4.4	16.2	25.0	5.9	13.2	8.8
505	7.03	23	4.3*	73.9	21.7	21.7	39.1	--	--	4.3*
505	8.00	60	6.7	81.7	11.7	11.7	13.3	8.3	8.3	13.3
505	9.00	35	--	94.3	5.7	8.6	11.4	5.7	5.7	5.7
505	10.00	51	11.8	82.4	5.9	23.5	29.4	3.9	2.0	5.9
...										
505	28.00	100	7.0	83.0	10.0	30.0	17.0	11.0	12.0	17.0
505	29.00	72	15.3	77.8	6.9	27.8	19.4	16.7	16.7	9.7
505	30.00	67	7.5	80.6	11.9	20.9	17.9	6.0	4.5	7.5
505	31.00	45	2.2	86.7	11.1	20.0	28.9	8.9	8.9	20.0
505	32.00	1	.	.	.	.	.	.	.	.
505	32.01	17	11.8	76.5	11.8	23.5	23.5	--	--	5.9*
505	32.02	56	3.6	83.9	12.5	16.1	26.8	14.3	14.3	3.6

LBW = LOW BIRTH WEIGHT (<2500 G)
PPN = PETIT POIDS A LA NAISSANCE = INSUFFISANCE PONDÉRALE (<2500 G)

PRE = PRÉMATURE (<37 WEEKS/SEMAINES)

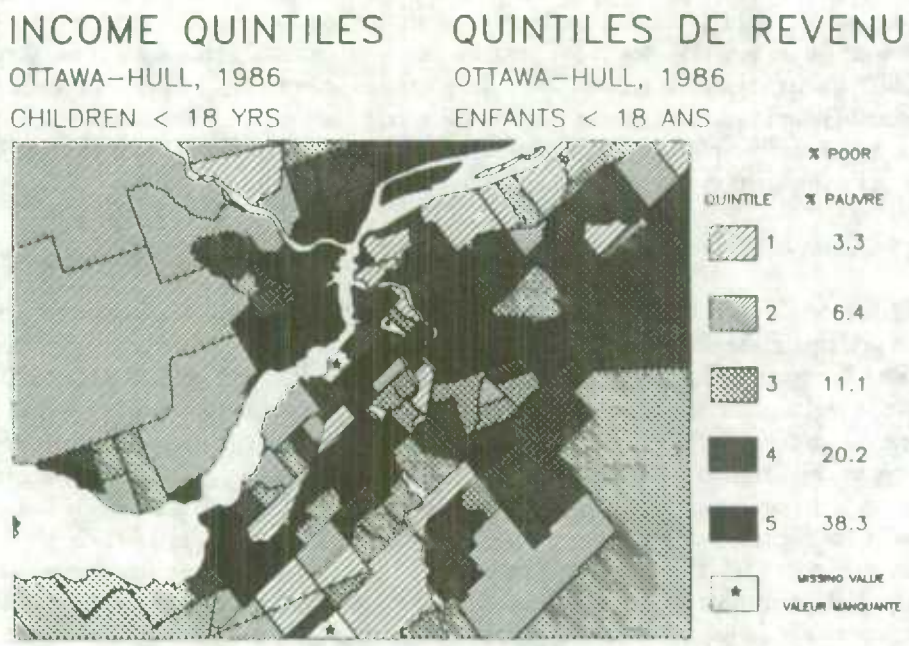
SGA = SMALL FOR GESTATIONAL AGE[1]
RDC = RETARD DE CROISSANCE[1]

[1] ARBUCKLE TE AND SHERMAN GJ, AN ANALYSIS OF BIRTH WEIGHT BY GESTATIONAL IN CANADA, CMAJ 1989 140:157-165, TABLES V-V1.

NOTE: * INDIQUE UN COEFFICIENT DE VARIATION DE 16.7% À 33.3%: UTILISER AVEC PRUDENCE.
. INDIQUE UN COEFFICIENT DE VARIATION QUI DÉPASSE 33.3%: DONNÉES SUPPRIMÉES.
-- INDIQUE UN POURCENTAGE NUL.

SOURCE: CENTRE CANADIEN D'INFORMATION SUR LA SANTÉ, STATISTIQUE CANADA. FILE = SUM1CTS (COUNTS/TOTAUX), SUM2CTS (%) (FICHIERS); PROGRAM = BBILON2 1990-03-30 (PROGRAMME)

Carte 1

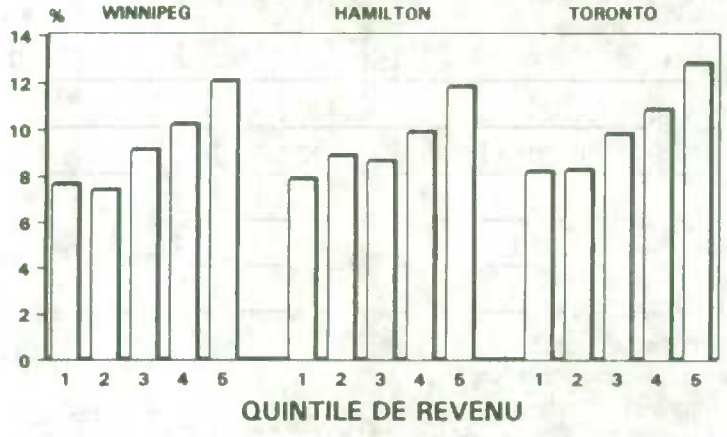


3.3 Taux de naissances avec retard de croissance dans trois régions métropolitaines

La figure 1 montre les résultats d'analyses pour trois RMR après regroupement des secteurs de recensement comme on l'a décrit plus haut (Wilkins, Sherman et Best 1991). À Winnipeg, Hamilton et Toronto, le pourcentage de naissances avec retard de croissance était d'environ 50% plus élevé dans les régions les plus pauvres (5e quintile) comparativement aux régions les moins pauvres (1er quintile). Dans l'étude de laquelle les présentes données sont tirées, d'autres analyses utilisaient la gamme complète de données sur les particuliers disponibles dans les actes de naissance, y compris le poids de naissance et l'âge de la grossesse de l'enfant, ainsi que l'âge, la parité (nombre d'accouchements antérieurs), l'état matrimonial et le lieu de naissance de la mère. La seule variable "écologique" employée était le pourcentage d'enfants vivant sous le seuil de pauvreté dans le quartier, renseignement utilisé pour combiner des secteurs de recensement semblables en quintiles.

Figure 1

NAISSANCES AVEC RETARD DE CROISSANCE
CANADA URBAIN, 1986



Note: RCI < 10ème rang centile
SOURCE: STATISTIQUE CANADA / SANTÉ ET BIEN-ÊTRE SOCIAL CANADA (WILKINS, SHERMAN & BEST, 1990)

Il faut remarquer que ce genre d'analyse -- où le quartier de résidence est utilisé comme substitut pour le statut socio-économique -- n'est pas approprié à l'extérieur de régions urbaines assez grandes. Cela est dû au fait que, dans les petites villes et les régions rurales, les plus petites régions résidentielles qui peuvent être définies d'une manière digne de foi en fonction des codes postaux ou des adresses postales sont la municipalité ou le groupe de municipalités qui partagent un même bureau de poste; dans un tel cas, toutes les classes socio-économiques seront groupées. Quand toutes les classes partagent le même code géographique, les études basées sur l'écologie ont tendance à embrouiller plutôt qu'à faire ressortir toutes les différences réelles en fonction du statut socio-économique qui pourraient exister au niveau de chaque classe.

3.4 Mortalité due au cancer du poumon selon la région et le district

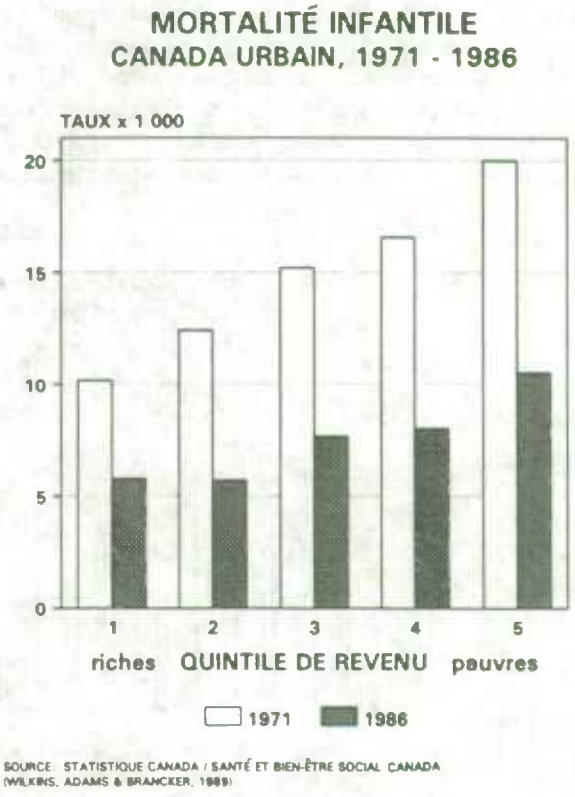
Des rapports de mortalité due au cancer du poumon, obtenus par standardisation indirecte, ont été calculés pour chacune des principales régions sanitaires du Québec (Hoey, Wilkins, Gagnon et O'Loughlin 1987). La différence entre la région avec le taux le plus élevé et celle avec le taux le plus faible était d'environ 50% et le classement de la région sanitaire du Montréal métropolitain (qui comprend toute la division de recensement de l'île de Montréal) n'était que légèrement supérieur à la moyenne. Toutefois, quand les taux de mortalité standardisés pour chacun des sept départements de santé communautaire (DSC) de la région de Montréal ont été calculés, on a trouvé que la différence entre les DSC de Montréal était de beaucoup supérieure à la différence entre les régions dans toute la province. Il faut remarquer que les DSC de Montréal comptent une population moyenne de 200 000 personnes et que le statut socio-économique des personnes qui y habitent est loin d'être complètement homogène. Néanmoins, les taux de mortalité dus au cancer du poumon étaient plus de deux fois plus élevés dans les DSC relativement plus pauvres (Verdun et St-Luc), comparativement aux DSC relativement riches (l'ouest de l'île et Ste-Justine). De plus, quand on a utilisé les districts des centres locaux de services communautaires (CLSC) plus petits et relativement plus homogènes comme unité d'analyse, les différences dans les indicateurs de l'issue des traitements étaient encore plus prononcées.

Pour effectuer des comparaisons entre plusieurs régions, je préfère utiliser des taux qui ont subi une standardisation directe plutôt qu'une standardisation indirecte, afin que le même ensemble de poids soit appliqué aux taux particuliers dans chaque région. (Dans l'exemple précédent, c'est seulement parce que les résultats devaient figurer avec des données régionales déjà publiées qui avaient été standardisées indirectement que l'on a utilisé des poids standardisés indirectement.) Il arrive parfois que les populations locales renferment des nombres disproportionnés d'hommes (ou de femmes), dans un tel cas, il est important d'effectuer une standardisation selon le sexe ainsi que selon l'âge à des fins de comparaison.

3.5 Tendances dans la mortalité infantile selon le revenu de 1971 à 1986

La figure 2 montre des taux de mortalité infantile selon le quintile du revenu en 1971 et en 1986 (Wilkins, Adams et Brancker 1989). En 1971, tout comme en 1986, les secteurs de recensement dans chaque région métropolitaine de recensement ont été combinés en cinq quintiles renfermant approximativement le même nombre de personnes, selon le pourcentage de la population vivant sous les seuils de pauvreté applicables à l'époque. Il faut remarquer que, bien que le même secteur de

Figure 2



SOURCE: STATISTIQUE CANADA / SANTÉ ET BIEN-ÊTRE SOCIAL CANADA (WILKINS, ADAMS & BRANCKER, 1989)

recensement ne soit pas nécessairement classé dans le même quintile de revenu au cours de chaque période, le principe utilisé pour la classification était le même dans les deux cas. Ce genre d'analyse fournit la seule preuve que nous ayons des tendances à long terme dans l'évolution des inégalités en matière de santé au Canada -- bien que la réduction des inégalités en matière de santé soit officiellement reconnue comme un des buts fondamentaux de la politique sanitaire canadienne. Dans ce cas, nous constatons que même si au cours des deux périodes les taux de mortalité infantile dans le 5e quintile étaient presque deux fois plus élevés que dans le 1er quintile (les rapports entre les taux étaient semblables), la différence en valeur absolue entre les taux en 1986 n'était que la moitié de ce qu'elle était en 1971.

3.6 Taux de lésions des enfants (piétons et cyclistes) à Montréal

On peut aussi analyser de cette façon d'autres ensembles de données recueillies localement ou spécialement -- par exemple, les données sur les lésions subies par les enfants, qu'ils soient piétons ou cyclistes, qui ont entraîné l'hospitalisation ou la production de rapports de police (Dougherty, Pless et Wilkins 1990). Après avoir codé le secteur de recensement du lieu de résidence des victimes à partir de leur code postal et de leur adresse, on a trouvé que les taux de lésions étaient presque cinq fois plus élevés dans les districts les plus pauvres de Montréal comparativement aux districts les moins pauvres. Dans ce cas, on s'attendrait à ce que les conditions dangereuses dans un quartier aient un effet sur toutes les personnes qui habitent le district, peu importe leur revenu familial, de sorte que l'emplacement géographique devrait être important en soi, et pas seulement comme indicateur du statut socio-économique d'un quartier. On devrait disposer simultanément des caractéristiques des particuliers (ou des familles) et des quartiers avant de pouvoir espérer démêler ces deux effets.

3.7 D'autres façons d'utiliser des données sur les régions locales tirées des fichiers de santé

Les analyses qui n'utilisent que les trois premiers caractères du code postal peuvent aussi fournir des renseignements très utiles au niveau infra-municipal dans les régions urbaines plus grandes. Cela peut être nécessaire, par exemple, quand on ne dispose pas du code postal complet à six caractères (Choinière 1991), quand les données sur l'adresse ne sont pas disponibles afin de permettre le traitement des codes postaux inexacts, manquants ou qui posent des problèmes (Gentleman, Wilkins, Nair et Beaulieu 1991) ou quand on veut effectuer une analyse rapide et pas trop précise d'un ensemble de données renfermant des adresses et des codes postaux complets. Dans ce genre d'analyse, les rapports observés sont généralement moins étroits que ceux que l'on constate dans des analyses basées sur un codage géographique plus précis.

Dans l'étude déjà mentionnée sur les issues de grossesse selon le revenu (Wilkins, Sherman et Best 1991), on a utilisé le rapport entre les naissances et la population du recensement âgée de moins d'un an pour estimer l'importance et la répartition du sous-dénombrement du recensement touchant les familles avec de jeunes enfants et pour effectuer cette évaluation selon la région métropolitaine de recensement et le revenu du quartier. Les résultats de cette opération se comparaient bien aux estimations basées sur la contre-vérification des dossiers (Germain 1988; Statistique Canada 1988), mais ils étaient disponibles à un niveau de regroupement beaucoup plus détaillé.

Dans l'étude sur la mortalité selon le revenu (Wilkins, Adams et Brancker 1989), on a aussi utilisé les codes postaux comme un moyen pour déterminer les personnes placées en établissement. De plus, quand les taux de mortalité étaient compilés selon le lieu de naissance et le revenu du quartier, il semblait que la mortalité généralement plus faible des immigrants, qui avaient aussi une probabilité plus élevée de vivre dans des quartiers relativement plus pauvres des grandes villes, créait une certaine confusion, au niveau du quartier, pour ce qui est du rapport entre le revenu et la mortalité.

4. CONCLUSION

On peut utiliser les données sur la santé codées selon les régions locales pour décrire l'état de santé de la population dans une région géographique donnée, regroupée selon les zones de services administratives locales ou de programmes ou regroupée selon des caractéristiques du quartier telles que le pourcentage de la population vivant sous le seuil de faible revenu.

Quand on établit un rapport entre des données sur la santé codées selon la région locale et des données du recensement ou autres sur des caractéristiques du quartier pour les mêmes régions, divers types d'analyses sont possibles. On peut donc s'attaquer, du moins indirectement, à des questions d'équité, d'environnement et d'efficacité administrative, sans avoir à recueillir d'autres données. Des exemples d'un tel travail comprennent des études des tendances relatives à la mortalité infantile et à l'insuffisance pondérale selon le revenu du quartier dans les régions urbaines du Canada. Pour la population, excluant les pensionnaires d'établissements institutionnels, vivant dans des régions urbaines, on peut aussi effectuer des analyses plus simples en n'utilisant que les trois premiers caractères du code postal, même quand on ne dispose pas de l'adresse complète.

BIBLIOGRAPHIE

Choinière, R. (1991). Les disparités géographiques de la mortalité dans le Montréal métropolitain, 1984-1988: étude écologique des liens avec les conditions sociales, économiques et culturelles, *Cahiers québécois de démographie*, 20, 1, 115-144.

Dougherty, G., Pless, B., et Wilkins, R. (1990). Social class and the occurrence of traffic injuries and deaths in urban children, *Revue canadienne de santé publique*, 81, 204-209.

Gentleman, J. F., Wilkins, R., Nair, C., et Beaulieu, S. (1991), Analyse de la fréquence des interventions chirurgicales au Canada, *Rapports sur la Santé*, 3, 4, 291-309.

Germain, M.-F. (1988). Taux de sous-dénombrement pour la variable revenu total par groupes d'âge-sexe (Note de service), Ottawa: Section de la qualité des données du recensement, Division des méthodes d'enquêtes sociales, Statistique Canada.

Hoey, J., Wilkins, R., Gagnon, G., et O'Loughlin, J. (1987). L'état de santé des Québécois: un profil par région socio-sanitaire et par département de santé communautaire, dans Commission d'enquête sur les services de santé et les services sociaux (Commission Rochon), *Programme de recherche: recueil des résumés*, Québec: Les Publications du Québec, 135-138.

Statistique Canada (1988). Taux de sous-dénombrement provenant de la contre-vérification des dossiers de 1986 (Recensement du Canada 1986, Bulletin d'information à l'intention des utilisateurs, numéro 2), Ottawa: Statistique Canada.

Statistique Canada (1991). Fichier de conversion des codes postaux (version de janvier 1991), Ottawa: Division de la géographie, Statistique Canada.

Wilkins, R., Sherman, G. J., et Best, P. A. F. (1991). Issues de grossesse et mortalité infantile selon le revenu dans les régions urbaines du Canada en 1986, *Rapports sur la Santé*, 3, 1, 7-31.

Wilkins, R., Adams, O., et Brancker, A. (1989). Évolution de la mortalité selon le revenu dans les régions urbaines du Canada entre 1971 et 1986, *Rapports sur la Santé*, 1, 2, 137-174.

Recueil du Symposium 91 de Statistique Canada
Questions spatiales reliées aux statistiques
Novembre 1991

APPLICATIONS SPATIALES ET STATISTIQUES DE DONNÉES GÉOCHIMIQUES ENVIRONNEMENTALES À DES QUESTIONS DE SANTÉ HUMAINE

D.R. Boyle[1]

RÉSUMÉ

On peut désigner par l'expression cartographie de la «sensibilité» géochimique les enquêtes géochimiques régionales qui permettent d'illustrer la répartition des éléments en trace et des facteurs géochimiques dont on sait qu'ils possèdent certaines composantes de risque (carence, toxicité) pour la santé. Des éléments ou facteurs de ce genre peuvent être mesurés dans divers milieux liés directement ou indirectement à la chaîne humain-aliment-eau. Dans ce type d'étude, l'accent est mis sur la caractérisation, tant spatiale que statistique, des régions où l'on relève des concentrations anormalement élevées ou faibles d'un paramètre particulier qui joue un rôle présumé (ou démontré) dans une maladie donnée.

La «transformation» des données géochimiques régionales relatives aux milieux absorbés directement, comme l'eau, en courbes de niveau liées à des effets dose-réponse connus, permet au géochimiste travaillant dans le domaine de l'environnement de mieux décrire l'incidence des données géochimiques aux responsables de la planification des services sanitaires et environnementaux.

MOTS CLÉS: Concept dose-réponse; cartographie de la «sensibilité» géochimique.

1. INTRODUCTION

La géochimie étudie l'abondance des éléments et des nucléides dans la croûte terrestre, la répartition des éléments dans les phases géochimiques de la terre ainsi que les lois qui gouvernent l'abondance et la répartition de ces éléments. La principale méthode appliquée est l'analyse géochimique détaillée des facteurs qui règlent l'abondance et la répartition des éléments dans les cinq sphères géochimiques terrestres (lithosphère, pédosphère, hydrosphère, biosphère, atmosphère).

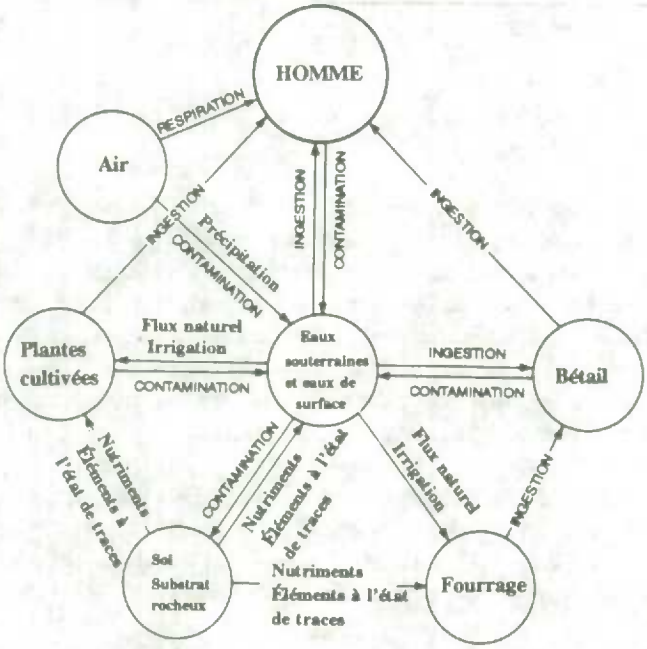
L'homme interagit avec toutes les sphères de son milieu naturel (Figure 1) et il reçoit ses nutriments de sources diverses (plantes cultivées, bétail, eau, etc.). Comme le montre la Figure 1, si les eaux peuvent recevoir un élément ou un composé donné en concentration anormale par suite de processus naturels ou d'une contamination, l'homme, lui, reçoit ces éléments ou contaminants non seulement de l'eau, mais de nombreuses autres sources. En conséquence, si un élément fait défaut dans un milieu géochimique particulier, l'homme, qui est placé au début de la chaîne alimentaire animale, retrouve la déficience dans son approvisionnement local en eau et en aliments.

Ainsi, le milieu géochimique peut jouer un rôle primordial dans la répartition géographique des troubles pathologiques et nutritionnels de l'homme. On en trouve des exemples manifestes dans : a) les carences régionales d'éléments essentiels comme l'iode (goitre, hypothyroïdisme, crétinisme), le fer (anémie), le calcium, le magnésium et le sodium (affections cardio-vasculaires, hypocalcémie), le chrome (régulation du glucose, diabète), le cobalt (anémie pernicieuse), le zinc (parakératose, troubles enzymatiques, mauvaise cicatrisation des plaies), le fluor (carie dentaire, ostéoporose ?) et le sélénium (cardiomyopathie, prévention du cancer) et b) les surabondances régionales, attribuables à certains processus géochimiques, de toxiques non essentiels comme le cadmium (dysfonction rénale, hypertension, affection cardiaque), le plomb (neuropathie, troubles psychotiques,

[1] D.R. Boyle, Énergie, Mines et Ressources, 601, rue Booth, Ottawa (Ontario), Canada K1A 0E8.

cadmium (dysfonction rénale, hypertension, affection cardiaque), le plomb (neuropathie, troubles psychotiques, hypertension), le mercure (neuropathie), l'aluminium (maladie d'Alzheimer ?), l'arsenic (cancer) et les éléments radioactifs (cancer).

Figure 1: Liens entre l'homme et le milieu. Toutes les voies mènent éventuellement à l'homme (d'après Boyle 1991).



La composition de l'eau potable, qui est réglée par certains facteurs chimiques, climatiques et géologiques du milieu géochimique, peut avoir un lien causal avec des affections comme les maladies cardio-vasculaires (dureté de l'eau), l'ostéoporose (disponibilité du calcium et du fluor), la mortalité infantile (concentrations élevées de nitrates, de cuivre ou de magnésium), les troubles neurologiques et psychotiques (mobilisation du Cd, du Pb, du Hg, du Tl et de l'Al par les pluies acides), l'hypertension (concentrations élevées de Na, de Ba, de Cd ou de Pb) et le cancer (concentrations élevées d'As, de Ra ou de Rn).

Lorsqu'elles sont exprimées sous forme de courbes d'égale abondance, de rapports ou de facteurs, les données géochimiques régionales provenant de divers milieux (p. ex., eau, sol, roches, végétation) peuvent être utilisées directement (concepts dose-réponse) et indirectement (cartographie de la sensibilité géochimique) pour aider le personnel sanitaire à réaliser :

a) des évaluations de l'exposition : - afin de déterminer le pourcentage d'une population qui est susceptible d'être atteint de maladies ou de troubles liés à la carence ou à la toxicité d'un élément ou d'un composé;

b) des évaluations du risque : - pour concevoir des modèles servant à prévoir les effets possibles sur la santé d'éléments ou de composés particuliers dans des milieux géologiques semblables qui n'ont pas fait l'objet d'enquêtes géochimiques;

c) des enquêtes épidémiologiques : - qui peuvent servir à établir un lien associatif ou causal entre les éléments ou les composés du milieu géochimique et certaines affections (p. ex., la présence de radium dans les eaux souterraines et la survenue du cancer osseux); elles peuvent également permettre de déterminer les conséquences possibles pour la santé de la perturbation par l'homme d'un milieu géochimique donné (p. ex., pluies acides et maladie d'Alzheimer);

d) des études de la migration du milieu rural vers le milieu urbain : - modifications de l'état de santé d'un groupe rural qui quitte un milieu géochimique ou géologique particulier pour s'établir dans un milieu urbain;

e) des études d'aménagement des terres : - étude des effets sanitaires possibles de l'établissement de populations dans des régions non favorables du point de vue géochimique (p. ex., aménagement des terres dans les régions de fortes concentrations de fluor du bassin sédimentaire carbonifère des Maritimes);

f) des études de la migration des groupes ethniques: - déplacement d'un groupe ethnique d'un milieu géochimique à un autre (p. ex., déplacement des Ukrainiens des steppes ukrainiennes vers les prairies canadiennes);

g) des études des voies chimiques : - études de l'abondance d'un élément et de son passage d'une espèce à une autre du point de vue des effets possibles sur la santé (p. ex., mobilisation, concentration et méthylation du mercure dans la chaîne roche-sol- eau-végétation-homme);

h) des corrections relatives aux concentrations naturelles : - normalisation des données sur l'exposition par addition ou soustraction des concentrations naturelles d'un paramètre environnemental (p. ex., addition des niveaux naturels de rayonnement dans les études sur les doses radioactives industrielles).

Dans toutes les études énumérées plus haut, il faut également prendre en compte les facteurs de perturbation autres que ceux qui sont attribuables à l'environnement géochimique (p. ex., mode de vie, régime alimentaire et tabagisme).

2. MÉTHODES STATISTIQUES ET SPATIALES D'ANALYSE GÉOCHIMIQUE

De nombreuses méthodes statistiques de représentation spatiale ont été appliquées à l'interprétation des données géochimiques; la plupart d'entre elles ont été employées en exploration minière. Les méthodes statistiques et spatiales de présentation des données géochimiques environnementales en sont encore à leurs débuts. Les présentations spatiales employées dans le présent article ont été générées par le système UNIRAS[2] : la représentation sur quadrillage a été effectuée à l'aide du progiciel d'interpolation GEOINT et la présentation cartographique finale a été réalisée au moyen du progiciel de traçage GEOPACK. Le système comprend également un sous-programme d'interpolation par la technique de krigeage (KRIGPAK) et une interface usager facultative (UNIMAP) employée en cartographie interactive. Bon nombre de systèmes semblables sont utilisés par les environnementalistes. Quel que soit le progiciel de traçage utilisé, le géochimiste doit s'assurer que les données (signes conventionnels ou courbes) renferment, selon l'application prévue, les concentrations maximales admissibles, les limites esthétiques, les concentrations dose-réponse ou toute autre limite prescrite par un règlement, et que ces données soient présentées sous forme d'éléments distincts, de manière que le personnel de la santé ou de l'environnement dispose de limites qu'il peut facilement reconnaître, et même faire appliquer en vertu de la loi, et avec lesquelles il est possible de délimiter les régions d'abondance ou de carence et, par conséquent, d'identifier les populations à risque. C'est ainsi que nous avons procédé dans le cas des éléments ou des paramètres auxquels les limites indiquées plus haut peuvent s'appliquer. En ce qui a trait aux autres éléments, les intervalles entre les courbes ont été choisis de façon à mieux faire ressortir les secteurs de déplétion et de richesse anormale à l'intérieur d'une région donnée.

3. CARTOGRAPHIE DE LA SENSIBILITÉ GÉOCHIMIQUE RÉGIONALE

De manière générale, avant de déterminer l'existence d'un lien de cause à effet dans l'étiologie d'une maladie ou d'un trouble, il faut établir les liens ou corrélations entre la maladie ou le trouble et les facteurs présumés de risque sanitaire. Pour y arriver, il faut établir des modèles d'étude et choisir les régions à étudier. En outre, il est essentiel de connaître la valeur des facteurs de risque en question dans le milieu naturel. Comme il a été indiqué précédemment, une partie au moins de l'étiologie ou du risque global lié à de nombreuses maladies peut être attribuée à une carence ou à un excès d'éléments majeurs ou présents à l'état de traces dans l'environnement. Pour que les études sanitaires détaillées soient efficaces, il faut disposer de méthodes

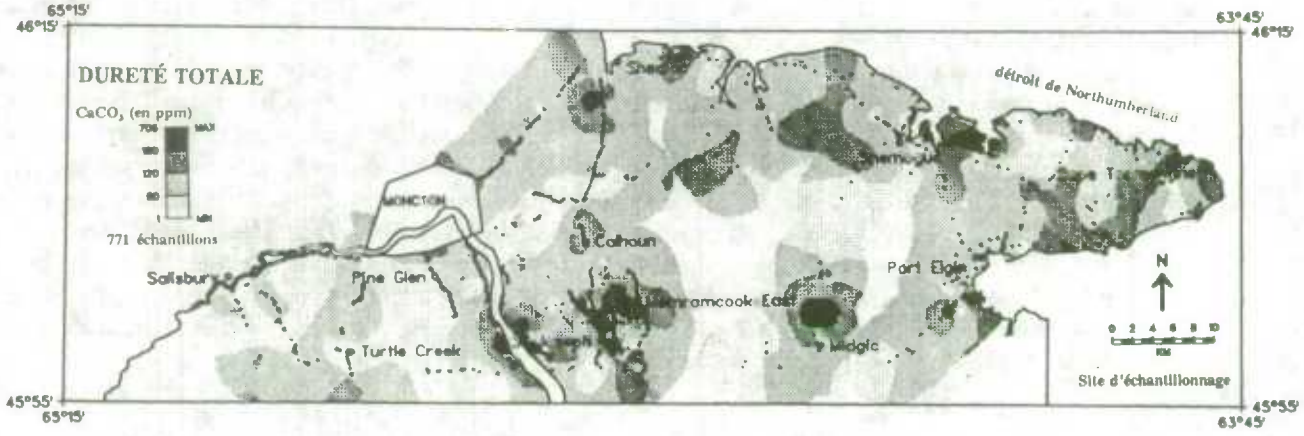
[2] European Software Contractors A/S, Norregade, Danemark.

permettant de cartographier la répartition des facteurs de risque sanitaire, soit, dans le cas présent, la répartition des éléments en quantité toxique ou en quantité insuffisante dans l'environnement.

Les enquêtes géochimiques régionales, qui illustrent sous forme de courbes ou de signes conventionnels, la répartition des éléments à l'état de traces ou des facteurs qui présentent effectivement certains risques (carence, toxicité) pour la santé humaine, peuvent être désignées par l'expression cartographie de la «sensibilité» géochimique. Ces éléments ou paramètres peuvent être mesurés dans divers milieux liés directement ou indirectement à la chaîne homme-aliments-eau-air. Ce genre d'étude met l'accent sur la délimitation des régions peuplées qui se caractérisent par des concentrations anormalement élevées ou anormalement faibles d'un paramètre jouant une rôle présumé (ou démontré) dans une maladie donnée. Les populations à risque de ces régions (population en général ou groupes d'âge, de sexe ou de race) de même que les populations naturelles adjacentes, peuvent être incluses dans un système d'échantillonnage d'épidémiologie environnementale afin d'étudier l'étiologie de la maladie en question. Il importera, au moment d'évaluer l'ampleur du risque sanitaire d'un paramètre donné, de connaître précisément les voies par lesquelles l'élément peut atteindre l'homme, de même que les formes sous lesquelles il lui est accessible. Des exemples de cartographie de la «sensibilité» géochimique sont présentés ci-après afin d'illustrer l'utilité de ce type d'enquête.

Une recherche considérable sur la relation existant entre les maladies cardio-vasculaires et la composition de l'eau potable a abouti à la conclusion qu'il existe, dans les eaux dures, un ou des «facteurs eau» qui offrent un certain degré de protection contre les maladies cardiaques et que les personnes qui consomment de l'eau douce courent un risque légèrement plus élevé de maladie cardio-vasculaire que celles qui consomment de l'eau dure (Shaper et coll. 1980; Calabrese et coll. 1980; Lacey et Shaper 1984). L'isolement du ou des «facteurs eau» contribuerait à la prévention des maladies cardiaques; même s'il se situe loin derrière le tabagisme et le régime alimentaire, ce facteur n'en demeure pas moins important.

Figure 2: Dureté des eaux souterraines de la région de Moncton, bassin sédimentaire carbonifère des Maritimes (données non publiées).

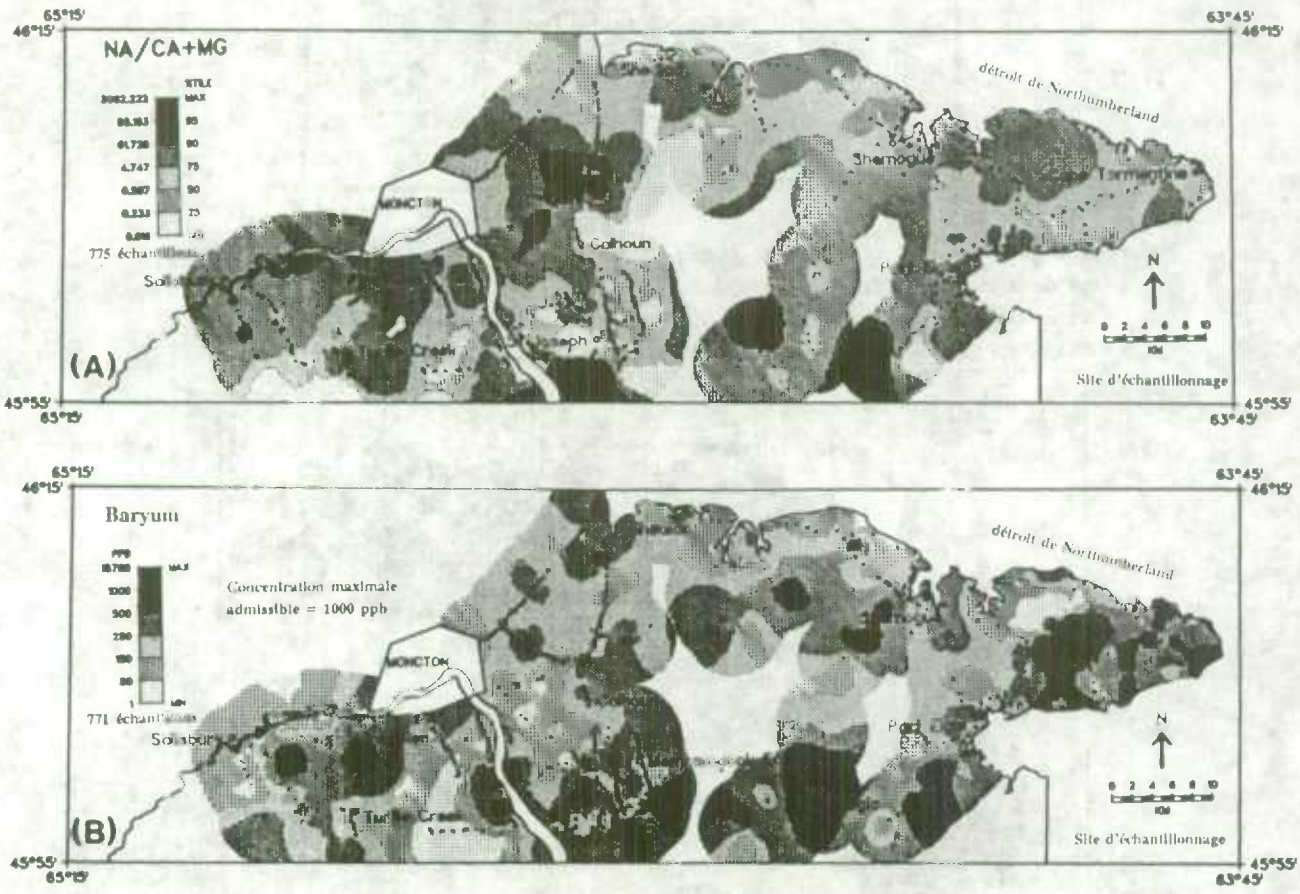


Les éléments chimiques les plus souvent liés à l'étiologie ou à la prévention des affections cardiaques sont le Ca, le Mg, le Na, le Ba, le Cd, le Pb et le Se. Le calcium et le magnésium sont des éléments essentiels à la fonction cardiaque et un pourcentage appréciable de l'apport quotidien de calcium, et plus particulièrement de l'apport de magnésium, peut provenir de l'eau potable (Hopps et Feder 1986). Ce sont ces deux éléments qui contribuent le plus à la dureté de l'eau; des cartes comme celle qui est présentée dans la figure 2, qui indique la dureté des eaux souterraines de la région de Moncton au Nouveau-Brunswick selon différent degrés de dureté (faible : 0-60 meq/L; moyen : 60-120 meq/L; élevé : >120 meq/L), peuvent être utilisées par les épidémiologistes pour choisir les groupes de population exposés et non exposés sur lesquels sera étudié l'effet du paramètre sur la survenue de la maladie. Le sodium, le baryum et, dans une moindre mesure, le cadmium et le plomb sont des «éléments hypertensifs»; un apport élevé d'un seul ou d'une combinaison de ces éléments peut entraîner une

élévation de la tension artérielle et peut-être une affection cardiaque. La répartition du Ba dans les eaux souterraines de la région de Moncton (Figure 3, A) et le rapport Na/Ca+Mg de ces eaux (Figure 3, B) peuvent tous deux être employés, le rapport Na/Ca+Mg plus efficacement peut être, en conjonction avec la répartition de la dureté de l'eau (Figure 2) pour obtenir une interprétation plus détaillée de l'exposition des groupes de population à certains «facteurs eau» qui peuvent être liés aux affections cardiaques. D'autres approches, comme l'étude des effets combinés du Na+Ba+Cd+Pb (données normalisées), pourraient être employées pour déterminer l'indice hypertensif des eaux potables chez les groupes de population étudiés. Le sélénium, qui est censé protéger l'homme contre certains types d'affections cardiaques, surtout les affections du myocarde (Chen et coll. 1980), peut aussi être inclus dans une étude approfondie de l'apport d'éléments à l'état de traces et des affections cardiaques.

Lorsque la charge atmosphérique de polluants acides excède la production de base issue de l'altération atmosphérique de la couverture et du substrat rocheux, le pH de même que la qualité des eaux souterraines peuvent être gravement touchés. Les eaux souterraines acides peuvent avoir un effet nuisible sur la santé humaine en raison de la solubilité et de la mobilité plus élevées des métaux lourds toxiques (p. ex., le Pb, le Cd, l'As, le Cu et l'Al). Ces métaux peuvent se dissoudre à partir de la lithosphère (altération atmosphérique) et des conduites de plomberie domestiques (solvabilité du plomb). Des cartes d'isocourbes du pH des eaux souterraines, comme la carte d'une partie des provinces Maritimes présentée dans la Figure 4, peuvent servir à délimiter les régions qui pourraient être sensibles aux charges atmosphériques acides et, en combinaison avec les cartes de densité de la population, peuvent être employées pour déterminer le pourcentage de la population à risque en raison de la plus grande mobilité des éléments toxiques. Il faudrait procéder à des études de la mobilité des éléments avant de porter un jugement global sur le sujet.

Figure 3: Répartitions du rapport sodium/calcium + magnésium (B), et du baryum (A), dans les eaux souterraines de la région de Moncton, bassin sédimentaire carbonifère des Maritimes (données non publiées).



La plupart des rayonnements auxquels l'homme est exposé proviennent du milieu naturel et se composent principalement de rayons gamma. C'est pourquoi les cartes qui indiquent les niveaux naturels de rayonnement jouent un rôle important dans la détermination des populations à risque en ce qui concerne l'exposition de la surface corporelle aux rayonnements et l'inhalation de composants radioactifs comme le radon. La Commission géologique du Canada a étudié le transport atmosphérique du rayonnement gamma sur plus de 30 % du territoire canadien. Les données relatives aux niveaux naturels de rayonnement au Canada sont présentées par Grasty et coll. 1984, et sont accompagnées de la méthode de calcul des doses moyennes estivales et annuelles de rayonnement absorbées en plein air par les Canadiens.

Figure 4: pH des eaux souterraines de la région centrale est du bassin sédimentaire carbonifère des Maritimes (données de Dyck 1976).

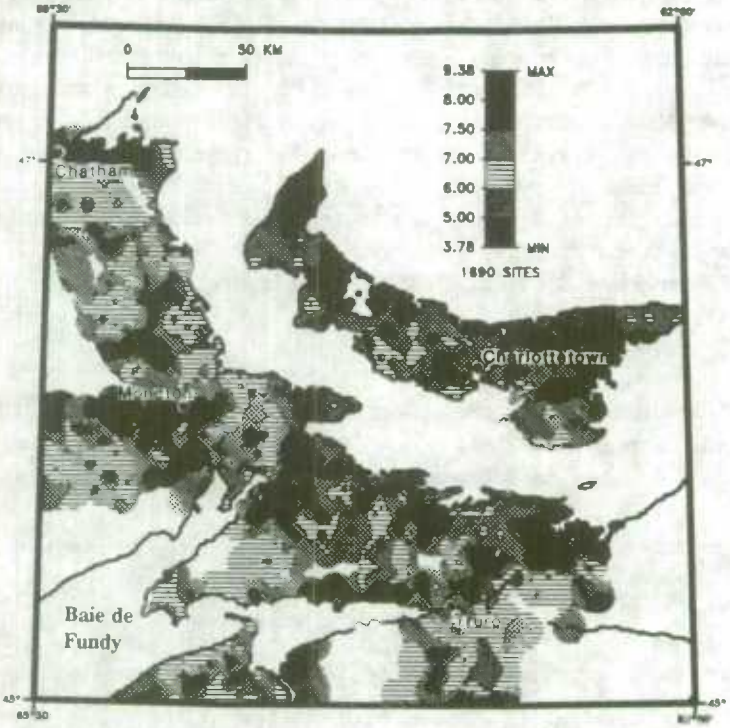
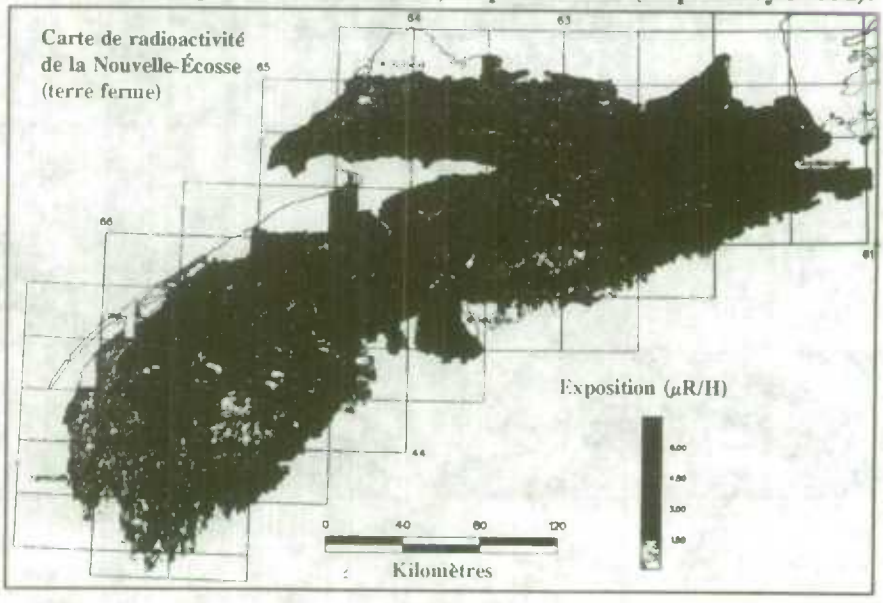


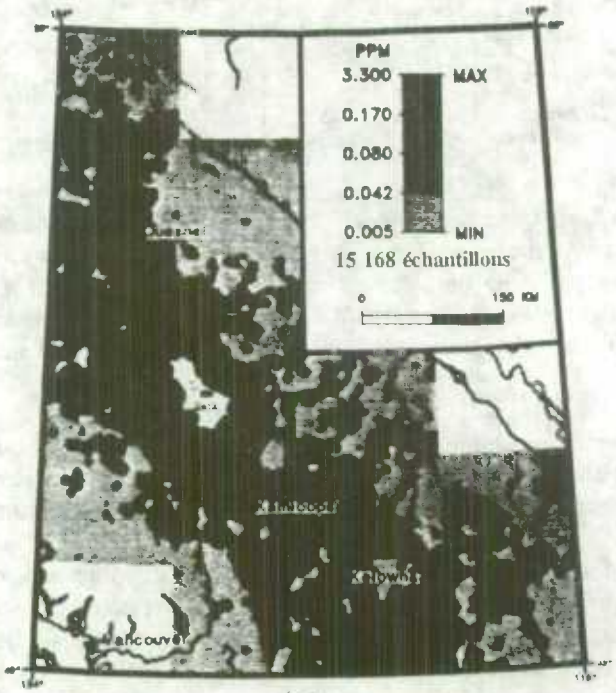
Figure 5: Exposition au rayonnement gamma (μR/H) en provenance du substrat rocheux et de la couverture en Nouvelle-Écosse (terre ferme). Les régions de forte et de faible expositions correspondent généralement aux régions sises sur des roches granitiques et sédimentaires, respectivement (d'après Boyle 1991).



Les agents sanitaires et les épidémiologistes étudiant les rayonnements et leurs effets sur la santé humaine peuvent utiliser des cartes comme celle qui est présentée dans la Figure 5 pour déterminer les charges naturelles de rayonnement reçues par un groupe de population à l'étude; ces cartes peuvent également servir à sélectionner les groupes de population à étudier. En outre, ce type de données peut être utilisé par les agents sanitaires et les responsables de la planification du code du bâtiment afin d'évaluer si les niveaux accrus de rayonnement domestique entraînent des risques pour la population et de déterminer les mesures nécessaires à réduire le risque en question. Les agents sanitaires qui étudient l'exposition au rayonnement en milieu de travail, à l'échelle régionale ou nationale, peuvent employer les cartes d'exposition comme celle de la Figure 5 pour calculer l'exposition totale au rayonnement gamma (milieu naturel et milieu de travail). Ainsi, les habitants de la Nouvelle-Écosse qui vivent sur un terrain granitique reçoivent de deux à quatre fois plus de rayonnement gamma à l'heure que ceux qui vivent sur des roches sédimentaires (voir la légende de la Figure 5); ces grands écarts de l'exposition naturelle pourraient modifier de façon importante l'interprétation des données relatives à l'exposition en milieu de travail.

Les exemples précédents mettent en évidence l'effet assez direct que les charges d'éléments et de rayonnement peuvent avoir sur la santé humaine. Cependant, les cartes de répartition de la concentration des éléments dans divers types de milieux géochimiques superficiels non ingérés par l'homme peuvent aussi agir comme indicateurs «sensibles» des maladies liées à la toxicité ou à la carence d'un élément. Ainsi, la répartition du fluor dans les eaux de ruisseau de la plus grande partie de la moitié sud de la Colombie-Britannique (Figure 6) indique une forte concentration de fluor dans un certain nombre de régions. Comme les caractéristiques géochimiques du fluor dans les eaux de surface et dans les eaux souterraines peu profondes sont semblables, les modèles de concentration de cet élément dans les eaux de surface sont, de manière générale, facilement applicables aux eaux souterraines. Dans la plupart des cas, dans les eaux souterraines peu profondes utilisées comme sources d'eau potable, la concentration de fluor est plus élevée que dans les eaux de surface adjacentes et ce, d'un facteur de dix environ. Par conséquent, il est possible d'appliquer un tel facteur aux données de la Figure 6 et ainsi, de délimiter les régions dans lesquelles peuvent survenir des affections liées à la toxicité ou à la carence de l'élément. Employées conjointement avec des cartes de densité de la population, les enquêtes de ce type peuvent être utilisées efficacement dans des évaluations de l'exposition. Cette méthode d'interprétation des données relatives aux concentrations de fluor dans les eaux de surface se compare à la méthode décrite ci-dessous pour l'interprétation de l'effet dose-réponse du fluor dans les sources souterraines rurales d'eau potable.

Figure 6: Répartition du fluor dans les eaux de ruisseau de la partie centrale sud de la Colombie-Britannique (les données sur le fluor proviennent de la base de données du Programme d'exploration géochimique préliminaire de la Commission géologique du Canada).



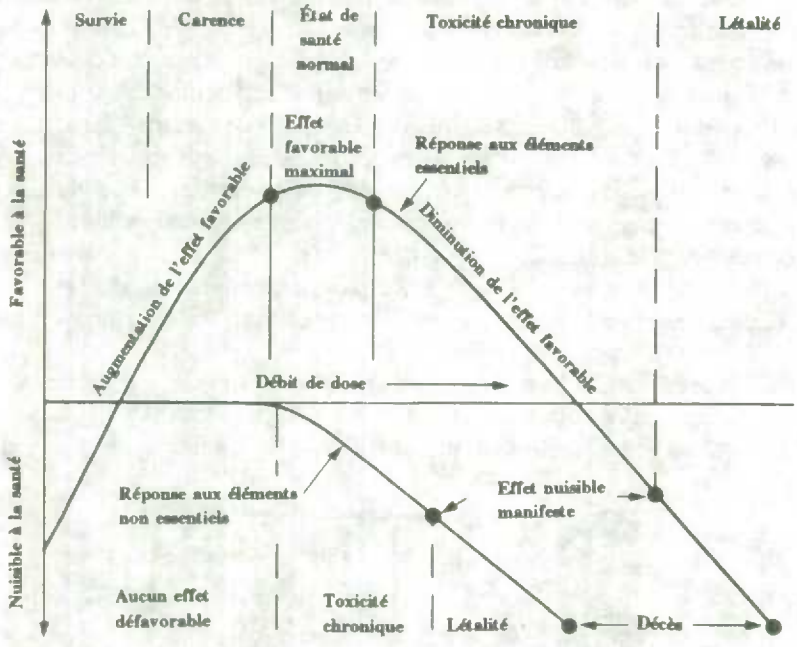
É.-U.

Les cartes présentées ici ne constituent que quelques exemples des nombreuses cartes de «sensibilité» géochimique qui pourraient être dressées. Ce genre de données géochimiques régionales pourrait également servir à cartographier des facteurs additifs et multiplicatifs susceptibles de présenter un risque pour la santé humaine.

4. APPLICATION DES CONCEPTS DOSE-RÉPONSE AUX DONNÉES GÉOCHIMIQUES

Les courbes dose-réponse (Figure 7) décrivent les attributs favorables et nuisibles à la santé qui touchent les populations humaines, selon la dose absorbée par unité de temps (habituellement 24 heures) pendant une période déterminée. Les attributs favorables peuvent se rapporter à l'état de santé général (faible morbidité) ou à une absence relative de certaines maladies (p. ex., carie dentaire, goitre, affections cardio-vasculaires). Les attributs nuisibles se manifestent par les maladies liées à la carence ou à la toxicité d'un élément. Ces courbes peuvent être employées pour décrire l'état de santé potentiel des populations humaines par rapport à l'ingestion ou à l'absorption de certaines concentrations d'un élément ou d'un composé (essentiel et non essentiel).

Figure 7: Représentation schématique des courbes dose-réponse relatives aux éléments essentiels et non essentiels (d'après Boyle 1991).



L'interprétation des données fondée sur des concepts dose-réponse doit prendre en compte le ou les modèles appliqués ainsi que toutes les hypothèses posées pour construire la courbe. Les courbes peuvent être dressées à l'aide des quantités ingérées, partielles ou totales, d'un élément ou d'un composé donné par unité de temps. Les quantités partielles ingérées (p. ex., le fluor dans l'eau potable) peuvent servir à des fins d'interprétation s'il est possible de déterminer l'apport alimentaire quotidien moyen en provenance d'autres sources (p. ex., aliments et boissons).

La «transformation» des données géochimiques régionales en courbes de niveau qui se rapportent à des effets dose-réponse connus permet au géochimiste travaillant dans le domaine de l'environnement de mieux décrire l'incidence des données géochimiques aux responsables de la planification sanitaire et environnementale.

Un exemple de l'application des concepts dose-réponse aux données géochimiques régionales est présenté dans les figures 8 et 9. Dans les deux régions, des eaux de sources domestiques ont été échantillonnées, à raison d'un prélèvement par 10 km^2 environ, et analysées en vue de déceler la présence de fluor et d'un certain nombre d'éléments et de paramètres relatifs à la qualité de l'eau (Dyck et coll. 1976 et Dyck 1980). Les données sur le

fluor sont exprimées à l'aide d'une échelle de grisés correspondant à une courbe dose-réponse schématique qui montre les effets potentiels sur la santé à divers intervalles de concentrations. Le plateau favorable représente l'intervalle de concentration du fluor qui est jugé suffisant au bon développement dentaire (et peut-être osseux) à la latitude de ces deux régions et d'après une consommation quotidienne de 1,5 L d'eau. Lors de l'interprétation de ces données, il est essentiel de mesurer ou de présumer la quantité de fluor fournie par le régime alimentaire. De manière générale, on appliquerait un apport alimentaire moyen au groupe de population (p. ex., nourrissons, enfants jusqu'à l'âge de 15 ans, vieillards, certains groupes prédisposés) auquel doivent s'appliquer les courbes de dose-réponse relatives à la présence du fluor dans l'eau. Utilisées conjointement avec les cartes de densité de la population, les interprétations des courbes dose-réponse des figures 8 et 9 peuvent servir à déterminer l'effet possible de la carence et de l'excès de fluor sur la santé de populations régionales données. Les régions dans lesquelles on juge que l'apport total de fluor (eau, aliments, additif pour la santé dentaire) s'écarte de façon importante de la norme (concentrations très élevées ou très faibles) pourraient faire l'objet d'études sanitaires plus détaillées. L'interprétation de la courbe dose-réponse du fluor dans les eaux souterraines des provinces Maritimes (Figure 8) montre que de vastes régions sont caractérisées par des concentrations très faibles de fluor et qu'un certain nombre de régions distinctes montrent des concentrations très élevées de fluor; très peu de sources se situent dans l'intervalle favorable à la santé. Ces données peuvent être comparées aux données de la partie sud de la Saskatchewan (Figure 9) où la plupart des concentrations de fluor se situent à la limite inférieure, ou juste au-dessous de cette limite, de l'intervalle favorable à la santé; très peu d'endroits montrent une concentration élevée de fluor. Dans les deux régions, la majorité de la population (100 % de la population à l'Île-du-Prince-Édouard) tire son approvisionnement en eau de sources souterraines.

Figure 8: Carte dose-réponse de la concentration de fluor dans les eaux souterraines de la région centrale est du bassin sédimentaire carbonifère des Maritimes (données sur le fluor, Dyck 1976).

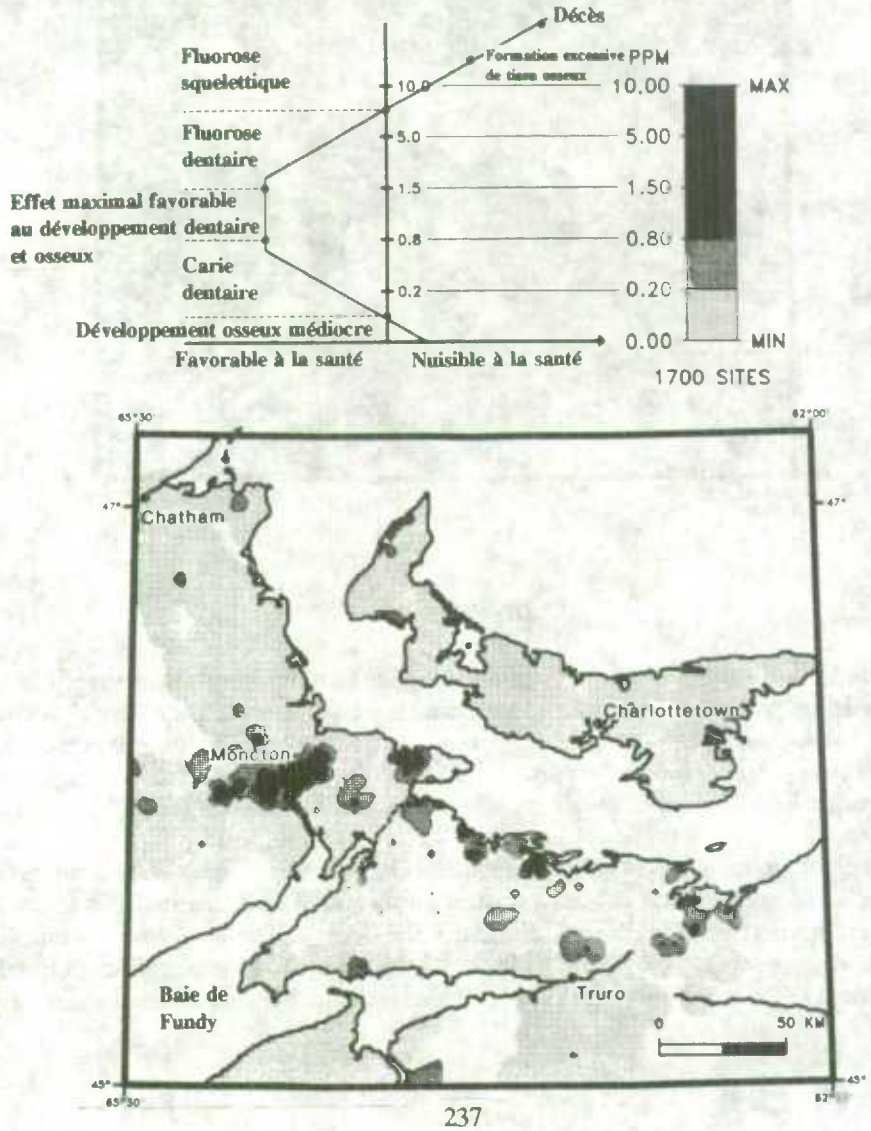
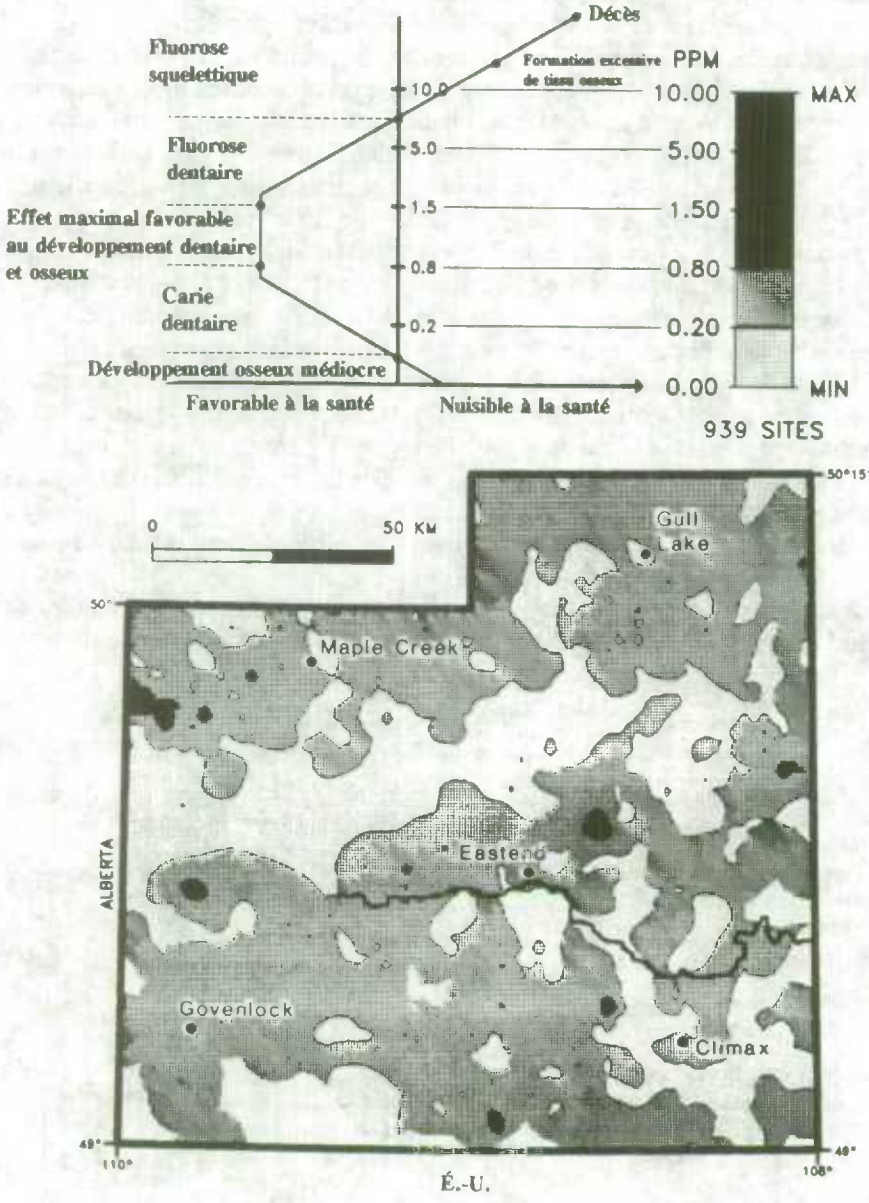


Figure 9: Carte dose-réponse du fluor dans les eaux souterraines de la région sud-ouest de la Saskatchewan (données sur le fluor, Dyck 1980).



5. DISCUSSION

Il est extrêmement important que les environnementalistes quantifient les paramètres susceptibles d'agir sur la santé humaine et qu'ils les présentent sous une forme permettant au personnel sanitaire d'évaluer le degré de risque sanitaire lié à chaque paramètre ou à chaque groupe de paramètres. De toute évidence, les répartitions globales de ces paramètres joueront un rôle important dans la détermination de la taille de la population à risque et de l'ampleur du risque lié à la carence ou à l'excès d'un élément ou d'un composé.

Les recherches en géographie médicale et en géologie médicale ont montré que tous les milieux géochimiques sont différents, du point de vue de l'état de santé général ou du risque lié à une maladie donnée. C'est ce qui ressort des diverses études portant notamment sur la longévité élevée et faible (Conseil national de recherches du Canada 1981), la carie dentaire (McClure 1970) et les maladies cardio-vasculaires (Conseil national de recherches du Canada 1979; Calabrese et coll. 1980; Shaper et coll. 1980). Par conséquent, il incombe à la

communauté scientifique de caractériser les milieux incapables de fournir à leurs habitants des conditions favorables à un état de santé maximal.

Les exemples d'applications de la cartographie de la «sensibilité» géochimique et du concept dose-réponse présentés plus haut mettent en lumière le rôle de premier plan que peut jouer la géochimie dans la caractérisation des milieux de forte ou de faible incidence de maladie, ce qui rend possible la comparaison des facteurs de risque sanitaire et la formulation de plans d'échantillonnage destinés aux études épidémiologiques.

6. REMERCIEMENTS

Nous aimerions remercier l'unité de cartographie informatisée de la Commission géologique du Canada, à qui nous devons les figures du présent article. Nous remercions également W. Spirito et K. Ford de la Division des ressources minérales de la Commission géologique du Canada qui ont élaboré les cartes géochimiques et radiochimiques. En outre, nous exprimons notre vive gratitude à C.E. Dunn de la Commission géologique du Canada qui nous a fait part de ses observations sur l'ébauche du présent article. Les points de vue exprimés ici sont ceux de l'auteur et ne constituent aucunement les politiques établies par la Commission géologique du Canada.

BIBLIOGRAPHIE

Boyle, D.R. (1991). The Canadian Geochemical Environment and its Relationship to the Development of Health Status Indicators. *Environmental Health Status Indicators*. University Waterloo Press, R.G. McColl (Ed), 1-35.

Calabrese, E.J., Moore, G.S., Tutthill, R.W., et Sieger, T.L. eds. (1980). Drinking water and cardiovascular disease. Pathotox Publishers, Inc., Illinois, 326p.

Chen, X., Chen,X., Yang, G., Wen, Z., Chen, J., et Ge, K. (1980). Relation of selenium deficiency to the occurrences of keshan disease. *Selenium in biology and medicine*. J.E. Spallholz et coll. (eds.), AVI Publ., Westport, Conn., 171-175.

Dyck, W., Garrison, E.W., Godoi, H.O. et Wells, G.S. (1976). Minor and trace element contents of well waters, Carboniferous basin, eastern Canada. Geological Survey of Canada, Open File 00340, NTS 011E, 011L, 021H, 021I et 021P.

Dyck, W. (1980). Regional well water geochemical reconnaissance data, Cypress Hills, Saskatchewan. Geological Survey of Canada Open File Report 678, 61pp.

Grasty, R.L., Carson, J.M., Charbonneau, B.W., et Holman, P.B. (1984). Natural background radiation in Canada. Geological Survey of Canada Bulletin 360, 39p.

Hopps, H.C. et Feder, G.L. (1986). Chemical qualities of water that contribute to human health in a positive way. *Science of the Total Environment*, 54, 207-216.

Lacey, R.F. et Shaper, A.G., (1984). Changes in water hardness and cardiovascular death rates. *International Journal of Epidemiology*, 13, 1, 18-24.

McClure, F.J., (1970). Water fluoridation. The search and victory. U.S. Department of Health Education and Welfare. National Institutes of Health. National Institute of Dental Research, Bethesda, Md., 354p.

National Research Council (1979). Geochemistry of water in relation to cardiovascular disease. National Academy of Sciences Publ., 98p.

National Research Council (1981). Aging and the geochemical environment. National Academy of Sciences Publ., 141p.

Shaper, A.G., Packham, R.F., et Pocock, S.J. (1980). The British regional heart study: Cardiovascular mortality and water quality. Jour. Environ. Pathology and Toxicology, 4-2, 3, 89-111.

LÉGENDES DES FIGURES

1. Liens entre l'homme et le milieu. Toutes les voies mènent éventuellement à l'homme (d'après Boyle 1991).

2. Dureté des eaux souterraines de la région de Moncton, bassin sédimentaire carbonifère des Maritimes (données non publiées).

3. Répartitions du rapport sodium/calcium + magnésium (B), et du baryum (A), dans les eaux souterraines de la région de Moncton, bassin sédimentaire carbonifère des Maritimes (données non publiées).

4. pH des eaux souterraines de la région centrale est du bassin sédimentaire carbonifère des Maritimes (données de Dyck 1976)

5. Exposition au rayonnement gamma (uR/H) en provenance du substrat rocheux et de la couverture en Nouvelle-Écosse (terre ferme). Les régions de forte et de faible expositions correspondent généralement aux régions sises sur des roches granitiques et sédimentaires, respectivement (d'après Boyle 1991).

6. Répartition du fluor dans les eaux de ruisseau de la partie centrale sud de la Colombie-Britannique (les données sur le fluor proviennent de la base de données du Programme d'exploration géochimique préliminaire de la Commission géologique du Canada).

7. Représentation schématique des courbes dose-réponse relatives aux éléments essentiels et non essentiels (d'après Boyle 1991).

8. Carte dose-réponse de la concentration de fluor dans les eaux souterraines de la région centrale est du bassin sédimentaire carbonifère des Maritimes (données sur le fluor, Dyck 1976).

9. Carte dose-réponse du fluor dans les eaux souterraines de la région sud-ouest de la Saskatchewan (données sur le fluor, Dyck 1980).

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

L'INCIDENCE DES DISTORSIONS GÉOGRAPHIQUES ATTRIBUABLES À LA RÈGLE DU SIÈGE D'EXPLOITATION

R. Burroughs[1]

RÉSUMÉ

Le recensement de l'agriculture recueille des renseignements sur l'exploitation d'une ferme en se rapportant seulement à l'emplacement du siège de la ferme. Puisqu'un nombre croissant de fermes sont réparties sur plusieurs lopins de terre distincts, une certaine distorsion est introduite dans les résultats, particulièrement pour les variables relatives à la terre. La recherche présentée dans cette communication tente de mesurer la distorsion à deux endroits différents: un dans l'Île-du-Prince-Édouard et l'autre près de Swift Current en Saskatchewan.

MOTS CLÉS: Règle du siège d'exploitation; distorsion positive; distorsion négative; superficie agricole totale.

1. INTRODUCTION

Le type d'erreur le plus fréquent dans le cas des données infraprovinciales du recensement de l'agriculture découle d'une convention baptisée règle du siège d'exploitation. En effet, la probabilité est grande que toute totalisation sous le niveau de la province comporte une part de ce type d'erreur. Malgré sa fréquence, peu d'utilisateurs en sont conscients ou peuvent déceler ce genre d'erreur sans un examen attentif des données.

L'utilisateur vigilant des données publiées pourrait bien se rendre compte que dans certaines divisions de recensement de la Saskatchewan, la superficie agricole totale est légèrement supérieure à la superficie géographique de l'ensemble de la division de recensement en question. Le fait est porté à l'attention du personnel du recensement de l'agriculture à deux reprises peut-être à l'intérieur du cycle de cinq ans du recensement. Dans cette étude, notre but est d'expliquer comment cette erreur est introduite dans les données, énoncer les facteurs qui peuvent l'amplifier et examiner les conséquences de cette erreur sur les données du recensement de l'agriculture de 1986 relatives à deux régions retenues pour les besoins de l'étude.

2. LA RÈGLE DU SIÈGE D'EXPLOITATION

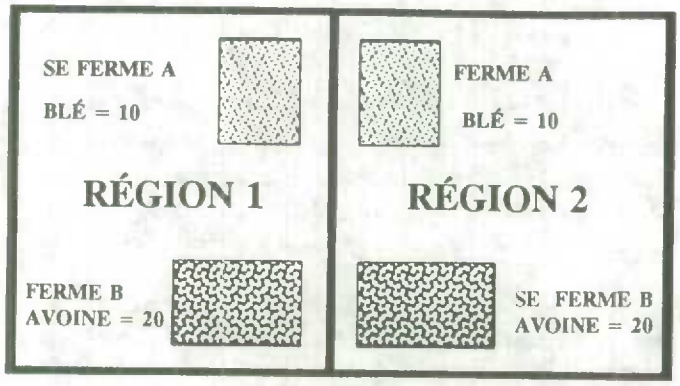
Souvent, les exploitations agricoles du Canada se composent de plusieurs parcelles de terre. Les données recueillies sur le questionnaire se rapportent à l'ensemble de l'exploitation et les variables ne font l'objet d'aucune répartition entre les diverses parcelles de terre. La seule information demandée au sujet de chaque parcelle est leur description officielle et leur superficie. Un problème fondamental se pose lorsqu'un point de référence géographique est associé aux données. Comment répartir les données entre les diverses parcelles de terre? En pratique, on a convenu d'attribuer toutes les données à la parcelle de terre désignée comme le siège de l'exploitation (SE) par l'exploitant, sans tenir compte de l'emplacement de toute autre parcelle de terre rattachée à l'exploitation.

---

[1] R. Burroughs, Statistique Canada, Division de l'agriculture, 12-C2, Immeuble Jean-Talon, Parc Tunney, Ottawa, (Ontario), Canada K1A 0T6.

S'il est vrai que cette façon de procéder permet d'éviter le travail complexe de répartition des données entre les parcelles de terre, elle a pour effet d'introduire des erreurs dans les totalisations faites à partir des données. La figure 1 donne un exemple de ce qui se passe. Prenons le cas d'une économie agricole simplifiée comprenant deux exploitations (A et B), deux régions (1 et 2) et deux types de cultures (blé et avoine). La ferme A, dont le siège est situé dans la région 1, exploite deux parcelles de terre, une dans chaque région et sur chaque parcelle, on a ensemencé de blé une superficie totale de 10 acres. La ferme B a son siège dans la région 2 et exploite deux parcelles de terre de 20 acres chacune, une dans chaque région, les deux parcelles étant ensemencées en totalité d'avoine. Si on détermine les superficies des terres et des cultures à partir de l'observation du diagramme (selon les terres agricoles comprises dans la région, comme le supposent la plupart des utilisateurs), on se rend compte que chacune des régions a une superficie agricole totale de 30 acres, répartie en 10 acres de blé et 20 acres d'avoine. Lorsque les superficies des terres et cultures sont calculées selon la convention de l'emplacement du siège de l'exploitation, on obtient des résultats quelque peu différents. Comme la ferme A est la seule dont le siège d'exploitation est situé dans la région 1, on peut uniquement attribuer à la région 1 les superficies correspondant aux deux parcelles de blé de 10 acres. De la même façon, dans la région 2, seules les parcelles de 20 acres d'avoine de la ferme B peuvent être prises en compte. Les différences entre les valeurs fondées sur les terres elles-mêmes et les valeurs fondées sur l'emplacement du siège d'exploitation correspondent à ce qu'on a appelé les distorsions attribuables à la règle du siège d'exploitation. Soulignons que la somme des distorsions des deux régions est égale à zéro pour chacune des variables (c'est-à-dire la terre, le blé et l'avoine).

Figure 1



	TOTAL	BLÉ	AVOINE
RÉGION 1			
SELON TERRES AGRICOLES	30	10	20
SELON EMPLACEMENT DU SE	20	20	0
ÉCART	-10	+10	-20
RÉGION 2			
SELON TERRES AGRICOLES	30	10	20
SELON EMPLACEMENT DU SE	40	0	40
ÉCART	+10	-10	+20

3. FACTEURS QUI INFLUENT SUR L'IMPORTANCE DES DISTORSIONS

Les facteurs qui déterminent l'ampleur des distorsions sont énoncés dans les paragraphes qui suivent.

Emplacement de la limite

La distorsion illustrée à la figure 1 découle du fait que les parcelles de terre rattachées aux exploitations agricoles en question sont séparées par la limite entre les deux régions. Si les parcelles de terre exploitées par une même ferme ne sont séparées par aucune limite territoriale, il n'y a pas de distorsion. Par exemple, il est peu probable que des distorsions se produisent dans le cas de la frontière entre le Manitoba et l'Ontario, alors que les probabilités sont grandes en ce qui concerne la frontière entre le Manitoba et la Saskatchewan.

Distance entre les parcelles

Plus la distance entre la parcelle de terre désignée comme le siège de l'exploitation et les autres parcelles exploitées est grande, plus une limite risque de les séparer.

Nombre de parcelles

Plus le nombre de parcelles est grand au sein d'une exploitation et plus le risque est grand que l'une ou plusieurs d'entre elles soient séparées par une limite.

Longueur d'une limite

Plus une limite entre deux régions est longue, plus elle risque de séparer des parcelles de terre de la parcelle considérée comme le siège de l'exploitation.

Valeur de la variable associée aux parcelles séparées

Le degré de distorsion liée à une variable donnée augmente en proportion de la valeur de la variable associée à la parcelle séparée par une limite. Dans le cas d'une exploitation qui fait pousser de l'avoine sur une parcelle de terre séparée des autres par une limite, plus la superficie ensemencée d'avoine est grande, plus la variable "avoine" cause de la distorsion. Si la parcelle séparée par une limité n'est pas ensemencée d'avoine, la distorsion liée à cette variable sera égale à zéro.

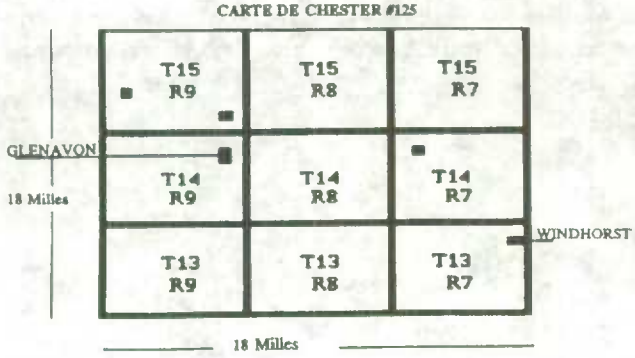
4. ÉTUDE DE CAS I - CHESTER, SASKATCHEWAN

Chester est une municipalité rurale du sud-est de la Saskatchewan. Au recensement de l'agriculture de 1986, le chiffre de superficie agricole publié (selon la règle du siège de l'exploitation) était légèrement supérieur à la superficie géographique de la municipalité, à l'intérieur de ses limites. C'est ce qui explique pourquoi cette municipalité a été choisie comme point de départ de notre étude.

Conformément à sa désignation officielle, Chester est la municipalité rurale n° 125 de la Saskatchewan et comprend neuf townships (townships 13, 14 et 15, rangs 7, 8 et 9 ouest) du deuxième méridien. Ses dimensions physiques sont de 18 milles par 18 milles exactement.

Pour déterminer la distorsion liée à la superficie agricole totale, deux mesures sont utilisées. La première a pour nom distorsion positive. Il s'agit de la superficie des parcelles de terre qui ne sont pas situées à Chester, mais qui sont associées à des exploitations agricoles dont le siège est situé à Chester. Cette mesure a été obtenue à la suite de l'enregistrement de la superficie de toutes les parcelles en question, d'après les renseignements fournis dans les questionnaires des exploitations dont le siège se situe dans la municipalité de Chester. La seconde mesure, baptisée distorsion négative, se rapporte à la superficie des parcelles de terre situées à Chester, mais rattachées à des exploitations dont le siège se situe à l'extérieur de la municipalité. Cette mesure aurait pu elle aussi être fondée sur les renseignements fournis dans les questionnaires, mais il aurait fallu en dépouiller des milliers dans toutes les municipalités avoisinantes. Il était beaucoup plus simple de tirer l'information de données déjà connues, méthode qui est expliquée en annexe.

Figure 2



	SUPERFICIE EN ACRES	EXPLOITATIONS DÉCLARANTES
SUPERFICIE DES TERRES	207 360	-
SUPERFICIE DES EXPLOITATIONS DONT LE SE EST DANS CHESTER	207 958	220
SUPERFICIE ATTRIBUÉE MAIS NON DANS CHESTER	22 326	44
SUPERFICIE NON ATTRIBUÉE MAIS DANS CHESTER	20 608	N/A
SUPERFICIE AGRICOLE SELON CELLE DES TERRES	206 240	-
VARIATION EN POURCENTAGE	I	

La figure 2 présente les résultats. Chester a une superficie de 207 360 acres. La municipalité était le siège de 220 exploitations dont la superficie agricole totalisait 207 958 acres. Ce chiffre est le résultat d'une distorsion

positive de 22 326 acres et d'une distorsion négative de 20 608 acres par rapport à une superficie agricole de 206 240 déterminée selon les terres agricoles comprises dans la municipalité. Les faits importants à souligner sont les suivants:

1) La superficie agricole est supérieure à la superficie de la municipalité parce que la distorsion positive n'est pas complètement neutralisée par la distorsion négative et parce que presque toute la superficie de la municipalité est à vocation agricole.

2) Bien que l'incidence nette de la distorsion sur la superficie agricole soit relativement petite, les distorsions positives et négatives sont en soi significatives (plus de 10% dans chaque cas).

5. ÉTUDE DE CAS II - LOT 19, ÎLE-DU-PRINCE-ÉDOUARD

Le cas du lot 19, à l'Île-du-Prince-Édouard, est à l'opposé de celui de Chester. On n'y pratique pas le même type d'agriculture et l'organisation cadastrale y est très différente. Les exploitations agricoles sont généralement plus petites en termes de superficie. Par conséquent, les distorsions y sont-elles aussi importantes?

La méthode que nous avons utilisée est très semblable à celle décrite pour Chester, sauf que la distorsion négative a été mesurée à l'aide des questionnaires plutôt que des équations. Cette façon de faire représentait beaucoup moins de travail que dans le cas des exploitations de Chester.

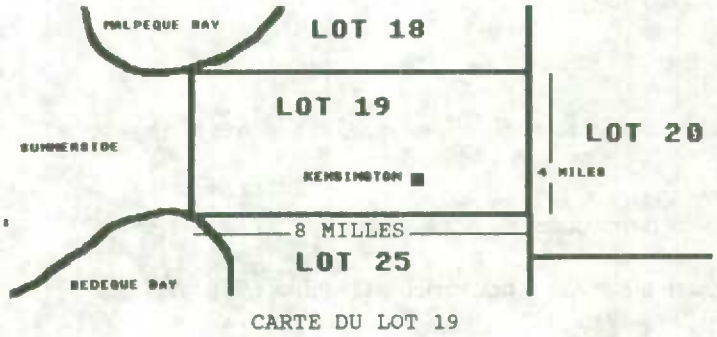
À l'Île-du-Prince-Édouard, les données municipales sont publiées par lot. Ces lots sont beaucoup plus petits que les municipalités rurales en Saskatchewan. Le lot 19 mesure approximativement 8 milles par 4 milles et est agricole dans une proportion d'environ 60%. Il est limité par la ville de Summerside à l'ouest et par trois autres lots ruraux au nord, à l'est et au sud.

Les résultats de l'étude sont présentés à la figure 3. La superficie totale du lot 19 est estimée à 20 352 acres. C'est l'emplacement du siège de 62 exploitations et sa superficie agricole totalise 12 598 acres. Nous avons calculé une distorsion positive de 3 355 acres et une distorsion négative de 3 336 acres.

Comme le nombre d'exploitations auxquelles des distorsions sont associées est relativement petit, nous avons décidé de pousser la recherche. Nous avons constaté que trois exploitations avaient une incidence significative. Celles-ci possédaient une part importante des terres du lot 19 ainsi que des terres des lots avoisinants. Leur siège d'exploitation était situé dans le lot 19. Les données sur la superficie des terres présentées à la figure 4 illustrent ce qui ce serait passé si ces exploitations avaient choisi comme siège une parcelle de terre située dans un autre lot. La superficie agricole, calculée selon la règle du siège d'exploitation, chute de 28% et un déséquilibre important se produit entre les distorsions positives et négatives.

Jusqu'ici l'étude a uniquement porté sur la variable "superficie agricole totale" puisqu'il s'agit de la seule information fournie au sujet de chaque parcelle sur le questionnaire du recensement. Pour l'étude de la distorsion liée aux autres variables, nous avions besoin de données régionales sur les terres provenant d'une autre source. Le système d'imagerie satellite de la Sous-section de la télédétection de la Section des cultures convenait parfaitement et a été mis à notre disposition.

Figure 3



CARTE DU LOT 19

	SUPERFICIE EN ACRES	EXPLOITATIONS DÉCLARANTES
SUPERFICIE DES TERRES	20 352	-
SUPERFICIE DES EXPLOITATIONS DONT LE SE EST DANS LE LOT 19	12 598	62
SUPERFICIE ATTRIBUÉE MAIS NON DANS LE LOT 19	3 355	13
SUPERFICIE NON ATTRIBUÉE MAIS DANS LE LOT 19	3 336	19
SUPERFICIE AGRICOLE SELON CELLE DES TERRES	12 579	-
VARIATION EN POURCENTAGE	0	

Figure 4

DÉPLACEMENT DU SIÈGE DE TROIS EXPLOITATIONS DU LOT 19 À DES LOTS AVOISINANTS

	SUPERFICIE EN ACRES	EXPLOITATIONS DÉCLARANTES
SUPERFICIE DES TERRES	20 352	-
SUPERFICIE DES EXPLOITATIONS DONT LE SE EST DANS LE LOT 19	9 064	59
SUPERFICIE ATTRIBUÉE MAIS NON DANS LE LOT 19	774	10
SUPERFICIE NON ATTRIBUÉE MAIS DANS LE LOT 19	4 289	22
SUPERFICIE AGRICOLE SELON CELLE DES TERRES	12 579	-
VARIATION EN POURCENTAGE	-28	

On y a trouvé une image satellite de l'Île-du-Prince-Édouard prise lors de la saison de végétation de 1986. De plus, ces dernières années, le personnel de la Sous-section a acquis une expérience considérable dans l'estimation de la superficie ensemencée de pommes de terre à partir de l'observation de ces images. Ils ont donc tracé les limites de chacun des champs de pommes de terre du lot 19 et sont parvenus à une estimation d'une superficie de 3 568 acres. À titre de comparaison, les chiffres du recensement de l'agriculture, fondés sur la règle du siège de l'exploitation, étaient de 3 764 acres.

Les faits importants à retenir au sujet de cette étude de cas sont les suivants:

1) Comme dans le cas de Chester, la distorsion nette associée à la superficie agricole était relativement faible. Les distorsions positives et négatives étaient également de proportions importantes, représentant même plus du double de celles calculées pour Chester.

2) L'incidence de trois grandes exploitations agricoles du lot 19 était suffisamment élevée pour modifier sensiblement la valeur de la superficie agricole ainsi que l'équilibre entre les distorsions positives et négatives.

3) L'écart de 196 acres (5%) entre les estimations selon la superficie des terres et les estimations selon l'emplacement du siège d'exploitation qui a été observé dans le cas des champs de pommes de terre démontre

qu'une mesure de la distorsion liée à une variable n'est pas un indicateur fiable de la distorsion liée à toute autre variable.

6. CONCLUSIONS

Les constatations de cette recherche sont les suivantes:

1) Le degré de distorsion au niveau des données régionales peut être significatif, même si l'incidence nette peut sembler négligeable.

2) Les grandes exploitations peuvent causer des distorsions importantes.

3) Il n'existe pas nécessairement de relation entre la distorsion associée à une variable et la distorsion liée à une autre.

4) La distorsion est fonction de plusieurs facteurs qui ne sont pas tous facilement mesurables.

Pour certains utilisateurs, les faibles distorsions nettes soulevées dans cette étude ont peu de conséquences. D'autres peuvent se montrer particulièrement inquiets de l'importance relative des distorsions positives et négatives. La solution que nous préconisons est une évaluation de chaque cas.

Par exemple, il n'y a pas lieu de se préoccuper des données au niveau provincial. Les risques tendent à augmenter à mesure que la désagrégation géographique devient plus fine. Dans le même ordre d'idées, plus les variables font l'objet d'une déclaration uniforme dans une région donnée, plus on peut s'attendre à un équilibre entre les distorsions positives et négatives. En outre, dans les régions où les exploitations sont de taille inférieure à la moyenne, les parcelles de terre multiples tendent à être moins courantes. Dans chaque cas, avant de se servir des données, l'utilisateur devrait donc évaluer les risques et essayer de mesurer l'incidence possible des distorsions.

ANNEXE

Méthode d'estimation de la distorsion négative utilisée pour l'étude de cas I

Dans les régions des provinces des Prairies que l'on retrouve sur un plan de township, la distorsion négative peut être estimée à l'aide des deux équations suivantes:

Équation 1

		STT	=	SNAG + SAG
	où	STT	=	superficie totale des terres
		SNAG	=	superficie non agricole
		SAG	=	superficie agricole

Équation 2

		SAG	=	SAG(SE) - DP + DN
	où	SAG(SE)	=	superficie agricole selon la règle du siège d'exploitation
		DP	=	distorsion positive
		DN	=	distorsion négative

En remplaçant les membres de droite de l'équation 2 par la SAG de l'équation 1, on détermine de la manière suivante la DN:

Équation 3

DN = STT - SNAG - SAG(SE) + DP

Les sources des données des membres de droite de l'équation 3 sont:

STT	-	désignation cadastrale officielle de la municipalité
SNAG	-	plan de township du recenseur
SAG(SE)	-	données publiées sur la superficie agricole de la municipalité
DP	-	compilation des enregistrements sur microfilms des questionnaires

SESSION 9

Cadres géographiques pour les données statistiques

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

NOUVEAUX CADRES CONCEPTUELS POUR LES DONNÉES RURALES

A.M. Fuller, D. Cook et J.G. FitzSimons[1]

RÉSUMÉ

Il est de plus en plus difficile, dans la 'société de centres', d'utiliser des concepts de ruralité clairement définis et communément acceptés. Bien que les configurations et les unités spatiales demeurent relativement fixes, les nouveaux comportements humains sont dynamiques, et on peut mieux les décrire en disant qu'ils constituent des différences de 'style de vie'. Pour redéfinir la ruralité, les auteurs explorent ici trois options en fonction (a) des concepts de région métropolitaine de recensement (RMR) et d'agglomération de recensement (AR), (b) d'une révision des définitions actuelles de la population et de la densité et (c) d'une méthode de classification multidimensionnelle.

MOTS CLÉS: Ruralité; société de centres; classification multidimensionnelle.

1. INTRODUCTION

"Grâce à l'industrie des transports, qui fait qu'il est devenu possible de **se rendre dans un lieu pour y travailler** puis de **revenir dans un autre pour y vivre**, nous pouvons maintenant travailler n'importe où mais nous ne pouvons plus vivre nulle part."

Gilbert 1960

La notion traditionnelle de ruralité a fait son temps. Bien que le monde de la recherche lance depuis longtemps cet avertissement, le terme, les vieilles définitions, la notion elle-même demeurent profondément ancrés dans nos esprits. Comme terme descriptif général, le mot 'ruralité' possède une signification et une valeur certaines. Il évoque dans la mentalité de la plupart des gens des conditions liées à la campagne, aux petites communautés, à la verdure et à l'éloignement. Ces éléments peuvent être perçus comme correspondant à la terre, aux habitants, à l'écologie et à l'espace. Toutefois, la combinaison, la nature ou le degré précis de ces caractéristiques fondamentales varie énormément dans la réalité objective et selon l'expérience de chacun. Fait plus important encore, ces conditions et ces perceptions changent avec le temps, de sorte qu'il devient de plus en plus difficile de maintenir des concepts de ruralité clairement définis et communément acceptés.

Pour les organismes publics qui doivent élaborer des politiques et gérer des programmes, les étiquettes décrivant des différences de situation entre les personnes sont une préoccupation constante. Ces organismes doivent obéir à des impératifs d'équité et de justice, de responsabilité fiscale et d'efficacité politique et répondre à des besoins spéciaux. Les analystes de politiques et les chercheurs qui fournissent aux organismes publics l'information sur les besoins, les coûts et les avantages doivent utiliser des données d'accès public perçues comme étant mises au service du bien commun. Toutes ces exigences et ces contraintes touchant l'élaboration et la mise en oeuvre de politiques et de programmes dans le secteur public font ressortir un énorme besoin de disposer de données pertinentes et exactes qui correspondent à la réalité. Bien que les conceptions populaires de la ruralité doivent conserver leur importance dans la société, elles sont insuffisantes pour la production de données cohérentes et significatives devant servir à l'élaboration de politiques.

---

[1] A.M. Fuller, D. Cook et J.G. FitzSimons, University School of Rural Planning and Development, University of Guelph, Guelph (Ontario), Canada N1G 2W1.

La caractéristique principale de la 'ruralité' est qu'elle est un concept géographique. La notion de ruralité évoque l'idée d'espace, c'est-à-dire de faible densité de la population, des habitations ou des activités. Par son aspect, cet espace rural s'oppose invariablement aux centres urbains ou métropolitains à forte densité. On assiste toutefois à un débat important sur la question de savoir si les habitants de ces espaces dits 'ruraux' sont différents, se comportent différemment ou ont une représentation différente de la réalité. Nous pensons non seulement que les régions rurales présentent un aspect distinct, mais que leurs habitants adoptent des styles de vie différents en fonction de leur environnement et de leurs valeurs. Ce qu'il importe de savoir, toutefois, c'est qu'il existe beaucoup de types d'espaces ruraux et, probablement, de modes de vie ruraux différents, de sorte qu'une comparaison simpliste entre la campagne et la ville est insuffisante et souvent trompeuse.

L'objet de cette communication est de montrer l'impropriété croissante du terme 'ruralité' tel qu'il est utilisé à Statistique Canada et le besoin de définir des concepts plus appropriés qui soient fondés sur la réalité d'aujourd'hui et sur l'avenir probable. Il s'agit donc avant tout d'une contribution théorique qui ne parle qu'accessoirement de manipulation des données et dont le but est de fournir des indices sur les perspectives et les problèmes associés à la modification des définitions, des méthodes de mesure et des unités d'analyse. Nous espérons susciter ainsi des études plus approfondies sur ce problème et apporter quelques idées qui pourraient donner lieu à des recherches et à des débats plus documentés sur la notion de ruralité.

2. LA NATURE CHANGEANTE DE LA RURALITÉ

2.1 La notion traditionnelle de ruralité a-t-elle fait son temps?

On note dans la littérature sur cette question un malaise évident à propos du terme 'ruralité'. Hoggart (1990), se fondant principalement sur l'expérience britannique, fait l'observation générale suivante:

> "...l'utilisation indifférenciée du terme 'ruralité' en recherche nuit à l'avancement des sciences sociales."

Il poursuit en établissant la distinction entre le terme 'rural', qui qualifie un type spécifique de milieu géographique, et le terme 'ruralité', qui désigne un type particulier de comportement lié à ces milieux. Comme ces deux notions sont en outre vagues et englobent des types fort différents et que les chercheurs reconnaissent eux-mêmes qu'ils continuent d'utiliser par commodité les définitions et les données existantes, il recommande de "laisser tomber la notion traditionnelle de ruralité".

Le nombre croissant de séminaires sur l'avenir des régions rurales dans une période de restructuration mondiale montre l'insuffisance de données appropriées qui permettraient d'analyser les tendances nouvelles en matière de transformation et de dynamisme ruraux. Le séminaire 'Agriculture and Beyond, Rural Economic Development', qui s'est tenu aux États-Unis en 1987, a soulevé des inquiétudes partout dans le monde en ce qui a trait à la mesure du dynamisme rural, au-delà de l'agriculture (Castle, Newby, Summers, de Janvrey et Deavers).

Un séminaire international sur des questions de politique rurale, tenu en Écosse en 1991, a donné lieu à des observations semblables quant à la difficulté de mesurer la nouvelle dynamique des régions rurales. Le séminaire organisé par l'Agricultural and Rural Restructuring Group (ARRG) et portant sur les collectivités rurales viables, qui a eu lieu à Saskatoon en 1989, ainsi que la conférence de l'ARRG précédant les réunions de l'Agricultural Economics and Farm Management Society, qui s'est tenue à Vancouver en 1990, ont repris le problème des données, particulièrement dans les communications de Fuller, Ehrensaft et Gertler, d'Ehrensaft et Freshwater ainsi que de Fuller, Bollman et Ahearn. Au symposium international 'Economic Change, Policies, Strategies and Research Issues', organisé par l'Aspen Institute en 1990, Bonnen (É.-U.), MacDowell (É.-U.), Capellin (Italie) et Fuller (Canada) ont parlé des difficultés qui entourent le problème des définitions. Cette série de conférences a abouti en 1990 à la Conférence de Statistique Canada sur le milieu rural et les petites villes du Canada; les communications qui y ont été faites ont montré à la fois les contraintes du point de vue des données et l'insuffisance de concepts efficaces pour définir la ruralité. En 1991, deux autres conférences ont confirmé cette opinion: celle de Galway, à laquelle ont participé des spécialistes européens du développement rural (par exemple Grohn, de Finlande, et Henrichsmeyer, d'Allemagne) et celle du Royaume-Uni, tenue en août, où se

sont rencontrés des géographes ruraux du Royaume-Uni, du Canada et des États-Unis préoccupés par la restructuration rurale (par exemple Munton, Hart et Bryant).

2.2 La ruralité n'est plus ce qu'elle était

Il ressort de ce débat que la notion de ruralité a changé de sens. Des définitions qui ont sans doute été utiles dans le passé ont perdu leur sens avec les années et, parce qu'elles ont été maintenues pour des raisons de commodité et de continuité, ont même gêné la recherche. On peut se faire une représentation simple, en trois étapes, de l'évolution de la ruralité pour illustrer la dynamique de la transformation des sociétés industrielles occidentales depuis cent cinquante ans.

La société de courte distance est fondée sur la primauté des économies primaires. Elle correspond à un ancien état de choses où était centralisée dans le village la plus grande part de l'activité du pays environnant. Il n'y avait qu'une distance relativement courte à parcourir en voiture hippomobile pour atteindre le centre où se procurer des biens et des services et avoir accès à certains services communautaires comme l'église et l'école. Sa dynamique principale est centripète, l'économie est tributaire des ressources et l'organisation sociale est relativement structurée et fermée. Dans la 'société de courte distance', l'unité de l'espace et des fonctions est très marquée.

La société industrielle représente l'élargissement de l'espace interactif, mais sa caractéristique principale demeure la collectivité centrale, dont le mode d'organisation et la fonction prennent un caractère industriel, même lorsqu'elle répond aux besoins agricoles et qu'elle transforme les produits agricoles. L'organisation sociale demeure communautaire malgré le caractère 'contractuel' plus accentué des relations économiques et sociales. La technologie est le moteur principal du changement, et un exode net de main-d'oeuvre caractérise cette étape dans la plupart des systèmes ruraux. La collectivité à ressource unique (ville monoindustrielle), qui appartient à la fois à la société de courte distance et à la société industrielle, est un exemple de ce stade d'évolution.

La société de centres est le reflet de la nouvelle réalité rurale. L'espace dans lequel s'exerce l'activité s'élargit considérablement et englobe plusieurs centres de commerce où se consomment des biens et des services personnels et ménagers et où s'opère la socialisation. L'économie est désormais liée à la production internationale et aux marchés financiers, la segmentation du marché du travail étant très prononcée. La technologie qui a pour effet d'abolir les distances a résolu le problème de l'isolement par la diffusion des informations et par les médias, mais les distances géographiques et l'effet de l'espace demeurent visibles et se traduisent par des coûts de transport élevés. La grande mobilité personnelle que procurent les véhicules automobiles est une caractéristique du réseau multicommunautaire local.

Il est primordial de reconnaître que ce concept des trois étapes, élaboré à l'origine par Persson (1991), est une simplification de l'évolution de la ruralité au cours des années. Fait important à souligner, les trois types de société 'rurale' peuvent coexister dans une même région ou dans un même pays. Ce concept tient compte des racines de notre ruralité actuelle, dont il subsiste encore des éléments, tout en permettant la détermination de certaines caractéristiques essentielles de la nouvelle réalité. La notion d'espace et la configuration physique du paysage de même que les infrastructures comme les routes et la forme de peuplement demeurent essentiellement les mêmes, mais la réalité économique et sociale de l'activité humaine s'est modifiée considérablement. **Cela laisse supposer qu'il s'est produit un changement fondamental d'échelle de grandeur dans la relation entre les unités spatiales et les fonctions socio-économiques.**

2.3 Si la ruralité n'est plus ce qu'elle était, alors quel sens a-t-elle?

La nouvelle réalité rurale est un mélange complexe d'interrelations qui s'établissent dans et entre les infrastructures créées pour les sociétés de courte distance au dix-neuvième siècle. Le besoin individuel de s'identifier à un lieu (foyer et collectivité), d'interagir avec plusieurs lieux (le réseau multicommunautaire), d'atteindre la qualité de vie escomptée de nos jours et d'être informé des opinions et des nouvelles internationales de la société de centres constitue la nouvelle norme. C'est dans ce contexte que les nouveaux marchés ruraux du travail apparaissent, que des réseaux de transport perfectionnés (vulnérables aux crises du pétrole) prennent une grande importance et que des économies se rattachent à des marchés internationaux ou s'en détachent. Bien que les configurations et les unités spatiales demeurent relativement fixes, les nouveaux

comportements humains sont dynamiques, et on peut le mieux les décrire en disant qu'ils constituent des différences de 'style de vie'.

3. REDÉFINITION DE LA RURALITÉ

Les résultats sont décevants lorsqu'on cherche dans la littérature sur la ruralité des idées valables qui pourraient aider à la reformulation du concept de ruralité. Seuls quatre auteurs à l'extérieur de Statistique Canada semblent avoir une opinion sur la question de la ruralité en ce qui a trait aux méthodes de mesure.

Paul Cloke a élaboré un indice de la ruralité pour l'Angleterre et le pays de Galles au milieu des années 1970 (Cloke 1977). Cet indice se fondait sur une analyse multidimensionnelle des facteurs qui permettait d'obtenir une échelle de valeurs au moyen de laquelle on pouvait représenter la ruralité sur une carte. Mis à jour en 1986, cet indice a permis d'observer que la ruralité et la pauvreté avaient tendance à augmenter dans les régions éloignées et les hautes terres comme le centre du pays de Galles (Cloke et Edwards 1986).

Marvel Lang a étudié diverses approches pour la redéfinition des notions de milieu rural et de milieu urbain en vue du recensement de la population des États-Unis de 1986. Il fait remarquer que les formes de peuplement et les modes de vie socio-culturels sont les deux transformations dignes de mention qui se sont produites dans la population américaine et il observe que "... le fait pour une personne de demeurer en milieu rural ne signifie plus nécessairement qu'elle ait adopté un mode de vie rural". Il préconise l'utilisation d'une méthode de regroupement des ménages, mais reconnaît que cette méthode ne tient pas compte des aspects socio-culturels de la population; ce qui fait qu'une fois de plus une idée est proposée, mais sans être assortie de moyens satisfaisants pour effectuer des mesures.

Beal, lui, est très précis au sujet des méthodes de mesure; il divise l'espace géographique des États-Unis en unités fondées sur la densité de population (Beal 1978). Il soutient qu'il y a un continuum du milieu rural à la grande ville et qu'une division en dix catégories permet de faire ressortir les différences essentielles entre les régions rurales et urbaines. Reposant sur une solide logique interne fondée sur l'idée de continuum, cette méthode a l'avantage d'être relativement simple. Toutefois, en appliquant au Canada les catégories de Beal, Ehrensaft a découvert que les seuils de population utilisés par Beal ne convenaient pas au contexte canadien, bien que la notion de distance par rapport aux grands centres soit fort utile (Ehrensaft et Beaman 1991). Dans sa tentative d'utilisation des dix catégories de Beal, Ehrensaft en a ajouté une onzième pour tenir compte des milieux nordiques et autochtones.

3.1 Les exigences de la mesure dans le domaine public

Deux exigences dominent dans la diversité des considérations nécessaires à la reformulation des concepts statistiques et géographiques. La première est la nécessité de rendre les nouvelles définitions compatibles avec celles des anciennes qui sont le plus valables, de façon à assurer la continuité dans le temps. L'analyse de l'évolution de la société est une des principales raisons pour lesquelles on recueille des données statistiques. Cette analyse serait impossible si, à chaque recensement, l'on adoptait de nouvelles définitions qui vaudraient seulement pour la période de ce recensement. On pourrait surmonter la difficulté en modifiant les concepts et les définitions à mesure qu'apparaissent de nouveaux phénomènes sociaux importants, c'est-à-dire en les adaptant aux modèles déjà établis et utilisés.

La seconde exigence réside dans le fait que tous les concepts et définitions doivent être simples et d'application universelle, la complexité étant à la fois coûteuse et susceptible de semer la confusion chez leurs utilisateurs. Ces concepts et ces définitions doivent aussi avoir une signification large et être universellement acceptés. À l'heure de la mondialisation des économies, des nouvelles alliances politiques et économiques et de la reconnaissance de la diversité culturelle, la nécessité de disposer de concepts et de définitions qui soient acceptés à l'échelle internationale devient de plus en plus importante.

Ces exigences de compatibilité dans le temps, de simplicité et d'universalité sont des facteurs contraignants dans la quête de méthodes de mesure améliorées de la ruralité.

3.2 Un nouveau concept de ruralité

Il y a trois facteurs qui entrent en ligne de compte dans la redéfinition de la ruralité: l'évolution du concept de ruralité jusqu'à la notion de société de centres; les idées et l'information provenant d'autres études, dont l'expérience de ceux qui essaient de mesurer la ruralité et de la représenter dans l'espace; les contraintes imposées par les organismes statistiques du point de vue de la continuité et de l'universalité. On peut formuler trois options correspondantes: **tirer le meilleur parti** de l'état de choses actuel; **adapter** le concept existant; ou **modifier** complètement le concept et la définition.

A. Le nouveau statu quo

Il importe de noter qu'on a déjà fait une tentative courageuse en vue d'améliorer la situation au Canada et que le modèle de la région métropolitaine de recensement (RMR) et celui de l'agglomération de recensement (AR) en sont le résultat. Adoptés pour le recensement de 1986, ces deux concepts géographiques, fondés sur la population, comportent chacun trois niveaux: le noyau urbanisé, la banlieue urbaine et la banlieue rurale. La RMR a dans son ensemble plus de 100 000 habitants et l'AR, plus de 10 000. Les deux concepts géographiques permettent de distinguer des sous-unités rurales dans les zones qu'ils définissent, de telle sorte que la ruralité peut être fonction d'une différence d'ordre de grandeur du chiffre de population.

L'hypothèse générale est que la définition de la RMR et celle de l'AR reposent sur une classification de l'espace en fonction des interactions parmi la population, c'est-à-dire sur une classification qui sert à décrire le marché du travail. Cette classification est conçue pour tenir compte de l'influence prépondérante des grands centres métropolitains sur la structure sociale et économique du pays environnant. Elle se fonde donc sur la notion traditionnelle de relation métropole/pays environnant, relation qui est illustrée par les caractéristiques du marché du travail.

Bien que cette hypothèse soit en elle-même valable, elle est aussi le point faible du système du fait qu'elle suppose que toutes les interactions humaines sont régies par les relations avec la région métropolitaine. Fait plus important, elle suppose la prépondérance d'**une seule** agglomération urbaine; or ce n'est pas là ce que nous observons dans la nouvelle société de centres, où il y a souvent plusieurs centres et modes d'interaction, en particulier dans les régions rurales.

Par essence, l'approche qui place l'agglomération urbaine au centre de tout, l'unipolarité, le caractère arbitraire des seuils de population et le recours exclusif à la notion de marché du travail sont autant de points faibles qui font que la classification RMR/AR n'est pas satisfaisante dans la perspective d'une étude de la ruralité.

B. L'adaptation des définitions

L'option B consiste à améliorer la situation en procédant à une révision des définitions existantes sans abandonner le principe de la continuité dans le temps. Nous avons donc décidé de conserver les unités géographiques (subdivisions de recensement) pour assurer la continuité spatiale, mais d'augmenter le nombre de catégories de ruralité, en recherchant les seuils de population qui révéleraient des différences significatives pour certaines variables clés sélectionnées en fonction de ce que nous savons des conditions actuelles de la vie rurale dans le sud de l'Ontario.

Nous avons donc testé cette approche en utilisant les profils 2A/2B du recensement de 1986 pour l'Ontario, au niveau de la subdivision de recensement. Les données ont été converties en pourcentages, puis en quintiles, en vue d'une comparaison des centres selon leur taille. Les variables indépendantes que nous avons choisies sont la population, la densité de population et la distance par rapport à des centres urbains de taille déterminée. Nous avons sélectionné comme variables dépendantes les paramètres suivants, jugés être les indicateurs essentiels de l'économie sociale de n'importe quelle unité géographique:

type de logement,
migration,
emploi,
éducation,

infrastructure,
revenu.

Un test de normalité de Kolmogorov-Smirnof a été effectué pour évaluer les différences entre les catégories de population selon les indicateurs choisis. Comme le test était effectué sur des données converties en quintiles, les résultats ont révélé des niveaux d'écart dans la structure des variables selon les unités, et non pas des différences de valeur entre les variables elles-mêmes.

Il est intéressant d'observer que lorsqu'on a intégré la densité dans l'analyse à l'aide de la définition officielle de la densité de population urbaine -- <1 000 et <400 --, 52% des cas n'étaient pas classés selon les paramètres donnés. Une nouvelle analyse effectuée sur le groupe statistique des 52% a établi la validité de la limite de 1 000 habitants comme seuil significatif, mais non celle de la limite de 400 habitants. Cela confirme qu'une faible population ne correspond pas nécessairement à une faible densité, et vice versa; de plus, un petit chiffre de population indique la taille et la situation des unités géographiques plutôt que leur ruralité.

La figure 1 montre les résultats des tests de différence entre les indicateurs sélectionnés pour six groupes de taille de population (1-999, 1 000-2 499, 2 500-4 999, 5 000-9 999, 10 000-19 999, 20 000 et plus). Tous les indicateurs révèlent un écart statistique entre les régions de moins de 1 000 habitants et celles de 1 000 à 2 499 habitants. On a observé des différences, quoique non systématiques, parmi les unités de 2 500 à 10 000 habitants. Deux indicateurs différaient pour les unités de 10 000 à 20 000 habitants, et les communautés de plus de 20 000 habitants différaient de nouveau entre elles. La tendance résultante, qu'on peut voir à la figure 1, est simple et révélatrice. Nous avons donné des noms aux groupes de taille de population étudiés pour en faire ressortir les différences caractéristiques quand on les examine dans leur réalité.

Figure 1

Régions rurales au Canada

0 - 1 000	1 000 - 10 000	10 000 - 20 000
Région rurale	Centre de district	Centre régional
Adjacente Éloignée	Adjacent Éloigné	Adjacent Éloigné

Une **région rurale** a moins de 1 000 habitants et peut être proche ou éloignée d'une grande agglomération. Un **centre de district** a une population de 1 000 à 10 000 habitants et peut être situé loin ou à proximité d'un grand centre urbain. Un **centre régional** a une population de 10 000 à 20 000 habitants.

Nous pouvons ainsi considérer que les trois groupes de subdivisions de recensement décrits plus haut forment l'ensemble des régions rurales du Canada rural étant donné que nous avons tenu compte de la plupart des variations spatiales et démographiques des régions rurales. Nous savons également que les divers groupes de taille de population diffèrent quant aux indicateurs de revenu, de migration et d'emploi, qui sont d'assez bons substituts des caractéristiques du marché du travail. Si nous ajoutons la variable de proximité d'un centre urbain, autre facteur caractéristique du marché du travail, nous pouvons raisonnablement admettre que la plupart des situations rurales sont couvertes dans cette classification à trois catégories.

C. L'approche multidimensionnelle

L'utilisation d'un processus à deux étapes faisant intervenir une classification multidimensionnelle des données de recensement du plus bas niveau comme fondement des indicateurs socio-économiques choisis, puis la classification de divisions de recensement plus importantes suivant la proportion et la combinaison de ces groupes socio-économiques se trouvant dans leurs limites, c'est là une approche qui mérite un examen plus approfondi. Nous pourrions l'utiliser en y associant les indicateurs plus généraux tels que la taille et la densité de la population et la proximité d'une métropole.

Fait également important, l'approche multidimensionnelle nous permet de décrire la ruralité au moyen de changements de style de vie. Des indicateurs judicieusement choisis peuvent faire ressortir efficacement les caractères essentiels de la société de centres tout en permettant de conserver leur pertinence aux caractéristiques résiduelles de la société de courte distance et de la société industrielle et de les faire entrer dans les calculs.

On peut avoir une idée de ce que serait la première étape de ce processus en examinant la classification Lifestyles (marque déposée) selon la segmentation des marchés mise au point par Compusearch au moyen des données du recensement de 1981 et remaniée en profondeur à l'aide des données de celui de 1986.

La classification Lifestyles de 1986 se fonde sur une analyse de grappes non hiérarchique de trente-cinq variables qui représentent le revenu, le niveau d'instruction, l'âge du chef de ménage, la taille du ménage, l'emploi et la profession, la mobilité du ménage, le type d'habitation et le mode d'occupation, le cadre résidentiel et la langue maternelle (Compusearch 1989). Des études distinctes ont été effectuées pour les communautés de plus de 25 000 habitants ('urbaines') et pour les autres petites villes, régions rurales et agglomérations ('rurales'). Ces analyses ont produit 48 grappes 'urbaines' et 22 grappes 'rurales' représentant des 'quartiers' (secteurs de dénombrement) comparativement homogènes à l'aide des variables incluses dans les études. Les 70 grappes ont ensuite été regroupées dans 13 grandes catégories.

Même si ce système de classification fait ressortir la richesse de la base de données, encore que seulement dans une perspective de segmentation de marchés, on pourrait, à partir d'une liste de variables plus restreinte, élaborer un système général plus simple pour la classification des SD, sans distinction a priori de l'"urbain' et du 'rural' en fonction d'un seuil de 25 000 habitants. On pourrait ensuite classer les plus grandes unités géographiques de recensement en fonction à la fois de la diversité et de la proportion relative des types de grappe qu'elles contiennent.

Une méthode de classification multidimensionnelle permet de contourner certaines des limitations du regroupement, de la classification et de la causalité qui sont inhérentes au système actuel. On résout le problème du regroupement en permettant l'organisation de l'espace recensé selon l'activité sociale et économique plutôt qu'en fonction de certaines unités prédéfinies. L'approche multidimensionnelle permet de classer les unités selon des types de comportement contemporains et donc d'explorer des questions nouvelles et peut-être plus pertinentes. Enfin, en prenant la population comme une simple variable plutôt que comme variable indépendante, cette approche permet à la recherche de dépasser une forme simpliste de relation de causalité en matière de comportement des populations.

Sur le plan théorique, une classification multidimensionnelle est plus conforme à la nouvelle réalité spatiale de la société de centres, réalité qui est plus mobile et dans laquelle l'activité sociale et économique est plus dispersée. La société de centres représente la tendance générale vers la mondialisation, où le changement technologique dans les communications a dispersé les valeurs et l'information et où l'amélioration des transports a réduit les distances. La taille de la population est de ce fait un facteur moins important dans les différences sociales et économiques. La méthode de classification multidimensionnelle nous permet donc de dépasser la relation implicite entre la taille de la population et son comportement.

Les principales limitations de la classification multidimensionnelle ont trait à son manque de continuité dans le temps et à sa subjectivité. La mise en oeuvre d'une approche multidimensionnelle dans de futurs recensements rendrait difficile la comparaison avec les données de recensements antérieurs aussi bien que postérieurs puisqu'il faudrait remanier la classification après chaque recensement pour qu'elle demeure applicable à la période visée par un recensement. La classification multidimensionnelle est ainsi plus subjectives que ne le sont des définitions plus simples, susceptibles de ne pas changer avec le temps.

4. SOMMAIRE ET CONCLUSIONS

Cette étude préliminaire montre clairement qu'en nous proposant de modifier le concept et la définition de la ruralité de façon qu'elle corresponde plus fidèlement à la réalité actuelle nous nous sommes donné une tâche difficile. L'exigence de la simplicité et de la continuité dans le temps a produit dans tous les pays industrialisés

de l'Occident des définitions de l'espace rural qui sont fondées sur la population et sur des mesures connexes comme la densité de population.

Bien qu'un système de classification fondé uniquement sur la population, la densité de population et la proximité de grands centres métropolitains présente l'avantage apparent de la simplicité, il soulève un certain nombre de problèmes, que l'on peut ranger dans trois catégories: le regroupement, la classification et la causalité. Le petit nombre des indicateurs et la généralité du niveau de représentation géographique ne permettent pas de faire ressortir la diversité socio-économique considérable des grandes unités géographiques recensées. Les unités spatiales actuellement utilisées, qui correspondent à des divisions administratives, reflètent également la société de courte distance et la société industrielle, et leur utilisation ne permet peut-être pas de faire ressortir suffisamment les nouveaux modes d'interaction. En outre, l'actuelle définition des milieux urbain et rural suppose l'existence d'une relation entre la taille et la densité d'une population, d'une part, et ses interactions et ses activités socio-économiques, d'autre part. Comme la nature et la forme de cette relation changent visiblement, de tels indicateurs secondaires simplistes peuvent n'être plus des descripteurs appropriés.

Nous pouvons dégager de notre exposé trois observations qui seraient les jalons de recherches et de discussions plus approfondies.

1. La réalité rurale au Canada est devenue un ensemble complexe de caractéristiques qui diffèrent autant l'une de l'autre qu'elles permettent de distinguer le rural de l'urbain. Un débat théorique en profondeur doit être entrepris pour obtenir un consensus sur le nombre de grands types d'espaces ruraux à définir et sur les principaux indicateurs qui doivent servir à les distinguer les uns des autres.

2. Il faut entreprendre une étude particulière de l'approche multidimensionnelle relative à la classification de la ruralité et de l'espace rural. C'est cette approche qui est la plus prometteuse parce que sa conception incorpore un élément dynamique et qu'elle permet des redéfinitions successives.

3. Il faut procéder, pour l'ensemble du Canada, à un examen approfondi de la méthode de révision des définitions en utilisant des seuils de population inférieurs à la limite de 20 000 à 25 000 habitants.

Si de telles études sont effectuées de façon sérieuse et structurée, nous pourrions avoir à temps pour le prochain recensement, au début du XXI[e] siècle, la possibilité de comprendre pleinement les options viables qui s'offrent à nous pour la redéfinition de la ruralité.

BIBLIOGRAPHIE

Beale, C. (1977). Quanti-dimensions of decline and stability among rural communities. Dans Richard Rodefeld (éd.) *Change in Rural America: Causes, Consequences and Alternatives*, Saint Louis: C.V. Mosby Co., 70-78.

Bryant, C. (1991). Community development and the restructuring of rural employment. Communication présentée à la conférence sur la Contemporary Social and Economic Restructuring of Rural Areas, Londres (R.-U.).

Cloke, P.J. (1977). An index of rurality for England and Wales, *Regional Studies*, 11, 31-46.

Cloke, P.J. (1986). Rurality in England and Wales, *Regional Studies*, 20, 289-306.

Compusearch (1989). *Lifestyles (TM) Reference Manual*, Toronto: Compusearch.

Ehrensaft, P., et Beeman, J. (éds.) (1991). Distance and diversity in non-metropolitan economies. Université du Québec à Montréal, document inédit.

Gertler, M., et Baker, H.R. (éds.) (1989). *Collectivités rurales viables au Canada*. Actes du séminaire n° 1 sur la Rural Policy, Saskatoon, Saskatchewan, 11 au 13 octobre.

Gilbert, E.W. (1960). The Idea of the Region, *Geography*, XLV, 157-175.

Grohn, K. (1991). Politique rurale en Finlande. Écrit à l'intention du secrétariat du programme rural de l'OCDE, ministère de l'intérieur - Développement municipal et régional, Finlande.

Hart, J.F. (août 1991). Part-ownership and farm enlargement. Communication présentée à la conférence sur la Contemporary Social and Economic Restructuring of Rural Areas, Londres (R.-U.).

Henrichsmeyer, W. Développement rural viable: objectifs et contraintes. Institut de la politique agricole, Université de Bonn, document inédit.

Hoggart, K. (1990). Let's do away with rural, *Journal of Rural Studies*, 6, 3, 245-257.

Lang, M. (1968). Redefining urban and rural for the U.S. Census of population: Assessing the need for alternative approaches, *Urban Geography*, 7, 2, 118-134.

Munton, R. (août 1991). Farm adjustment in a period of uncertainty: Some aspects of the British experience. Communication présentée à la conférence sur la Contemporary Social and Economic Restructuring of Rural Areas, Londres (R.-U.)

Persson, L.O., et Westholm, E. (1991). Changing macro conditions reflected by rural households. Communication présentée dans le cadre du Changement rural en Europe, réunion de la 5e revue à Sila Greca, Calabre (Italie).

Summers, G.F., Bryden, J., Deavers, K., Newby, H., et Sechler, S. (éds.) (1988). *Agriculture and Beyond, Rural Economic Development*, Madison, Wisconsin: University of Wisconsin.

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

LA DICHOTOMIE ENTRE LES POPULATIONS URBAINE ET RURALE: APERÇU DES CRITÈRES ACTUELS ET PERSPECTIVES DE RECHERCHE

N. Torrieri et J. Sobel[1]

RÉSUMÉ

La division de la géographie (GEO) du U.S. Bureau of the Census réexamine présentement, en vue du recensement de l'an 2000, les critères et les méthodes qu'elle utilise pour définir et délimiter les agglomérations urbaines et les régions urbanisées (RU). La GEO envisage diverses solutions -- y compris le fait de ne plus s'en remettre aux limites des agglomérations -- pour délimiter les RU. De plus, elle entend examiner diverses techniques de délimitation, comme la méthode des moyennes mobiles. Elle étudiera aussi la possibilité de créer de nouvelles régions géographiques afin de classifier de façon plus précise les grappes de population urbaine aux États-Unis. Ces activités visent à accroître l'aptitude du Census Bureau à distinguer les populations urbaines des populations rurales.

MOTS CLÉS: Population urbaine; population rurale; région urbanisée; agglomération urbaine; densité de la population.

1. INTRODUCTION

Le Census Bureau définit la population urbaine des États-Unis avant chaque recensement décennal. La population urbaine se compose de toutes les personnes qui vivent dans les agglomérations[2] d'au moins 2 500 habitants et de celles qui vivent dans des unités statistiques géographiques appelées "régions urbanisées" ou RU. Une RU comprend une ou plusieurs agglomérations ainsi que le territoire avoisinant à forte densité de population où l'on compte en tout 50 000 personnes ou plus. Pour le Census Bureau, la population rurale se compose de toutes les personnes qui ne font pas partie de la population urbaine.

Les définitions de population urbaine et de population rurale du Census Bureau peuvent servir de base à des études savantes sur le peuplement, la répartition géographique de la population et les mouvements de population. Elles servent aussi à la mise en oeuvre de divers programmes. De nombreuses organisations du secteur privé fondent leurs décisions en matière de marketing et de localisation sur les caractéristiques économiques et sociales de certaines RU ou sur le fait qu'une RU se trouve à proximité. En outre, la RU sert d'unité géographique de base pour l'application de dizaines de programmes gouvernementaux, depuis l'établissement de normes d'émission de gaz d'échappement jusqu'au calcul des sommes versées aux hôpitaux par l'administration fédérale à titre de remboursement.

---

[1] N. Torrieri et J. Sobel, Division de géographie, U.S. Bureau of the Census, Washington (DC) 20223-0001, É.-U.

[2] Dans les critères servant à définir une RU, le terme "agglomération" désigne aussi bien des municipalités, comme des villes et des villages, que des agglomérations de recensement (census designated places -- CDP). Une agglomération de recensement est une grappe de population non constituée en municipalité pour laquelle le Census Bureau définit des limites en collaboration avec des organismes de l'État et les administrations locales. En ce qui concerne Porto Rico, les autorités locales s'entendent avec le Census Bureau pour reconnaître les "zonas urbanas" et les "communidades" comme les équivalents d'une agglomération; ces zones sont utilisées dans l'application des critères servant à définir une RU.

Des millions de dollars sont alloués aux régions qui sont reconnues comme RU au titre de la voirie et du transport en commun. Les fonctionnaires des administrations locales savent très bien que la dénomination "RU" peut faire la différence entre un excédent et un déficit budgétaires, entre la prestation de services essentiels et leur suppression. Ils tentent parfois d'influencer le processus de désignation en communiquant directement avec des membres du Census Bureau ou en faisant des représentations auprès des membres du Congrès. Compte tenu des enjeux, le Census Bureau prend toutes les précautions nécessaires pour que les critères de délimitation des RU soient appliqués de façon équitable et cohérente à la grandeur du territoire.

Après chaque recensement décennal, la division de la géographie revoit les critères et les méthodes qu'elle utilise pour définir et circonscrire la population urbaine du pays. Ce processus de révision est déjà engagé en vue du recensement de l'an 2000.

2. UN SURVOL HISTORIQUE DES DÉFINITIONS URBAINES ET RURALES AU CENSUS BUREAU

Les premières définitions de la population urbaine et de la population rurale au Census Bureau remontent à 1790, année du premier recensement. À cette époque-là, les personnes qui vivaient dans les municipalités de 2 500 habitants ou plus faisaient l'objet d'une totalisation indépendante mais n'étaient pas encore désignées collectivement comme population "urbaine". La première publication du Census Bureau où l'on désigna les localités de 2 500 habitants ou plus comme des agglomérations "urbaines" fut un Supplément au recensement de 1900, daté de 1906[3]. On ne donnait aucune raison particulière pour le choix de ce seuil sinon d'affirmer qu'il constituait une ligne de démarcation plus réaliste entre les populations urbaine et rurale.

Le recensement des manufactures de 1900 a été l'occasion de définir des districts industriels -- ce qu'on appelle aujourd'hui les régions métropolitaines -- autour des quatre plus grandes villes des États-Unis: New York, Chicago, Philadelphie et St-Louis. Ces districts comprenaient la population des banlieues et étaient constitués des divisions civiles secondaires (DCS) à forte densité de population qui étaient situées à moins de 10 milles d'une ville centrale. Or, les DCS contenaient de grandes portions de territoire rural et d'importants îlots de population rurale, et les limites de certaines de ces divisions avaient été modifiées sensiblement au fil des années. Pour le recensement de 1910, les districts industriels devinrent des "districts métropolitains"; par la suite, on définit des districts métropolitains pour les recensements de 1920, 1930 et 1940.

Les formes de peuplement observées aux États-Unis au début du dix-neuvième siècle, celles-là même qui ont été à l'origine de la définition de population urbaine dans les recensements, étaient caractérisées généralement par des centres de population uniques et reconnaissables. Vers la fin du dix-neuvième siècle, le développement des moyens de transport et le coût raisonnable des terres et des habitations rurales ont amené les gens à s'établir à l'extérieur des limites des villes et des villages. Au même moment, on a commencé à observer la présence d'importantes collectivités et d'établissements commerciaux à l'extérieur des villes; bon nombre de ces villes étaient elles-mêmes en mutation grâce aux projets d'annexion et de remembrement qui leur permettaient d'acquérir de grandes portions de territoire à faible densité de population. Ainsi, la ville en tant qu'entité politique ne suffisait plus pour définir une région urbaine.

Pour tenir compte de cette évolution, le Census Bureau s'est efforcé de revoir sa définition de la population urbaine en vue des recensements de 1930 et de 1940 en révisant à plusieurs reprises les définitions de population urbaine et de population rurale. La révision la plus notable a été celle par laquelle on a fixé à 1,000 personnes au mille carré (p.m.c.) la densité de population minimum pour qu'une subdivision administrative puisse être reconnue comme urbaine. En 1948, le Census Bureau a organisé une conférence sur la banlieue urbaine afin d'envisager des modifications plus profondes aux définitions de population urbaine et de population rurale. Cette conférence a permis l'adoption de deux nouvelles unités statistiques. Premièrement, le Bureau of the Budget (aujourd'hui l'Office of Management and Budget), en collaboration avec d'autres organismes fédéraux dont le Census Bureau, a créé la "standard metropolitan area" (SMA) pour définir la zone métropolitaine rattachée aux grandes villes. Deuxièmement, le Census Bureau a élaboré la définition de RU pour désigner la zone urbaine située aux alentours des grandes villes ainsi que la population de cette zone.

---

[3] Twelfth Census of the United States, 1900, Supplementary Analysis, p. 20.

La SMA était un moyen de définir une zone pratique d'intégration économique et sociale autour d'une ou de plusieurs agglomérations centrales. La RU, au contraire, servait à mesurer l'étendue d'une agglomération urbaine, y compris la portion bâtie d'un noyau urbain et les quartiers périphériques à forte densité de population. Le Census Bureau a appliqué pour la première fois sa nouvelle définition de la population urbaine lors du recensement de 1950. Ce fut le premier recensement où le Census Bureau classait comme population urbaine les groupes suivants: les personnes vivant dans les agglomérations urbaines (agglomérations d'au moins 2,500 habitants, constituées ou non en municipalité, qui se distinguent de l'habitat suburbain dispersé et qui ne font pas partie d'une RU) et celles vivant dans les RU.

3. CRITÈRES SERVANT À DÉFINIR UNE RU ET MÉTHODE DE DÉLIMITATION DES RU

Nous allons exposer dans leurs grandes lignes les critères qui servaient à définir les RU en 1990 ainsi que la méthode de délimitation des RU. La figure 1 illustre l'application de ces critères pour une agglomération hypothétique et les zones périphériques à forte densité de population.

Minimum de 50 000 habitants

Une RU se compose d'une agglomération et des zones avoisinantes fortement peuplées qui, ensemble, comptent au moins 50 000 habitants.

Densité de population minimum

L'agglomération qui forme le noyau de la RU n'a pas de critère minimum à respecter en ce qui a trait à la densité de population, bien qu'en règle générale les noyaux de RU aient une densité de population d'au moins 1 000 p.m.c. Les zones limitrophes fortement peuplées sont constituées normalement d'un ou de plusieurs îlots de recensement contigus qui ont une densité de population d'au moins 1 000 p.m.c.. Le Census Bureau se sert de la densité de population comme indice d'urbanisation parce que cette mesure est facile à calculer et qu'elle constitue une base solide au point de vue démographique pour déterminer l'importance des centres de population et leur répartition. De plus, la valeur des facteurs qui entrent dans le calcul de la densité de population, à savoir les chiffres de population et la superficie des unités géographiques, est connue dans les six mois qui suivent le jour du recensement.

Inclusion d'agglomérations "complètes"

À l'heure actuelle, le Census Bureau définit les RU en fonction d'agglomérations complètes; une agglomération située dans une zone périphérique à forte densité de population sera incluse dans la RU si elle compte au moins 2 500 habitants et si au moins 50% de sa population vit dans des îlots qui ont une densité de population de 1 000 p.m.c. ou plus.

Pour mieux distinguer la population urbaine de la population rurale pour les agglomérations de ce genre, le Census Bureau a adopté la notion de "prolongement urbain" en vue du recensement de 1970. Selon cette notion, il peut exister des zones rurales à l'intérieur du territoire d'une agglomération, pourvu que la densité de population de ces zones soit inférieure à 100 p.m.c.

Inclusion de zones à faible densité de population et de zones admissibles non contiguës

Une RU peut comprendre des "poches" à faible densité de population, comme des parcs ou des cours de triage. Ces espaces peuvent se trouver aussi bien au sein de l'agglomération centrale que dans la périphérie. Les critères actuels prévoient aussi l'inclusion de zones admissibles non contiguës qui sont séparées du territoire principal de la RU par des zones à faible densité de population mais qui y sont reliées par une route.

Exclusion de certaines superficies du calcul de la densité de population

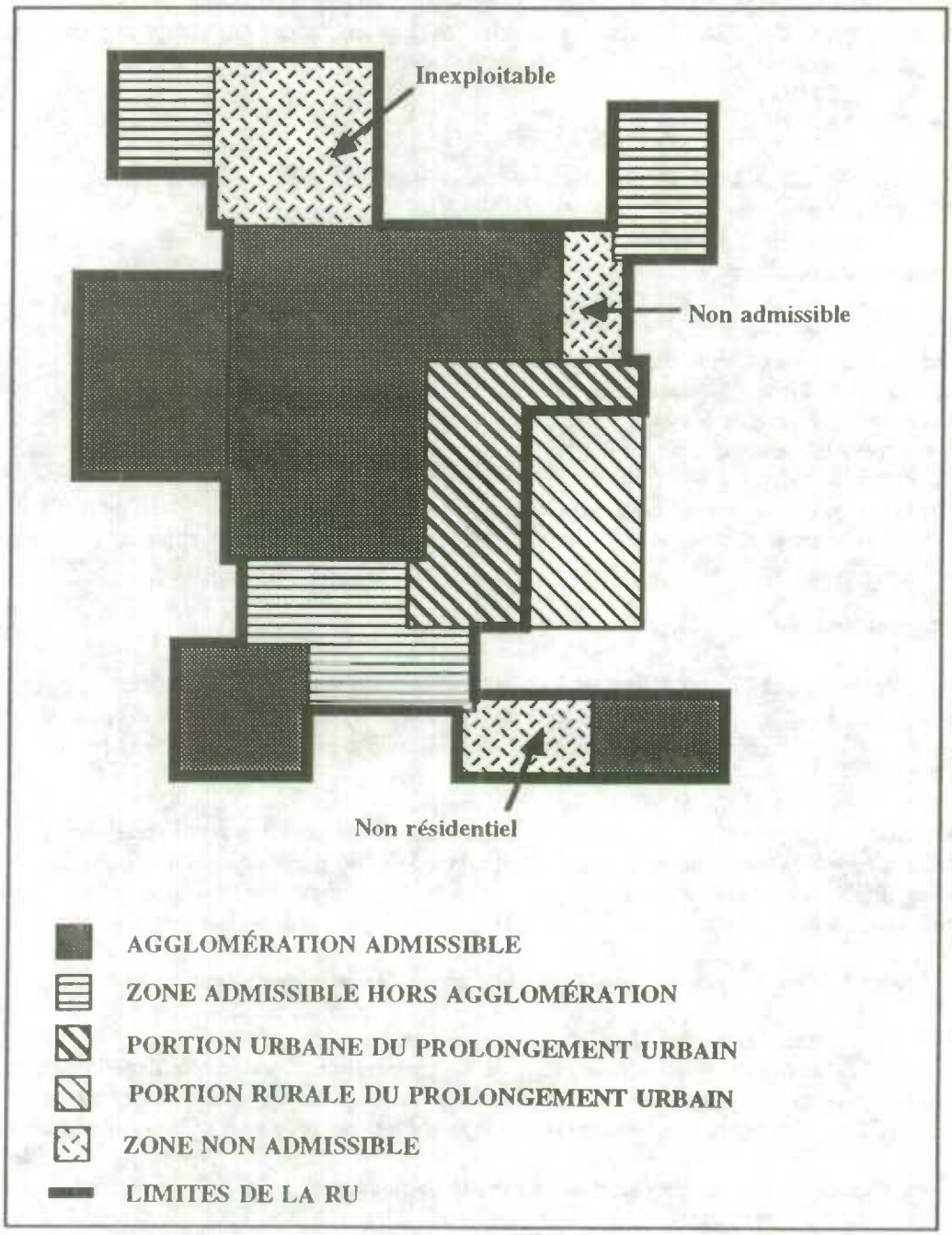
Les superficies consacrées à un usage industriel, commercial ou récréatif ou à l'industrie du transport (par ex.: grands parcs industriels, parcs nationaux ou aéroports) et les superficies inexploitables (par ex.: terres

marécageuses ou slikkes) peuvent être exclues du calcul de la densité de population si le Census Bureau dispose de renseignements prouvant leur existence. Lorsque des superficies de ce genre longent une portion d'une RU, elles ne font pas partie pour autant de cette RU, à moins qu'elles permettent de rejoindre des zones admissibles non contiguës qui ont une densité de population d'au moins 1 000 p.m.c.

Fusion de RU

En règle générale, le Census Bureau fusionne des RU adjacentes lorsque celles-ci sont incluses en majeure partie dans la même ZSM (zone statistique métropolitaine) ou la même ZSMM (zone statistique métropolitaine majeure).

Figure 1. Région urbanisée hypothétique



Pour le recensement de 1990, la délimitation des RU était largement automatisée. Dans les douze centres régionaux du recensement (CRR) du Census Bureau, des géographes traçaient les limites probables des RU de 1990 à l'aide d'un logiciel conçu par la division de la géographie. Des géographes de cette division revoyaient ensuite le travail de leurs collègues et apportaient des corrections au besoin. Dans un deuxième temps, ils retenaient toutes les RU qui comptaient au moins 50 000 habitants; ces dernières sont les régions pour lesquelles le Census Bureau a produit des totalisations en 1990.

4. PERSPECTIVES DE RECHERCHE

Nous sommes en train d'évaluer les définitions, les critères et les méthodes qu'utilise le Census Bureau pour définir la population urbaine. Quatre grands sujets de recherche attirent particulièrement notre attention: 1) concepts et définitions, 2) techniques de délimitation, 3) variabilité des critères et 4) nouvelles régions géographiques.

4.1 Concepts et définitions

Agglomérations urbaines

Le Census Bureau classe comme urbaines des agglomérations complètes (municipalités ou agglomérations de recensement) qui sont à l'extérieur de RU si ces agglomérations comptent au moins 2 500 habitants. Nous faisons de même pour des agglomérations situées à l'intérieur de RU, y compris les îlots qui n'ont pas la densité de population requise (1 000 p.m.c.). Selon les critères du Census Bureau, les îlots qui ne font pas partie d'une agglomération urbaine ou d'une RU ne peuvent être définis comme urbains, peu importe la population ou la densité de population de ces îlots. Par conséquent, beaucoup d'îlots peu peuplés à l'intérieur comme à l'extérieur de RU sont définis comme urbains, alors que de nombreux îlots populeux à l'extérieur de RU sont définis comme ruraux du seul fait qu'ils ne font pas partie d'une agglomération. Nous devrions revoir notre définition de l'agglomération urbaine pour corriger cette anomalie.

Classification selon l'utilisation des terres

Selon les normes du Census Bureau, deux catégories de terrain à population éparse peuvent faire partie d'une RU: le terrain inexploitable et le terrain urbain non résidentiel. En classant un îlot de recensement dans l'une ou l'autre de ces catégories, nous pouvons inclure les îlots voisins à forte densité de population dans une RU qui, normalement, ne pourrait les contenir en raison de leur éloignement (figure 1). Plusieurs régions ont pu être reconnues comme RU en 1990 du seul fait que nous avions pu y introduire des îlots additionnels en nous fondant sur ce critère. Inversement, certaines régions n'ont pu être reconnues comme RU parce que nous ne pouvions appliquer le même critère.

L'emploi de cette classification pose plusieurs problèmes. Premièrement, nous n'avons pas été capables d'élaborer des définitions claires et objectives qui peuvent être appliquées de façon cohérente à la grandeur du pays. Deuxièmement, le Census Bureau ne dispose pas de sources de données complètes qui décrivent les caractéristiques de toutes les parcelles de terre aux États-Unis. En vue du recensement de l'an 2000, nous voulons créer des méthodes plus objectives pour la classification des terres, déterminer s'il est possible d'obtenir et d'exploiter les données de référence nécessaires à l'application de ces méthodes ou encore, déterminer si nous sommes en mesure d'élaborer de nouvelles techniques de délimitation qui rendraient inutile ce genre de classification.

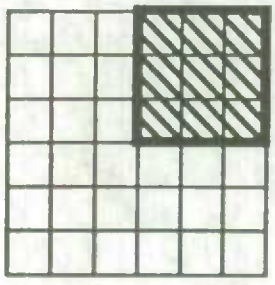
4.2 Techniques de délimitation

Moyenne mobile

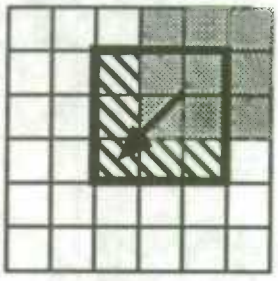
La manière classique de délimiter une RU consiste à examiner un îlot (ou, parfois, des groupes d'îlots semblables et contigus) tout en tenant compte des limites des agglomérations et de l'utilisation des terres. Si on détermine que ces îlots peuvent faire partie de la RU, on examine l'îlot (ou le groupe d'îlots) suivant. Nous envisageons actuellement l'utilisation d'une "technique de moyenne mobile", par laquelle nous pourrons tracer une limite

entre les populations urbaine et rurale grâce à une suite d'opérations fondée sur le chevauchement de groupes d'îlots. Selon cette technique, nous constituons des groupes d'îlots contigus et calculons la densité de population pour ces groupes; le résultat obtenu détermine si les îlots en question doivent être inclus ou non dans la RU. Nous constituons ensuite un autre groupe d'îlots contigus avec des îlots que nous venons d'inclure dans la RU et d'autres qui n'ont pas encore été examinés (figure 2) et nous refaisons le calcul. Tant que les groupes d'îlots étudiés successivement répondent aux critères d'inclusion, nous continuons de former de nouveaux groupes avec un certain nombre d'îlots du groupe étudié précédemment.

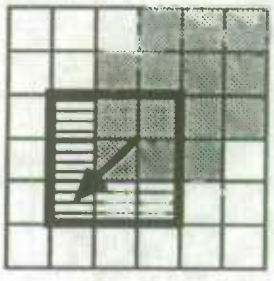
Figure 2. Technique de la moyenne mobile



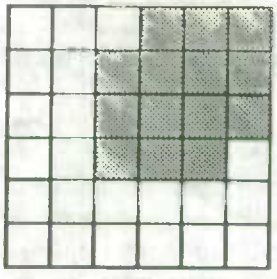
1. Un groupe d'îlots contigus forme l'unité d'analyse de base d'une RU projetée. Cette unité peut faire partie de la RU projetée parce qu'elle a une densité de population suffisante.



2. Quatre îlots qui font déjà partie de la RU projetée sont inclus dans une nouvelle unité d'analyse. Compte tenu de la densité de population, cette nouvelle unité peut faire partie de la RU projetée.



3. Quatre îlots qui font déjà partie de la RU projetée sont inclus dans une nouvelle unité d'analyse. Compte tenu de la densité de population, cette nouvelle unité ne peut faire partie de la RU projetée.



4. Finalement, la RU projetée se compose de tous les îlots des deux unités d'analyse admissibles.

LIMITES DE L'UNITÉ D'ANALYSE

ÎLOTS AJOUTÉS À LA RU PROJETÉE

ÎLOTS DÉJÀ INCLUS DANS LA RU PROJETÉE

ÎLOTS QUI NE SONT PAS AJOUTÉS À LA RU PROJETÉE

Grâce à cette technique, nous n'avons plus à faire intervenir des classifications subjectives dans le processus de délimitation. De plus, en déterminant les limites d'une RU en fonction des densités de population de groupes d'îlots chevauchants, nous nous trouvons à éliminer les variations mineures et non représentatives de la densité de population entre les îlots d'une agglomération urbaine puisque ces variations sont "fondues" dans le calcul de la moyenne de densités de population d'îlots contigus. Nous pouvons ainsi faire une meilleure analyse des densités de population des RU et déterminer la limite réelle qui sépare la population urbaine de la population rurale.

Zones de densité de population

Nous envisageons aussi d'étudier la possibilité de définir des zones correspondant à des tranches de densité de population dans des régions déjà reconnues comme des RU. Nous déterminerons en même temps si le fait d'adopter de nouveaux critères de densité pour pouvoir délimiter ces zones avec plus de précision aura un effet quelconque. Cette dernière affirmation donne à entendre qu'il est peut-être temps de cesser de se fonder sur un seul critère de densité de population pour délimiter les RU.

4.3 Variabilité des critères servant à définir une RU

Bon nombre des facteurs qui influent considérablement sur la délimitation des RU varient beaucoup d'une région à l'autre du pays. Par exemple, dans les villes du Nord-Est, plus anciennes, les tracés de rues sont souvent plus denses qu'ailleurs au pays, notamment dans l'Ouest, où les îlots sont plus grands à cause d'un tracé de rues moins serré et du manque d'objets hydrographiques pouvant servir convenablement de limite d'îlot. Dans de nombreux États, les municipalités ont toute la latitude voulue pour annexer du territoire tandis que dans d'autres États, il est à peu près impossible de le faire. À cause de ces différences géographiques et morphologiques, nous sommes tentés de définir des critères différents selon les régions.

Pour la première fois depuis que le Census Bureau définit des RU, des régions qui avaient été reconnues comme RU pour le recensement de 1980 ne l'étaient plus pour le recensement de 1990[4]. Le Census Bureau n'avait pas prévu de dispositions qui auraient permis de reconnaître ces régions comme des RU même si elles ne répondaient pas aux critères. Nous devons nous prononcer sur l'opportunité d'accorder un "statut spécial" aux RU, que ce soit pour des raisons théoriques ou pratiques, ou de définir un "délai de grâce" (par ex.: un recensement) pendant lequel la population d'une RU dûment reconnue tomberait sous le cap des 50 000 sans que la région en question ne perde son titre de RU. Par conséquent, nous allons aussi envisager de fixer un critère de population moindre pour des régions qui ont déjà été reconnues comme RU.

4.4 Nouvelles réglons géographiques

Étant donné les nombreuses applications des programmes de désignation des régions urbaines et de délimitation des RU et la grande importance qu'accordent de nombreux analystes à la comparabilité des données, nous pourrions envisager de ne pas modifier réellement nos critères et nos méthodes en vue du prochain recensement. Nous pourrions, à la place, définir simplement de nouvelles unités géographiques entre les recensements et laisser les utilisateurs choisir les produits qui, selon eux, répondent le mieux à leurs besoins.

5. CONCLUSION

Les questions qui ont été soulevées dans cette communication ne sont pas faciles à résoudre; les solutions proposées devront répondre aux besoins des personnes et des organisations qui sont le plus touchées par nos définitions des populations urbaine et rurale. Nous ne modifierons pas ces définitions sans avoir obtenu un large consensus de la part du public. Par ailleurs, nous reconnaissons la nécessité d'adopter une méthode dynamique et pénétrante pour améliorer nos définitions, une méthode qui procurera aux utilisateurs de données une classification plus utile des populations urbaine et rurale.

---

[4] Danville (Illinois) et Enid (Oklahoma) ont perdu leur statut de RU par suite du recensement de 1990.

SESSION 10

Analyse des données sous des perspectives géographiques

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

ANALYSE STATISTIQUE DE DONNÉES SPATIALES URBAINES DU RECENSEMENT ÉTANT DONNÉ DES VALEURS MANQUANTES

D.A. Griffith[1]

RÉSUMÉ

L'auteur présente une méthode d'estimation des valeurs manquantes dans une analyse de la distribution géographique du revenu médian de la famille par secteur de recensement en 1986 dans la région métropolitaine d'Ottawa-Hull. Cette méthode repose sur le principe du maximum de vraisemblance. On établit une comparaison entre les résultats d'un modèle de réponse autorégressif obtenus à l'aide des données du recensement de 1986 au Canada, les résultats d'un modèle de régression classique obtenus à l'aide des mêmes données et les résultats d'un modèle autorégressif conditionnel obtenus à l'aide des données du recensement de 1980 aux États-Unis pour la région de Houston. On s'intéresse en même temps à l'amélioration de la précision des estimations de valeurs manquantes.

MOTS CLÉS: Autorégressif; algorithme E-M; valeurs manquantes; précision; autocorrélation spatiale.

1. INTRODUCTION

Les valeurs manquantes peuvent gêner considérablement l'analyse visuelle et l'analyse ordinaire des données spatiales. Elles sont la cause de distributions géographiques incomplètes et nuisent à la généralisation de secteurs cartographiques. Le but de cette communication est de décrire une méthode de traitement des valeurs manquantes parmi des données urbaines du recensement au Canada. Une version de cette méthode a déjà été appliquée à une série de données urbaines du recensement aux États-Unis (voir Griffith, Bennett et Haining 1989).

On totalise souvent des données d'enquête sur échantillon selon des unités spatiales déterminées (par ex.: secteur de recensement) et on publie ensuite les chiffres agrégés. Les règles de protection du secret statistique nous obligent à supprimer souvent certains chiffres car avec ce genre de totalisation *a posteriori*, il n'est pas sûr que l'on ne pourra jamais associer des données publiées à un ménage identifiable et que, par conséquent, on réussira à préserver le caractère confidentiel des données relatives à chaque ménage. Il n'y a donc aucune raison *a priori* de croire que les valeurs manquantes sont le résultat d'un facteur systématique quelconque; c'est pourquoi elles sont considérées comme des variables aléatoires dans cette communication.

1.1 Contexte

Griffith, Bennett et Haining (1989) ont étudié la distribution géographique du revenu médian de la famille par secteur de recensement en 1980 dans la région métropolitaine de Houston. La région avait été divisée en 363 secteurs de recensement et on avait supprimé les données relatives au revenu pour trois de ces secteurs. Comme pour Ottawa-Hull, la méthode de censuration appliquée pour Houston était déterminée par le nombre de ménages présents dans un secteur de recensement et était donc considérée comme indépendante du revenu médian de la famille.

---

[1] D.A. Griffith, Département de géographie et Programme de statistique interdisciplinaire, Université de Syracuse, Syracuse, New York, États-Unis, 13244-1160.

Dans le cas de la variable de revenu de Houston, on a observé une autocorrélation spatiale positive relativement forte, ce qui a donné le paramètre $\hat{\rho} = 0.1747$ ($\hat{\rho}_{max} = 0.1755$) pour le modèle autorégressif conditionnel. Cette distribution géographique est caractérisée par une surface moyenne non uniforme; on a pu déterminer que le revenu dans le secteur de recensement i était une fonction des coordonnées cartésiennes (u_i, v_i) du centroïde de ce secteur. La précision des estimations de valeurs manquantes de type spatial était supérieure à celle des estimations de valeurs manquantes de type classique dans des proportions de 10, de 30 et de 4%. Malheureusement, les intervalles de confiance correspondants étaient très étendus (faible degré de précision) et dans un cas en particulier, la limite inférieure était négative, ce qui a obligé les expérimentateurs à tronquer l'intervalle à zéro pour des raisons pratiques. On a conclu notamment qu'il fallait inclure des variables explicatives additionnelles dans la spécification du modèle.

1.2 Région métropolitaine d'Ottawa-Hull

Les données que nous allons analyser dans cette communication sont tirées du recensement de la population du Canada de 1986 -- les données sous forme électronique nous ont été fournies par Statistique Canada -- et ont trait aux 192 secteurs de recensement qui composent la région métropolitaine d'Ottawa-Hull. Cette région compte au total 819 263 habitants et a une superficie de 5138.33 kilomètres carrés, pour une densité moyenne de population de 159.44 habitants au kilomètre carré. Les centroïdes des secteurs de recensement sont exprimés en coordonnées U.T.M. (Universal Transverse Mercator -- Projection universelle transverse de Mercator); comme ces coordonnées se trouvent sur une échelle d'intervalles, elles ont été transformées pour des raisons de commodité. La moyenne du revenu médian de la famille par secteur de recensement -- pour les secteurs où la valeur de cette variable est connue -- pour l'ensemble de la région métropolitaine est de $40 522, avec des bornes inférieure et supérieure de $16 908 et de $59 902; le revenu médian réel de la famille pour la région urbaine est de $41 775.

Une particularité complexe de la région d'Ottawa-Hull est que celle-ci se compose de deux grandes "régions culturelles". La limite qui sépare ces deux régions correspond grosso modo à la rivière des Outaouais, qui est aussi une frontière politique. L'une des régions est située dans la province d'Ontario tandis que l'autre est située dans la province de Québec. Cette dualité sera prise en considération dans l'analyse grâce à l'introduction d'une variable indicatrice; cette variable prendra la valeur 1 dans le cas d'un secteur de recensement situé en Ontario et la valeur -1 pour un secteur situé au Québec (ce paramétrage rend possible l'exécution de tests de différence de moyennes).

2. DÉTERMINANTS DE LA DISTRIBUTION GÉOGRAPHIQUE DU REVENU DANS UNE RÉGION URBAINE

Selon les spécialistes de l'économie urbaine, trois facteurs fondamentaux influent sur la distribution géographique du revenu dans les villes (voir Richardson 1977); ce sont la densité de la population, les gradients et les externalités spatiales. Premièrement, le revenu est inversement proportionnel à la densité de la population. L'élasticité-revenu de la demande d'espace chez les ménages bien nantis est telle qu'ils ont une préférence marquée pour les grands espaces. Dans leur cas, la maximisation de l'utilité a pour conséquence de créer le rapport inverse mentionné ci-dessus, puisque cette classe de ménages est prête à conserver la facilité d'accès au centre-ville en échange de frais de transport plus élevés. Idéalement, une variation minime du revenu compenserait une variation des frais de transport qui découlerait du choix d'un endroit donné. Ainsi, étant donné que les classes à revenu élevé ont une préférence pour les secteurs à faible densité de population et qu'elles jouissent des moyens financiers voulus, le revenu tend à augmenter lorsque la densité de la population diminue.

Deuxièmement, il existe des gradients de revenu pour chaque région urbaine. La surface de revenu d'une région peut rappeler par exemple la forme d'un chapiteau ou celle d'une toile d'araignée en trois dimensions, avec de nombreux sommets et de nombreux creux. Le sommet le plus haut est souvent observé au centre de l'agglomération. Des sommets secondaires sont observés dans les quartiers résidentiels. Les creux sont imputables à des utilisations concurrentes du sol ou à l'existence de quartiers délabrés. Cette caractérisation est conforme aux modèles de configuration spatiale des zones urbaines de Burgess et de Hoyt.

Troisièmement, le lieu de résidence est aussi fonction des attraits du quartier, de la préférence pour un cadre de vie agréable et des caractéristiques socio-économiques du voisinage. Ces éléments créent des externalités spatiales, qui viennent de ce que certains lieux d'habitation procurent des avantages particuliers, outre la facilité d'accès au centre-ville. Par conséquent, les ménages qui appartiennent à une même catégorie de revenu tendent à occuper le même quartier, les ménages à revenu élevé étant disposés à supporter le coût des externalités positives. Ces "effets secondaires" ont pour conséquence d'intégrer la notion d'autocorrélation spatiale à la distribution géographique du revenu. Cette caractérisation est conforme à la notion de configuration spatiale des zones urbaines basée sur le modèle des noyaux multiples de Harris et Ullman.

3. SPÉCIFICATION D'UN MODÈLE AUTORÉGRESSIF POUR OTTAWA-HULL

Deux caractéristiques de modèle sont en cause ici. Il s'agit, premièrement, du jacobien de la transformation d'un domaine autocorrélé en un domaine non autocorrélé, particulièrement lorsque n = 192 et que la surface est décomposée en secteurs de forme irrégulière, et, deuxièmement, du choix de la spécification du modèle autorégressif spatial.

Le jacobien pour Ottawa-Hull implique une matrice 192 X 192, dont le déterminant doit être calculé à plusieurs reprises. Griffith (1992) propose une approximation numérique utile pour ce déterminant. Si le modèle autorégressif est exprimé en fonction de la version normalisée de la matrice d'adjacence binaire C ($c_{ij} = 1$ si les secteurs de recensement i et j sont contigus et $c_{ij} = 0$ dans le cas contraire; chaque ligne est rajustée de telle manière que la somme de ses éléments égale un), par exemple W ($w_{ij} = c_{ij} / \sum_{j=1}^{n} c_{ij}$), les valeurs propres de cette matrice sont telles que

$$\lambda_{max} = 1 \text{ and } \lambda_{min} = 0.62371 \rightarrow -1.601706 < \hat{\rho} < 1.$$

En prélevant un échantillon systématique de taille $n = 22$ dans cet espace de paramètres admissible, on obtient l'approximation jacobienne suivante ($SCR = 0.000099$):

$$\begin{aligned} \hat{J}_w^{1/n} &= 0.237169 \bullet ln(1.873080) + 0.144759 \bullet ln(1.159028) - \\ & 0.237169 \bullet ln(1.873080 + \rho) - 0.144759 \bullet ln(1.159028 - \rho). \end{aligned} \tag{3.1}$$

Cette formule d'approximation a servi à l'étude qui est résumée dans cette communication.

Nous avons choisi le modèle de réponse autorégressif parmi le groupe des modèles autorégressifs spatiaux possibles à cause principalement de sa capacité de réduire au maximum le nombre de variables spatiales de décalage qui doivent être prises en considération. Considérons un ensemble de p variables, dont les valeurs composent la matrice X. En outre, un vecteur de uns, désigné par 1, doit être rattaché à ces p variables de manière à inclure une ordonnée à l'origine. Soient Y le vecteur des variables de réponse et ξ le vecteur d'erreurs. Le modèle statistique linéaire classique qui peut être construit à l'aide de ces termes a la forme $Y = X\beta + \xi$, où β est un vecteur $(p+1)$-by-1 de coefficients de régression. Le modèle de réponse autorégressif spatial peut donc s'écrire

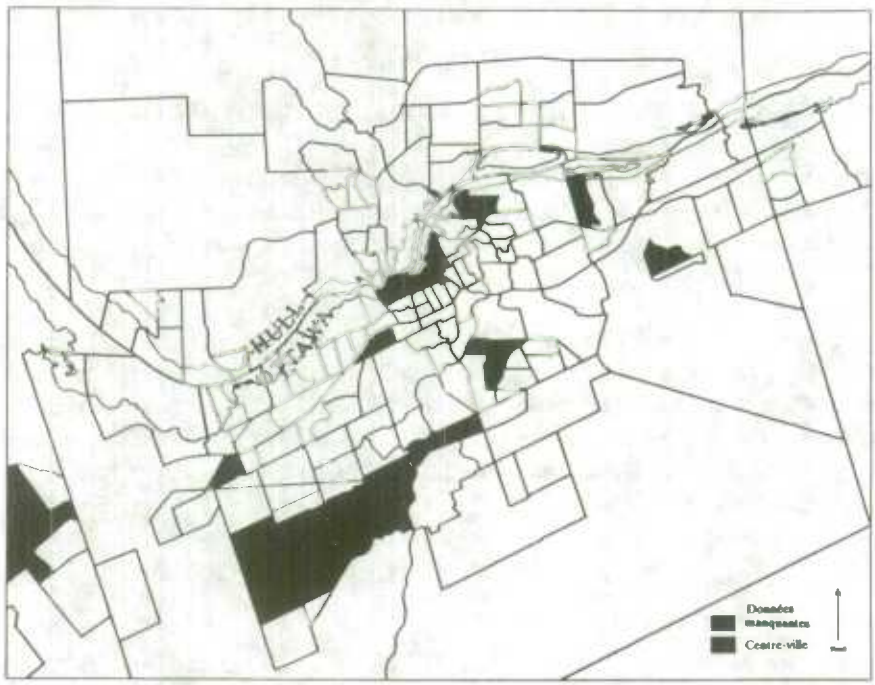
$$(I - \rho W)Y = X\beta + \xi \quad \text{or} \quad Y = \rho WY + X\beta + \xi,$$

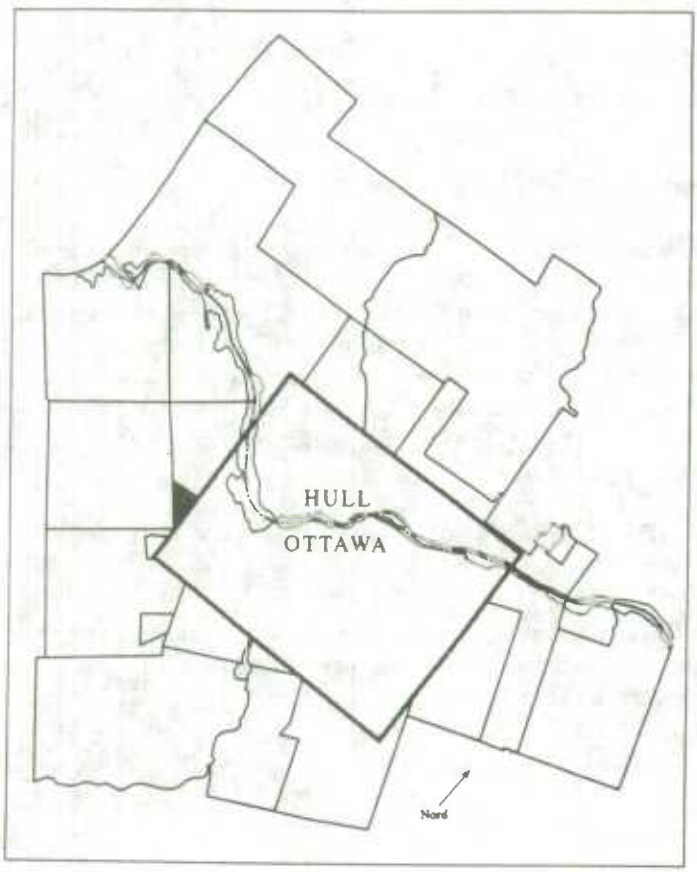
où I est la matrice unité. C'est la spécification que nous allons utiliser dans cette communication. Elle a été décrite notamment par Upton et Fingleton (1985). On peut en estimer les paramètres au moyen de progiciels statistiques courants en la réécrivant sous la forme:

$$Y\exp(\hat{J}_w^{1/n}) = \rho WY\exp(\hat{J}_w^{1/n}) + X\beta\exp(\hat{J}_w^{1/n}) + \xi\exp(\hat{J}_w^{1/n}). \tag{3.2}$$

Griffith (1988b) décrit une méthode qui permet d'estimer les paramètres de la spécification au moyen de l'équation (3.2).

Figure 1 : Carte montrant les secteurs de recensement pour lesquels on ne connaît pas le revenu médian de la famille en 1986.





4. ESTIMATION DES VALEURS MANQUANTES DU REVENU MÉDIAN PAR SECTEUR DE RECENSEMENT POUR LA RÉGION D'OTTAWA-HULL

Dans la région métropolitaine d'Ottawa-Hull, trois secteurs de recensement[2] comptent respectivement 13, 45 et 64 personnes. Un échantillon aléatoire national sans contrainte permettra très difficilement d'obtenir pour ces unités spatiales des échantillons dont la taille respecte les normes minimum en matière de protection du secret statistique. À Statistique Canada, on croit que l'effectif d'un secteur de recensement doit même être supérieur à 250 pour qu'on puisse obtenir un échantillon conforme aux règles de protection du secret statistique. Par ailleurs, les chiffres du revenu médian de huit autres secteurs de recensement[3] ont été supprimés. Comme les populations respectives de ces secteurs varient de 1,766 à 6,083, il est permis de croire que cette suppression est la conséquence de l'utilisation par Statistique Canada d'un plan d'échantillonnage stratifié qui fait abstraction de la dimension spatiale. La figure 1 contient une carte de la région qui indique les secteurs de recensement pour lesquels il y a des valeurs manquantes. À l'exception d'une grappe de trois secteurs contigus, les secteurs touchés sont répartis un peu partout sur le territoire de la municipalité.

4.1 Solution fondée sur l'algorithme E-M classique

Pour estimer les valeurs manquantes du revenu médian de la famille pour la région d'Ottawa-Hull au moyen de l'algorithme E-M classique (estimation des valeurs manquantes par itération, suivie d'une maximisation de la fonction de vraisemblance; voir Little et Rubin 1987), on procède de la façon suivante:

Étape 1: calculer $b_{\tau=0} = (X_o^t X_o)^{-1} X_o^t Y$

Étape 2: calculer $\hat{Y}_{m,\tau} = X_m b_\tau$

Étape 3: calculer $b_{\tau+1} = (X^t X)^{-1} X^t \hat{Y}$, où $\hat{Y}^t = \langle Y_o^t \vdots \hat{Y}_{m,\tau} \rangle$,

Étape 4: répéter les étapes 2 et 3 jusqu'à ce que $|b_{\tau+1} - b_\tau| \approx 0$.

Ces opérations ont été exécutées à l'aide de PROC NLIN de SAS (voir section 5.1).

L'exécution de l'étape 1 exigeait la spécification de l'équation de régression. Griffith, Bennett et Haining (1989) ont observé que la surface du revenu médian de la famille pour Houston pouvait être décrite en partie par un modèle de gradient linéaire (ou de surface de dérive). De plus, en économie urbaine, les surfaces de revenu urbain sont définies comme une fonction de la densité de population, conformément à ce que nous disions dans la section 2. Par ailleurs, en ce qui concerne la région d'Ottawa-Hull, il est possible de tenir compte des différences culturelles en introduisant, comme nous l'avons dit dans la section 1.2, une variable indicatrice. Les trois facteurs ont été analysés à l'origine. Les résultats de cette analyse figurent dans le tableau 1. Selon le principe de la somme des carrés additionnelle, aucun des termes de coordonnées U.T.M. ne s'est avéré significatif. En fait, l'axe de coordonnées nord-sud linéaire est devenu important lorsqu'on a introduit la densité de population dans le modèle de régression. Ainsi, la moyenne conditionnelle du revenu médian de la famille par secteur de recensement a été établie à $42 419, le revenu médian pour les secteurs de recensement d'Ottawa se situant, en moyenne, à $4 578 au-dessus et le revenu médian pour les secteurs de Hull se situant, en moyenne, à $4 578 au-dessous. En outre, lorsque la densité de la population augmente, le revenu médian de la famille par secteur de recensement tend à diminuer. La différence culturelle et la densité de la population expliquent environ le cinquième de la variation géographique du revenu médian de la famille dans la région d'Ottawa-Hull. Le diagnostic de modèle a produit une statistique de Shapiro-Wilk non significative de 0.9910, ce qui confirme que la population-mère des résidus de régression suit une distribution normale. L'analyse graphique de l'hétéroscédasticité indique que la variance d'erreur n'est peut-être pas toujours constante, mais cette déviation est sans gravité (peut-être est-ce l'indice de la présence de plusieurs valeurs aberrantes négligeables). Enfin, le coefficient de Moran significatif observé pour les résidus de régression (0.1876) dénote l'existence d'un faible degré d'autocorrélation spatiale positive.

---

[2] Ceux portant les numéros 47, 140.01, et 160.03.

[3] Ceux portant les numéros 2.01, 6, 34, 110, 120.01, 125.01, 130.01, et 137.03.

Tableau 1: Résultats de l'estimation étant donné des valeurs manquantes

Modèle de régression linéaire classique sans les valeurs manquantes

Ordre du modèle	modèle de régression d.l.	modèle de régression somme des carrés	résidus de régression d.l.	résidus de régression somme des carrés	df	somme des carrés additionnelle	R^2
Zéro	2	4033904135.2	178	15167150599	2	4033904135*	0.21
Linéaire	4	4322552193.5	176	14878502541	2	288648058	0.23
Quadratique	7	4612198066.2	173	14588856668	3	289645873	0.24
Cubique	11	4620744265.8	169	14580310469	4	8546199	0.24

* représente une proportion appréciable de la somme des carrés de la régression par comparaison à l'erreur quadratique moyenne du modèle au complet.

Estimations calculées à l'aide d'un modèle de régression linéaire classique avec des valeurs manquantes

Méthode	β_0	β_1	β_2	MSE
Pas à pas	42419.22165	4577.94548	-1.37701	85208711
Par itération	42419.19599	4577.93092	-1.37700	80249474

Estimation calculées à l'aide d'un modèle d'autorégression spatial avec des valeurs manquantes

Méthode	ρ	β_0	β_1	β_2	MSE
E-M	0.44900	23451.43946	3051.74277	-0.94187	71722838
Par itération	0.47327	22456.38101	2958.47903	-0.92201	71425230

Les étapes 2, 3 et 4 ont donné des résultats à peine différents de ceux obtenus à l'étape 1 (voir tableau 1). L'erreur quadratique moyenne (EQM) doit être multipliée par 189/178 pour équivaloir à celle calculée à l'étape 1. Nous nous sommes servis des coefficients recalculés à ces étapes pour estimer les valeurs manquantes du revenu médian de la famille.

Les tableaux 2 et 3 contiennent les estimations des valeurs manquantes du revenu médian de la famille. On obtient les résultats du tableau 2 en considérant les valeurs manquantes comme des paramètres. À chaque estimation correspond un intervalle de confiance à 95% (voir section 6). Dans presque tous les cas, la limite inférieure est beaucoup plus grande que la valeur minimum observée du revenu médian; la valeur estimée est toujours supérieure à la moyenne ou à la moyenne conditionnelle du revenu médian et la limite supérieure est, elle aussi, toujours au-dessus de la valeur maximum observée (l'écart peut atteindre 10%). Par contre, en ce qui concerne les espérances conditionnelles qui figurent dans le tableau 3, les limites inférieures sont beaucoup plus près de la valeur minimum observée (deux de ces limites étant au-dessous de cette valeur), les valeurs moyennes sont égales aux estimations de paramètres et les limites supérieures sont encore plus grandes que celles des estimations du tableau précédent (toutes sont plus élevées que la valeur maximum observée).

Tableau 2: Estimations des valeurs manquantes du revenu médian de la famille pour les onze secteurs de recensement de la région d'Ottawa-Hull: valeurs manquantes considérées comme des paramètres

	Modèle de régression classique			Modèle d'autorégression spatial		
Numéro de secteur de recensement	Limite inf. d'un I.C. à 95%	Estimation de la valeur manquante	Limite sup. d'un I.C. à 95%	Limite ind. d'un I.C. à 95%	Estimation de la valeur manquante	Limite sup. d'un I.C. à 95%
2.01	26734.14	45046.94	63359.74		44392.6	
6.00	26302.88	44606.87	62910.86		48817.9	
34.00	24234.79	42519.35	60803.91		36794.2	
47.00	28540.47	46908.65	65276.82		35006.3	
110.00	27067.52	45388.29	63709.05		45050.0	
120.01	27216.36	45541.00	63865.65		50668.6	
125.01	23475.30	41761.96	60048.63		40427.1	
130.01	28142.80	46496.19	64849.58		49171.1	
137.03	26313.69	44617.88	62922.07		42260.0	
140.01	28624.61	46996.10	65367.59		53205.4	
160.03	28620.85	46992.19	65363.53		50336.2	

Tableau 3: Estimations des valeurs manquantes du revenu médian de la famille pour les onze secteurs de recensement de la région d'Ottawa-Hull: valeurs manquantes considérées comme des espérances d'observations

	Modèle de régression classique			Modèle d'autorégression spatial		
Numéro de secteur de recensement	Limite inf. d'un I.C. à 95%	Estimation de la valeur manquante	Limite sup. d'un I.C. à 95%	Limite inf. d'un I.C. à 95%	Estimation de la valeur manquante	Limite sup. d'un I.C. à 95%
2.01	19285.36	45046.94	70808.53	29260.1	45144.1	61028.1
6.00	18845.28	44606.87	70368.46	30136.9	46024.7	61912.4
34.00	16757.76	42519.35	68280.94	23318.4	39247.8	55177.2
47.00	21147.06	46908.65	72670.23	22223.3	38586.0	54948.7
110.00	19626.70	45388.29	71149.87	28655.8	44551.0	60446.2
120.01	19779.42	45541.00	71302.59	32037.2	47964.0	63890.8
125.01	16000.38	41761.96	67523.55	23936.2	39828.6	55721.1
130.01	20734.60	46496.19	72257.78	32260.5	48193.3	64126.1
137.03	18856.29	44617.88	70379.47	27645.6	43530.1	59414.6
140.01	21234.51	46996.10	72757.69	34357.7	50353.9	66350.1
160.03	21230.60	46992.19	72753.78	32822.8	48770.6	64718.5

4.2 Solution fondée sur une version statistique spatiale de l'algorithme E-M: fondements algébriques

Martin (1984) présente l'équation générale qui sert à estimer les valeurs manquantes dans les modèles autorégressifs spatiaux. Dans son argumentation, il note que le jacobien doit être pondéré de telle sorte que, pour une structure de covariance inverse générale définie par la matrice V, il devienne:

$$J = -1n(\det|V|/\det|V_{mm}|)/n_o,$$

où $\det|V_{mm}|$ est la matrice de covariance inverse pour la configuration des valeurs manquantes et n_o est le nombre de valeurs non manquantes. L'équation (3.1) est une approximation de $-1n(\det|I - \rho W|)/192$. Sur les onze valeurs manquantes pour la région d'Ottawa-Hull, huit sont dispersées tandis que les trois autres sont disposées linéairement. Par conséquent, $\det|I - \rho W_{mm}| = \det|I| \cdot (1-\rho/24)(1+\rho/24)$ puisque les valeurs propres de la matrice W_{mm} sont 1/24, neuf zéros et -1/24. Les résultats obtenus par Martin impliquent que l'équation (3.1) doit être modifiée de la façon suivante:

$$J_w^{Vs} = (192/181)[0.237169 \bullet 1n(1.873080) + 0.144759 \bullet 1n(1.159028) - 0.237169 \bullet 1n(1.873080 + \rho) - 0.144759 \bullet 1n(1.159028 - \rho)] + [1n(1-\rho/24) + 1n(1+\rho/24)]/181. \quad (4.1)$$

Soit dit en passant, il est possible d'estimer $\det(|V_{mm}|)$ lorsque les valeurs manquantes forment des grappes assez appréciables; sa valeur devient 1 lorsque les valeurs manquantes sont toutes dispersées.

Griffith, Bennett et Haining (1989) ont décrit une méthode d'estimation fondée sur le modèle autorégressif conditionnel, où

$$V = I - \rho C.$$

Griffith (1988a), pour sa part, a décrit une méthode d'estimation fondée sur le modèle autorégressif simultané, où

$$V = (I - \rho W)^t (I - \rho W).$$

À l'heure actuelle, il est très difficile d'appliquer ces modèles à l'aide de progiciels statistiques courants comme le SAS. En revanche, le modèle de réponse autorégressif qui a été choisi pour cette étude a pour équation d'estimation des valeurs manquantes

$$\hat{Y}_m = [(I - \rho W_{mm})^t (I - \rho W_{mm}) + \rho^2 W_{om}^t W_{om}]^{-1} \{[(I - \rho W_{mm})^t X_m - \rho W_{om}^t X_o]\beta + \rho[W_{om}^t (I - \rho W_{oo}) + (I - \rho W_{mm})^t W_{mo}] Y_o\}. \quad (4.2)$$

Cette équation se ramène à $\hat{Y}_m = X_m\beta$ lorsque les données sont non autocorrélées spatialement (c.-à-d., $\rho = 0$), ce qui correspond à l'algorithme classique.

Les équations (4.1) et (4.2) ont servi à calculer les estimations géographiques présentées dans cette communication.

4.3 Solution fondée sur une version statistique spatiale de l'algorithme E-M: estimations

Pour estimer les valeurs manquantes du revenu médian de la famille pour la région d'Ottawa-Hull à l'aide d'une version statistique spatiale de l'algorithme E-M (c.-à-d. d'un modèle de réponse autorégressif spatial), on procède de la façon suivante:

Étape 1: calculer $b_{\tau=0} = (X_o^t X_o)^{-1} X_o^t Y$,

Étape 2: calculer $\hat{Y}_{m,o} = X_m b_{\tau=0}$,

Étape 3: estimer $\hat{\rho}_\tau$ et b_τ pour $\hat{Y} = \rho W\hat{Y} + X\beta + \xi$, où $\hat{Y}^t = \langle Y_o^t : \hat{Y}_{m,\tau}^t \rangle$,

Étape 4: calculer $\hat{Y}_{m,\tau+1}$ à l'aide de l'équation (4.2),

Étape 5: répéter les étapes 3 et 4 jusqu'à ce que $|b_{\tau+1} - b_\tau| \approx 0$.

Les étapes 1, 2, 3 et 5 ont été, elles aussi, exécutées à l'aide de PROC NLIN de SAS (voir section 5.2); l'étape 4 l'a été au moyen du code MINITAB à chaque itération.

Les étapes 2, 3, 4 et 5 ont donné des résultats très peu différents de ceux obtenus à l'étape 1 (voir tableau 1). L'erreur quadratique moyenne (EQM) doit être multipliée à la fois par le jacobien et par le facteur n/n_o pour être comparable à l'EQM calculée à l'étape 1. On note, comme dans la solution précédente, un degré d'autocorrélation spatiale positive modeste. La moyenne conditionnelle du revenu médian de la famille est tombée à $22 456 et l'écart de revenu médian entre Ottawa et Hull est réduit d'environ $3 239.

Les tableaux 2 et 3 renferment les estimations des valeurs manquantes du revenu médian de la famille. On a obtenu les estimations du tableau 2 en considérant les valeurs manquantes comme des paramètres. Ces estimations diffèrent des valeurs prédites du tableau 3, mais pas de façon appréciable. Pour ce qui a trait aux espérances conditionnelles du tableau 3, on observe des intervalles de confiance beaucoup plus étroits pour le modèle spatial que pour le modèle classique. Dans le cas du modèle spatial, les bornes inférieures dépassent largement la valeur minimum observée; par ailleurs, la non-linéarité fait que les valeurs moyennes s'écartent des estimations de paramètres et les limites supérieures sont plus près -- par rapport aux limites supérieures des estimations du modèle classique -- de la valeur maximum observée, bien qu'un grand nombre d'entre elles la dépassent.

5. LE CODE SAS

Le code SAS a été conçu pour estimer des valeurs manquantes. Il utilise PROC NLIN, qui produit les solutions itératives décrites plus haut. Ce code est exposé dans les annexes I et II.

5.1 Code pour l'estimation selon l'algorithme classique

Le code SAS pour la solution décrite dans la section 4.1 figure à l'annexe I. Les lignes 1 et 2 définissent la voie d'accès aux fichiers d'entrée. Les lignes 3-12 permettent l'accès aux données d'attribut contenues dans le fichier "OTT-HULL DATA". Les lignes 6 et 7 permettent la transformation des coordonnées U.T.M.. La ligne 8 crée la variable indicatrice "culturelle". La ligne 11 calcule la variable de densité de population. Les lignes 13-16 calculent les coefficients de régression linéaire pour les valeurs connues (étape 1 de la section 4.1). Les valeurs prédites, YHAT, servent alors de premières estimations pour les valeurs manquantes. Les lignes 17-21 substituent ces estimations aux valeurs manquantes. Les lignes 23-33 définissent des variables indicatrices pour chaque valeur manquante, du fait que l'on considère les valeurs manquantes comme des paramètres. Les lignes 36-38 initialisent les paramètres du modèle. La ligne 39 est le modèle de régression linéaire simple. Les lignes 40-42 permettent le remplacement des valeurs manquantes par leurs estimations (paramètres BM_j) à chaque itération dans PROC NLIN. Les lignes 43-89 sont les dérivées premières et les dérivées partielles croisées utilisées par PROC NLIN pour optimiser la fonction objectif construite à partir de l'énoncé de modèle de la ligne 39. La ligne 90 génère les espérances des valeurs manquantes comme s'il s'agissait d'observations; des intervalles de confiance à 95% sont calculés par la même occasion.

5.2 Code pour l'estimation selon une version spatiale de l'algorithme classique

Le code SAS pour la solution décrite dans la section 4.2 figure à l'annexe II. Comme dans le cas précédent, les lignes 1 et 2 définissent la voie d'accès aux fichiers d'entrée. Les lignes 3-11 permettent l'accès aux données d'attribut stockées dans le fichier "OTT-HULL DATA". La ligne 6 crée la variable indicatrice "culturelle" tandis que la ligne 9 calcule la variable de densité de population. La ligne 10 permet l'élimination des coordonnées U.T.M., étant donné qu'elles ne servent pas dans cette analyse. Les lignes 12-23 permettent l'accès à la matrice d'adjacence C, contenue dans le fichier "OTT-HULL CONN", pour la région métropolitaine d'Ottawa-Hull, calculent $\sum_{j=1}^{n} C_{ij}$ ainsi que la première partie de la variable de décalage spatiale CY. Les lignes 24-27 terminent le calcul de la variable de décalage et les lignes 28-30 transposent les résultats en un vecteur colonne. La ligne 47 transforme CY en WY. Les lignes 35-45 définissent des variables indicatrices pour chaque valeur manquante, du fait, là encore, que l'on considère les valeurs manquantes comme des paramètres. La ligne 53 initialise les paramètres du modèle. Les lignes 59-69 sont le produit du code MINITAB, qui calcule l'équation (4.2) (voir annexe III). Les lignes 70 et 71 définissent le jacobien de la transformation d'un espace autocorrélé en un espace

non autocorrélé; c'est l'équation (4.1). Les lignes 72-77 permettent le remplacement des valeurs manquantes par leurs estimations (paramètres BM_j) à chaque itération dans PROC NLIN. La ligne 78 définit le modèle de réponse autorégressif spatial. Les lignes 79-84 sont les dérivées premières utilisées par PROC NLIN pour optimiser la fonction objectif construite à partir de l'énoncé de modèle. La ligne 86 génère les espérances des valeurs manquantes comme s'il s'agissait d'observations; des intervalles de confiance à 95% sont calculés par la même occasion. Les lignes 87-91 permettent de calculer le quotient des valeurs prédites par le jacobien.

5.3 Code MINITAB

On s'est servi de MINITAB pour calculer par itération les estimations de Y_m à cause des possibilités qu'offre ce logiciel sur le plan des opérations matricielles; on aurait pu aussi bien utiliser SAS IML. Le code MINITAB est décrit dans l'annexe III. La ligne 2 enregistre les estimations SAS de $\hat{\rho}_{\tau-1}$ et de $b_{\tau-1}$. La ligne 8 enregistre les composantes d'estimation qui reposent sur la décomposition de la matrice W en blocs. La ligne 9 construit la matrice I_m. La ligne 10 calcule $W'_{mm} + W_{mm}$. La ligne 11 calcule $W'_{mm}W_{mm} + W'_{om}W_{om}$. La ligne 12 calcule $\hat{\rho}_{\tau-1}(W'_{mm} + W_{mm})$. La ligne 14 calcule $\hat{\rho}^2_{\tau-1}(W'_{mm}W_{mm} + W'_{om}W_{om})$. Les lignes 15-17 calculent $[I_m - \hat{\rho}_{\tau-1}(W'_{mm} + W_{mm}) + \hat{\rho}^2_{\tau-1}(W'_{mm}W_{mm} + W'_{om}W_{om})]^{-1}$. La ligne 18 calcule $W'_{mm}X_m$. La ligne 19 calcule $W'_{om}X_o$. La ligne 24 calcule $W'_{mm}X_m b_{\tau-1}$. La ligne 25 calcule $W'_{om}X_o b_{\tau-1}$. La ligne 26 calcule $X_m b_{\tau-1}$. La ligne 27 calcule $X_m b_{\tau-1} - \hat{\rho}_{\tau-1}(W'_{mm}X_m b_{\tau-1} + W'_{om}X_o b_{\tau-1}) + \hat{\rho}_{\tau-1}(W'_{om}Y_o + W_{mo}Y_o) - \hat{\rho}^2_{\tau-1}(W'_{om}W_{oo}Y_o + W'_{mm}W_{mo}Y_o)$. La ligne 28 calcule la valeur de l'équation (4.2). Enfin, la ligne 32 permet d'enregistrer un fichier qui sert à l'itération suivante.

6. PRÉCISION DES ESTIMATIONS DE VALEURS MANQUANTES

Les tableaux 2 et 3 nous renseignent sur la précision des estimations des valeurs manquantes en donnant des intervalles pour chaque estimation. Il y a deux façons d'envisager ces intervalles. Premièrement, on peut considérer les valeurs manquantes Y_m comme des paramètres et les estimer. Deuxièmement, on peut considérer ces mêmes valeurs comme des espérances conditionnelles d'observations, pour lesquelles sont construits soit des intervalles de confiance ou des intervalles de prédiction pour nouvelles observations.

Si on procède à une analyse classique, dans laquelle on suppose que les données codées suivant une grille géographique sont indépendantes et où les valeurs manquantes sont posées égales à la moyenne d'échantillon des valeurs connues, disons $\bar{x}_o$, les équations nécessaires pour le calcul des erreurs types asymptotiques sont les suivantes:

$$-E[\partial^2 ln(L)/\partial\mu^2] = n/\sigma^2,\ -E[\partial^2 ln(L)/\partial(y_{m_j})^2] = 1/\sigma^2,$$

$$-E[\partial^2 ln(L)/(\partial\mu\partial y_{m_j})] = -1/\sigma^2, \text{ et } -E[\partial^2 ln(L)/(\partial y_{m_j}\partial y_{m_k})] = 0.$$

Puisque $-E[\partial^2 ln(L)/(\partial\sigma^2\partial y_{m_j})] = 0$ et $-E[\partial^2 ln(L)/(\partial\sigma^2\partial\mu)] = 0$, le résultat de $-E[\partial^2 ln(L)/\partial(\sigma^2)^2]$ est superflu, ce qui simplifie le problème de l'inversion de matrice obligatoire. Comme prévu, ces valeurs donnent σ^2/n_o comme variance asymptotique pour $\bar{x}_o$. La variance asymptotique de l'estimation de y_{m_j}, lorsque celle-ci est considérée comme un paramètre, devient $n\sigma^2/[(n-1) - n_m/(n-1)]$. En revanche, lorsqu'on considère l'estimation de y_{m_j} comme la valeur prédictive d'une nouvelle observation, la variance est $\sigma^2(1 + 1/n_o)$. Dans ce cas, la statistique t nécessaire est $t_{.975,180} = 1.9732$. Par conséquent, les intervalles d'estimation sont à peu près les mêmes dans ce cas, soit environ 40521.8 ± 20536.1.

Si on choisit de faire une analyse fondée sur l'algorithme E-M classique, dans laquelle on suppose encore que les données codées suivant une grille géographique sont indépendantes et où les valeurs manquantes sont posées égales à $X_m b$, où les coefficients de régression reposent sur la solution itérative qui fait intervenir à la fois les valeurs connues et les valeurs manquantes, les équations nécessaires pour le calcul des erreurs types asymptotiques sont les suivantes:

$$-E[\partial^2 ln(L)/(\partial\beta\partial\beta)] = X^tX\sigma^2, \quad -E[\partial^2 ln(L)/(\partial\beta\partial Y_m)] = -X/\sigma^2, \text{ et } \quad -E[\partial^2 ln(L)/(\partial Y_m\partial Y_m)] = I_m/\sigma^2.$$

Cette fois-ci encore, le résultat de $-E[\partial^2 ln(L)/\partial(\sigma^2)^2]$ est sans importance, ce qui simplifie le problème de l'inversion de matrice obligatoire. La variance asymptotique de l'estimation de Y_m, lorsque celle-ci est considérée comme un paramètre, est définie par les éléments diagonaux 2 à $m+1$ de $\sigma^2[I_m - X_m(X^tX)^{-1}X_m^t]^{-1}$. Par contre, lorsque l'estimation de Y_m est considérée comme la valeur prédictive d'une nouvelle observation, la variance est $\sigma^2[I_m + X_m(X^tX)^{-1}X_m^t]$; l'expression entre crochets renferme les deux premiers éléments d'un développement de l'inversion de matrice mentionnée précédemment. Il ne faut pas s'étonner de ce que ces deux estimations puissent, elles aussi, être semblables, comme le montrent les résultats des tableaux 2 et 3. Dans ce cas, la statistique t nécessaire est $t_{.975,178} = 1.9734$.

Si on opte pour une analyse fondée sur une version spatiale de l'algorithme E-M, dans laquelle on suppose que les données codées suivant une grille géographique sont dépendantes, on peut poser les valeurs manquantes égales aux estimations calculées à l'aide de l'équation (4.2). Upton et Fingleton (1985, p. 353) ont déjà défini les formules nécessaires pour calculer les erreurs types asymptotiques dans le cas d'un modèle autorégressif spatial avec un ensemble complet de données codées suivant une grille géographique:

$$-E[\partial^2 ln(L)/\partial(\sigma^2)^2], \; -E[\partial^2 ln(L)/\partial\rho^2], \; -E[\partial^2 ln(L)/(\partial\beta\partial\beta)],$$
$$-E[\partial^2 ln(L)/(\partial\sigma^2\partial\beta)], \; -E[\partial^2 ln(L)/(\partial\sigma^2\partial\rho)], \text{ et } -E[\partial^2 ln(L)/(\partial\rho\partial\beta)].$$

Ces formules pourraient devoir être modifiées légèrement lorsque l'ensemble de données est incomplet. Voici les termes qu'il serait nécessaire d'ajouter:

$$-E[\partial^2 ln(L)/(\partial\beta\partial Y_m)] = [\rho W_{om}^t X_o - (I_m - \rho W_{mm})^t X_m]/\sigma^2,$$
$$-E[\partial^2 ln(L)/(\partial Y_m\partial Y_m)] = [\rho^2 W_{om}^t W_{om} + (I_m - \rho W_{mm})^t (I_m - \rho W_{mm})]/\sigma^2,$$
$$-E[\partial^2 ln(L)/(\partial\sigma^2\partial Y_m)] = \{\beta^t[X_m(I_m - \rho W_{mm}) - \rho X_o^t W_{om}] + \rho E(Y_o^t)[(I_o - \rho W_{oo})^t W_{om} + W_{mo}^t(I_m - \rho W_{mm})]$$
$$- [\rho^2 W_{om}^t W_{om} + (I_m - \rho W_{mm})^t (I_m - \rho W_{mm}]E(Y_m)\}/\sigma^4, \text{ et}$$
$$-E[\partial^2 ln(L)/(\partial\rho\partial Y_m)] = \{\beta^t(X_m^t W_{mm} + X_o^t W_{om}) + E(Y_o^t)[2\rho(W_{oo}^t W_{om} + W_{mo}^t W_{mm}) - (W_{mo}^t + W_{om})]$$
$$+ [2\rho(W_{mm}^t W_{mm} + W_{om}^t W_{om}) - (W_{mm} + W_{mm}^t)]E(Y_m)\}/\sigma^2.$$

où

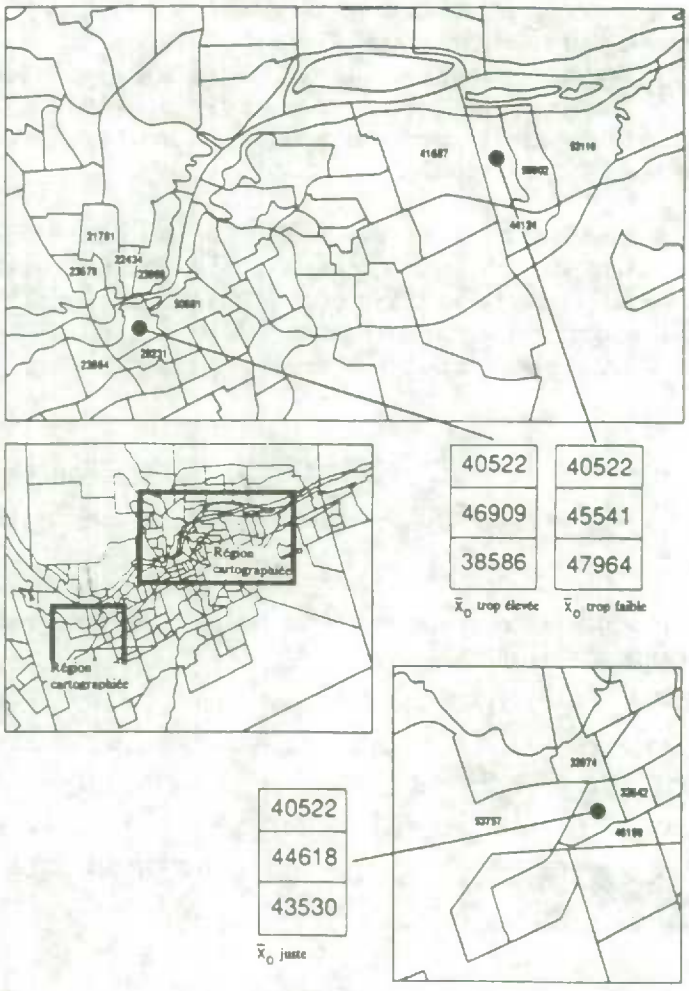
$$E(Y_o) = [(I_o - \rho W_{oo}) - \rho^2 W_{om}(I_m - \rho W_{mm})^{-1} W_{mo}]^{-1}[X_o + \rho W_{om}(I_m - \rho W_{mm})^{-1} X_m]\beta, \text{ et}$$
$$E(Y_m) = \{\rho(I_m - \rho W_{mm})^{-1} W_{mo}[I_o - \rho W_{oo}) - \rho^2 W_{om}(I_m - \rho W_{mm})^{-1} W_{mo}]^{-1} X_o + [(I_m - \rho W_{mm})$$
$$- \rho^2 W_{mo}(I_o - \rho W_{oo})^{-1} X_m\}\beta.$$

Dans ce cas, la statistique t nécessaire est $t_{.975,177} = 1.9735$.

La complexité numérique de ce dernier cas, qu'illustrent bien les résultats ci-dessus, est imputable à l'existence d'une autocorrélation spatiale non nulle et de valeurs non nulles pour $-E[\partial^2 ln(L)/(\partial\sigma^2\partial\rho)]$ et $-E[\partial^2 ln(L)/(\partial\sigma^2\partial Y_m)]$ -- l'inversion de matrice est plus complexe ici. Par conséquent, afin de pouvoir établir des comparaisons, nous nous limitons ici aux intervalles pour les valeurs prédites. La figure 2 contient une représentation graphique de ces comparaisons. Il convient de souligner que lorsque $\hat{y}_m = \bar{x}_o$, les intervalles de prédiction sont les mêmes pour toutes les valeurs manquantes; dans certains cas ils sont trop larges; dans d'autres, trop étroits. On introduit la variabilité, et par conséquent l'incertitude, en posant $\hat{Y}_m = X_m b$. Or, les intervalles sont plus étendus dans les cas où on s'éloigne des réponses moyennes de l'espace des attributs. En outre, les estimations proprement dites sont supérieures à $\bar{x}_o$ dans les onze cas d'absence de valeur. Enfin,

l'équation (4.2) réduit sensiblement la largeur des intervalles, ce qui donne des estimations de valeurs manquantes qui sont plus en rapport avec les valeurs des secteurs avoisinants.

Figure 2: Comparaison d'intervalles de prédiction obtenus à l'aide de trois méthodes d'estimation.

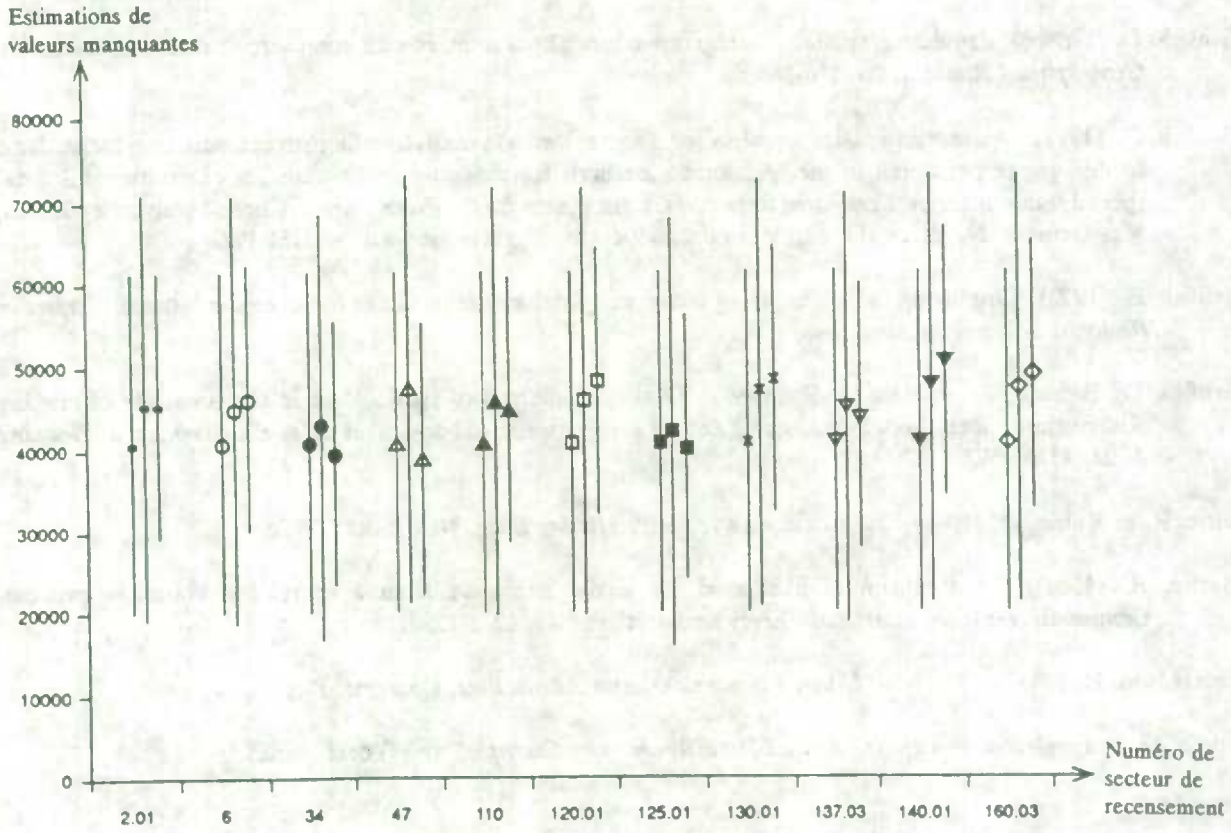


7. CONCLUSIONS

Cette étude a produit des résultats intéressants et nous a permis de tirer un certain nombre de conclusions. Premièrement, l'introduction de variables explicatives accroît sensiblement la précision des estimations de valeurs manquantes. L'étude qui avait été faite sur Houston avait permis d'observer des intervalles de confiance très étendus et même, dans un cas, exagérément larges. On avait utilisé à cette occasion un modèle du gradient, qui, selon toute vraisemblance, remplaçait les variables explicatives. Dans le cas d'Ottawa-Hull, une fois que la densité de population a été incluse dans la spécification du modèle, non seulement le diagnostic de régression a commencé à produire des résultats assez conformes à la norme, mais aussi les termes de coordonnées étaient devenus superflus. Incidemment, il y aurait moyen d'obtenir une mesure plus précise de la densité de la population en calculant le rapport de l'effectif de population à la superficie de la zone résidentielle plutôt que le rapport de l'effectif de population à la superficie totale du secteur de recensement. Deuxièmement, la spécification d'un modèle autorégressif spatial accroît davantage la précision des estimations de valeurs manquantes. La figure 3 met en relief cette affirmation. La première série d'estimations choisies illustre les cas où la moyenne estimée est très inférieure à la moyenne des revenus médians des secteurs de recensement voisins, est à peu près égale à la moyenne des revenus médians des secteurs voisins, et est supérieure à la moyenne des revenus médians des secteurs voisins. La deuxième série d'estimations provient d'une analyse E-M classique. Les estimations ne sont toujours pas conformes aux valeurs des secteurs avoisinants; dans le premier cas, où la moyenne estimée est trop élevée, la moyenne conditionnelle est encore plus élevée tandis que dans le troisième cas, elle reflète déjà plus la réalité. La troisième série d'estimations, tirées de l'équation (4.2), sont

plus en rapport avec les valeurs des secteurs avoisinants; dans le premier et le troisième cas, il s'agit de la meilleure estimation parmi les trois, tandis que dans le deuxième cas, elle se situe entre les deux autres estimations. Cette constatation vaut aussi pour le cas de Houston, où on a observé une autocorrélation spatiale positive beaucoup plus forte.

Figure 3 : Cartes de valeurs manquantes.



L'équation d'estimation (4.2) s'est révélée utile dans notre étude, surtout parce qu'elle accroît l'éventail d'estimateurs qui peuvent être utilisés lorsqu'il manque des données codées suivant une grille géographique. La méthode que décrivent Griffith, Bennett et Haining (1989) équivaut au krigeage, d'après la théorie des variables régionalisées (par exemple, le modèle exponentiel; voir Griffith 1991). Le krigeage s'appuie sur au moins quatre modèles différents (linéaire, exponentiel, gaussien, sphérique); certains conviennent mieux que d'autres pour définir un ensemble particulier de données. Nous ne voyons pas pourquoi il n'en serait pas de même pour les données spatiales! D'où la nécessité d'élaborer des cas particuliers de la solution générale de Martin (1984) et d'en étudier les propriétés. L'équation (4.2) est également précieuse en ceci qu'elle permet d'appliquer la technique au moyen de progiciels commerciaux courants comme SAS (voir annexes I et II) sans exiger une trop grande quantité de ressources informatiques. Au contraire, l'étude sur Houston avait malheureusement nécessité beaucoup de programmation en FORTRAN ainsi que des sous-programmes IMSL.

Enfin, la production d'estimations fiables pour des valeurs manquantes parmi des données urbaines du recensement ne met pas en péril la protection du secret statistique. Statistique Canada peut continuer de censurer les données confidentielles tout en offrant aux chercheurs des séries de données essentiellement complètes; cette méthode rappelle vaguement la méthode de l'arrondissement aléatoire, utilisée couramment pour préserver le caractère confidentiel des réponses contenues dans les questionnaires produits par les particuliers. Elle gagnerait énormément en utilité, toutefois, si Statistique Canada faisait à l'interne des analyses

confidentielles de la fiabilité des estimations de valeurs manquantes. De telles analyses, appliquées à un échantillon national plutôt qu'aux totalisations de certaines régions urbaines, fourniraient des renseignements utiles sans pour autant violer le caractère confidentiel des données. Les résultats obtenus pour Houston laissent supposer que des expériences semblables devraient être faites au U.S. Bureau of the Census.

BIBLIOGRAPHIE

Griffith, D. (1988a). *Advanced Spatial Statistics*, Boston: Kluwer.

Griffith, D. (1988b). Estimating spatial autoregressive model parameters with commercial statistical packages. *Geographical Analysis*, 20, 176-186.

Griffith, D. (1991). Advanced spatial statistics for geographic data analysis using supercomputing technology. Invited paper presented in the "Advanced methods for mapping and visualizing environmental data" special seminar series, *Ecosystem Research Center, Center for the Environment*, Cornell University, Ithaca, NY, October 24, under the auspices of a cooperative agreement with the USEPA.

Griffith, D. (1992). Simplifying the normalizing factor in spatial autoregressions for irregular lattices. *Papers in Regional Science*, 71, sous presse.

Griffith, D., Bennett, R. et Haining, R. (1989). Statistical analysis of spatial data in the presence of missing observations: a methodological guide and an application to urban census data. *Environment & Planning A*, 21, 1511-1523.

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

Martin, R. (1984). Exact maximum likelihood for incomplete data from a correlated Gaussian process. *Communications in Statistics: Theory and Methods*, 13, 1275-1288.

Richardson, H. (1977). *The New Urban Economics: And Alternatives*, London: Pion.

Upton, G. et Fingleton, B. (1985). *Spatial Data Analysis by Example*, New York: Wiley.

ANNEXE I. CODE SAS POUR L'ESTIMATION DE VALEURS MANQUANTES AU MOYEN D'UN MODÈLE DE RÉGRESSION LINÉAIRE

```
CMS FILEDEF ATTRIBUT DISK OTT-HULL DATA;                                    1
CMS FILEDEF CONN DISK OTT-HULL CONN;                                        2
DATA STEP1;                                                                 3
    INFILE ATTRIBUT;                                                        4
    INPUT NUM U V POP AREA INCOME;                                          5
U = U/10000;                                                                6
V = V/1000000;                                                              7
IF NUM < 5050500 THEN IND=1; ELSE IND=-1;                                   8
INCOME0 = INCOME;                                                           9
IF INCOME=0 THEN INCOME='.';                                                10
DEN = POP/AREA;                                                             11
RUN;                                                                        12
PROC REG SIMPLE DATA=STEP1;                                                 13
    MODEL INCOME = IND DEN/VIF CORRB COLLIN;                                14
    OUTPUT OUT=OUTREG1 P=YHAT R=YRESID;                                     15
RUN;                                                                        16
DATA STEP3;                                                                 17
    SET OUTREG1;                                                            18
    IF INCOME='.' THEN INCOME='0';                                          19
    IF INCOME=0 THEN INDMIS=1; ELSE INDMIS=0;                               20
INCOME = INCOME0 + INDMIS*YHAT;                                             21
DROP YHAT;                                                                  22
IF NUM=5050002.01 THEN IM1=1; ELSE IM1=0;                                   23
IF NUM=5050006.00 THEN IM2=1; ELSE IM2=0;                                   24
IF NUM=5050034.00 THEN IM3=1; ELSE IM3=0;                                   25
IF NUM=5050047.00 THEN IM4=1; ELSE IM4=0;                                   26
IF NUM=5050110.00 THEN IM5=1; ELSE IM5=0;                                   27
IF NUM=5050120.01 THEN IM6=1; ELSE IM6=0;                                   28
IF NUM=5050125.01 THEN IM7=1; ELSE IM7=0;                                   29
IF NUM=5050130.01 THEN IM8=1; ELSE IM8=0;                                   30
IF NUM=5050137.03 THEN IM9=1; ELSE IM9=0;                                   31
IF NUM=5050140.01 THEN IM10=1; ELSE IM10=0;                                 32
IF NUM=5050160.03 THEN IM11=1; ELSE IM11=0;                                 33
RUN;                                                                        34
PROC NLIN DATA=STEP3 RHO=0.1 METHOD=MARQUARDT;                              35
    PARM B0=0, B1=0, B2=0,                                                  36
    BM1=0, BM2=0, BM3=0, BM4=0, BM5=0, BM6=0,                               37
    BM7=0, BM8=0, BM9=0, BM10=0, BM11=0;                                    38
    MODEL INCOME = B0 + B1*IND + B2*DEN;                                    39
    INCOME = INCOME0 + BM1*IM1 + BM2*IM2 + BM3*IM3                          40
     + BM4*IM4 + BM5*IM5 + BM6*IM6 + BM7*IM7                                41
     + BM8*IM8 + BM9*IM9 + BM10*IM10 + BM11*IM11;                           42
    DER.B0 = 1;                                                             43
    DER.B1 = IND;                                                           44
    DER.B2 = DEN;                                                           45
    DER.BM1 = -IM1;                                                         46
    DER.BM2 = -IM2;                                                         47
    DER.BM3 = -IM3;                                                         48
    DER.BM4 = -IM4;                                                         49
    DER.BM5 = -IM5;                                                         50
    DER.BM6 = -IM6;                                                         51
    DER.BM7 = -IM7;                                                         52
```

```
    DER.BM8 = -IM8;                                                    53
    DER.BM9 = -IM9;                                                    54
    DER.BM10 = -IM10;                                                  55
    DER.BM11 = -IM11;                                                  56
    DER.B0.BM1 = 1 - IM1;                                              57
    DER.B1.BM1 = IND - IM1;                                            58
    DER.B2.BM1 = DEN - IM1;                                            59
    DER.B0.BM2 = 1 - IM2;                                              60
    DER.B1.BM2 = IND - IM2;                                            61
    DER.B2.BM2 = DEN - IM2;                                            62
    DER.B0.BM3 = 1 - IM3;                                              63
    DER.B1.BM3 = IND - IM3;                                            64
    DER.B2.BM3 = DEN - IM3;                                            65
    DER.B0.BM4 = 1 - IM4;                                              66
    DER.B1.BM4 = IND - IM4;                                            67
    DER.B2.BM4 = DEN - IM4;                                            68
    DER.B0.BM5 = 1 - IM5;                                              69
    DER.B1.BM5 = IND - IM5;                                            70
    DER.B2.BM5 = DEN - IM5;                                            71
    DER.B0.BM6 = 1 - IM6;                                              72
    DER.B1.BM6 = IND - IM6;                                            73
    DER.B2.BM6 = DEN - IM6;                                            74
    DER.B0.BM7 = 1 - IM7;                                              75
    DER.B1.BM7 = IND - IM7;                                            76
    DER.B2.BM7 = DEN - IM7;                                            77
    DER.B0.BM8 = 1 - IM8;                                              78
    DER.B1.BM8 = IND - IM8;                                            79
    DER.B2.BM8 = DEN - IM8;                                            80
    DER.B0.BM9 = 1 - IM9;                                              81
    DER.B1.BM9 = IND - IM9;                                            82
    DER.B2.BM9 = DEN - IM9;                                            83
    DER.B0.BM10 = 1 - IM10;                                            84
    DER.B1.BM10 = IND - IM10;                                          85
    DER.B2.BM10 = DEN - IM10;                                          86
    DER.B0.BM11 = 1 - IM11;                                            87
    DER.B1.BM11 = IND - IM11;                                          88
    DER.B2.BM11 = DEN - IM11;                                          89
    OUTPUT OUT=ITEROUT P=YHAT R=EHAT L95=LOW U95=UP;                   90
RUN;                                                                   91
PROC PRINT; VAR NUM INCOME0 LOW YHAT UP EHAT; RUN;                     92
ENDSAS;                                                                93
```

ANNEXE II. CODE SAS POUR L'ESTIMATION DE VALEURS MANQUANTES AU MOYEN D'UN MODÈLE D'AUTORÉGRESSION SPATIAL

```
CMS FILEDEF ATTRIBUT DISK OTT-HULL DATA;                               1
CMS FILEDEF CONN DISK OTT-HULL CONN;                                   2
DATA STEP1;                                                            3
    INFILE ATTRIBUT;                                                   4
    INPUT NUM U V POP AREA INCOME LAMBDAC LAMBDAW;                     5
IF NUM < 5050500 THEN IND=1; ELSE IND=-1;                              6
INCOME0 = INCOME;                                                      7
IF INCOME=0 THEN INDMIS=1; ELSE INDMIS=0;                              8
DEN = POP/AREA;                                                        9
```

```
DROP NUM U V POP AREA LAMBDAC LAMBDAW;                                  10
RUN;                                                                    11
DATA STEP2;                                                             12
    SET STEP1;                                                          13
    INFILE CONN;                                                        14
        INPUT CTN C1-C192;                                              15
    ARRAY CONN{192} C1-C192;                                            16
    ARRAY ICONN{192} IC1-IC192;                                         17
    RSUM = 0;                                                           18
    DO I=1 TO 192;                                                      19
        RSUM = RSUM + CONN{I};                                          20
        ICONN{I} = INCOME*CONN{I};                                      21
    END;                                                                22
RUN;                                                                    23
PROC MEANS DATA=STEP2 NOPRINT;                                          24
    VAR IC1-IC192;                                                      25
    OUTPUT OUT=ICOUT1 SUM=ICS1-ICS192;                                  26
RUN;                                                                    27
PROC TRANSPOSE DATA=ICOUT1 PREFIX=IW OUT=ICOUT2;                        28
    VAR ICS1-ICS192;                                                    29
RUN;                                                                    30
DATA STEP3;                                                             31
    SET ICOUT2;                                                         32
    SET STEP2;                                                          33
DROP IC1-IC192;                                                         34
IF CTN=2.01 THEN IM1=1; ELSE IM1=0;                                     35
IF CTN=6.00 THEN IM2=1; ELSE IM2=0;                                     36
IF CTN=34.00 THEN IM3=1; ELSE IM3=0;                                    37
IF CTN=47.00 THEN IM4=1; ELSE IM4=0;                                    38
IF CTN=110.00 THEN IM5=1; ELSE IM5=0;                                   39
IF CTN=120.01 THEN IM6=1; ELSE IM6=0;                                   40
IF CTN=125.01 THEN IM7=1; ELSE IM7=0;                                   41
IF CTN=130.01 THEN IM8=1; ELSE IM8=0;                                   42
IF CTN=137.03 THEN IM9=1; ELSE IM9=0;                                   43
IF CTN=140.01 THEN IM10=1; ELSE IM10=0;                                 44
IF CTN=160.03 THEN IM11=1; ELSE IM11=0;                                 45
DROP C1-C181;                                                           46
IW1 = IW1/RSUM;                                                         47
RUN;                                                                    48
*******************************************                             49
* DÉBUT DE L'ANALYSE DES DONNÉES SPATIALES *                            50
*******************************************;                            51
PROC NLIN METHOD=MARQUARDT;                                             52
    PARMS RHO=0.44900 B0=23451.43946 B1=3051.74277 B2=-0.94187;         53
    BOUNDS -1.601706 < RHO < 1.0;                                       54
    A1 = 0.228626;                                                      55
    A2 = 0.133974;                                                      56
    D1 = 1.859198;                                                      57
    D2 = 1.120741;                                                      58
    BM1=44392.6;                                                        59
    BM2=48817.9;                                                        60
    BM3=36794.2;                                                        61
    BM4=35006.3;                                                        62
    BM5=45050.0;                                                        63
    BM6=50668.6;                                                        64
    BM7=40427.1;                                                        65
```

```
    BM8=49171.1;                                                          66
    BM9=42260.0;                                                          67
    BM10=53205.4;                                                         68
    BM11=50336.2;                                                         69
    JHAT = EXP((192/181)*(A1*LOG(D1) + A2*LOG(D2) - A1*LOG(D1+RHO)        70
        - A2*LOG(D2-RHO)) + (LOG(1-RHO/24) + LOG(1+RHO/24))/181);         71
    TEMPINC = (INCOME0 + BM1*IM1 + BM2*IM2 + BM3*IM3 + BM4*IM4 +          72
          BM5*IM5 + BM6*IM6 + BM7*IM7 + BM8*IM8 + BM9*IM9 +               73
          BM10*IM10 + BM11*IM11)*JHAT;                                    74
    IW2 = IW1 + (BM1*C182 + BM2*C183 + BM3*C184 + BM4*C185 +              75
        BM5*C186 + BM6*C187 + BM7*C188 + BM8*C189 + BM9*C190 +            76
        BM10*C191 + BM11*C192)/RSUM;                                      77
    MODEL TEMPINC = (RHO*IW2 + B0 + B1*IND + B2*DEN)*JHAT;                78
    DER.B0 = JHAT;                                                        79
    DER.B1 = IND*JHAT;                                                    80
    DER.B2 = DEN*JHAT;                                                    81
    DER.RHO = ((RHO*IW2 + B0 + B1*IND + B2*DEN - TEMPINC/JHAT)*           82
          ((192/181)*(-A1/(D1 + RHO) + A2/(D2 - RHO)) +                   83
          (-1/(24-RHO) + 1/(24+RHO))/181) + IW2)*JHAT;                    84
    ID IW2 JHAT;                                                          85
    OUTPUT OUT=ITEROUT P=YHAT L95=LOW U95=UP R=EHAT;                      86
DATA STEP4 (REPLACE=YES);                                                 87
    SET ITEROUT;                                                          88
LOW = LOW/JHAT;                                                           89
YHAT = YHAT/JHAT;                                                         90
UP = UP/JHAT;                                                             91
PROC PRINT; VAR CTN INCOME0 IND DEN IW2 LOW YHAT UP; RUN;                 92
RUN;                                                                      93
ENDSAS;                                                                   94
```

ANNEXE III. CODE MINITAB POUR L'ÉQUATION (4.2)

```
NOECHO                                                                    1
READ 'PARM-EM DATA' C1                                                    2
PRINT C1                                                                  3
LET K1 = C1(1)                                                            4
LET K2 = C1(2)                                                            5
LET K3 = C1(3)                                                            6
LET K4 = C1(4)                                                            7
READ 'MIS-AR-W DATA' C1-C35                                               8
DIAG C33 M1                                                               9
COPY C1-C11 M2                                                            10
COPY C12-C22 M3                                                           11
MULT K1 M2 M2                                                             12
LET K5 = K1**2                                                            13
MULT K5 M3 M3                                                             14
SUB M2 M1 M2                                                              15
ADD M3 M2 M2                                                              16
INVERT M2 M2                                                              17
COPY C23-C25 M3                                                           18
COPY C26-C28 M4                                                           19
COPY C33-C35 M5                                                           20
LET C50(1) = K2                                                           21
LET C50(2) = K3                                                           22
LET C50(3) = K4                                                           23
```

```
MULT M3 C50 C46                                                    24
MULT M4 C50 C47                                                    25
MULT M5 C50 C48                                                    26
LET C53 = C48 - K1*(C46+C47) + K1*(C29+C32) - (K1**2)*(C30+C31)    27
MULT M2 C53 C54                                                    28
SET C55                                                            29
1:11                                                               30
END                                                                31
WRITE 'TEMP BM' C55 C54                                            32
END                                                                33
```

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

UTILISATION DE DONNÉES RÉGIONALES ET ADMINISTRATIVES POUR L'ÉTUDE DES EFFETS D'AGRÉGATION EN ANALYSE DÉMOGRAPHIQUE

C.G. Amrhein[1]

RÉSUMÉ

L'ensemble de données sur les déclarants de la Division des données régionales et administratives de Statistique Canada renferme des données ayant un niveau de décomposition géographique suffisamment grand pour que l'on puisse faire toutes les analyses nécessaires à la recherche d'effets d'agrégation. Dans une étude antérieure, on a examiné la nature de l'effet d'agrégation en se servant de données sur la migration produites à partir de données fiscales canadiennes. La présente étude pousse plus loin l'étude de Amrhein et Flowerdew (1990) en décomposant la table de flux aux 260 divisions de recensement selon l'âge et le sexe des migrants. Notre but n'est pas de définir un modèle de la migration au Canada mais de traiter des questions méthodologiques liées aux effets d'agrégation. Nous analysons trois algorithmes d'agrégation en utilisant, tour à tour, 130 et 65 régions et nous comparons les résultats de ces algorithmes aux données des 260 régions initiales et aux données provinciales.

MOTS CLÉS: Effets d'agrégation; algorithmes d'agrégation; migration.

1. INTRODUCTION

La nature et l'existence des effets d'agrégation font l'objet d'une recherche modeste mais soutenue depuis plusieurs années (voir Openshaw 1984 et Amrhein et Flowerdew 1990 pour des analyses). Dans une étude récente (Amrhein et Flowerdew 1990) où l'on utilisait des données sur la migration tirées d'un échantillon de déclarations d'impôt, on n'a pas prouvé l'existence d'un effet d'agrégation important. Un certain nombre de raisons pouvaient expliquer cette absence d'effet. Par exemple, comme Amrhein et Flowerdew avaient utilisé un modèle de Poisson, non linéaire, l'absence d'effet était peut-être attribuable à la différence entre le modèle non linéaire et les modèles linéaires utilisés dans les études antérieures. Cette absence était aussi peut-être attribuable à la nature linéaire du système urbain au Canada, selon lequel chaque province renferme une grande région urbaine (ou une grappe de régions de ce genre) de telle sorte que dans la mesure où on conserve les entités provinciales, le réseau de migration est stable. Enfin, il se peut que le caractère global des données de migration et le regroupement d'un grand nombre d'unités déclarantes (260 divisions de recensement) en un plus petit nombre d'unités (130, 65 ou 10) aient fait s'entremêler un certain nombre de processus de migration liés à l'échelle spatiale (intra-urbain, interurbain ou interrégional) et à la structure démographique (travail, retraite, etc.). La présente étude traite un aspect de cette dernière possibilité, soit l'agrégation selon l'âge et le sexe.

On reconnaît depuis longtemps l'importance de l'âge et du sexe comme variables causales en migration (voir, par exemple, Shaw 1975; Greenwood 1981; Mueller 1982; Greenwood 1985 et Clark 1986 pour des analyses) et on continue de s'intéresser à ces variables (voir, notamment, Ledent et Liaw 1989; Clark et White 1990; et Haurin et Haurin 1990). Dans cette étude, nous reconnaissons l'importance du processus de migration présumé décrit dans ces ouvrages en décomposant selon l'âge et le sexe la table de migration nationale construite à partir d'un échantillon de déclarations canadiennes de revenus de 1986. Flowerdew et Amrhein (1989) font une analyse détaillée de cet ensemble de données.

---

[1] C.G. Amrhein, Département de géographie, Université de Toronto, Toronto (Ontario), Canada M5S 1A1. National Center for Geographic Information and Analysis State University of New York at Buffalo, Amherst (New York), 14261, É.-U.

Notre étude vise principalement à rechercher un effet d'agrégation attribuable à trois facteurs. Les deux premiers facteurs, à savoir la division de l'espace, représentée par le nombre d'unités spatiales, et le groupement de régions, représenté par les règles d'agrégation, sont analysés dans Amrhein et Flowerdew (1990). Nous nous occupons de combiner l'effet du troisième facteur -- répartition de la population selon l'âge et le sexe -- à celui des deux autres. Le modèle utilisé a été choisi pour sa simplicité et sa parcimonie et ne se veut pas un modèle de la migration au Canada. Il existe des modèles beaucoup plus complexes (Shaw 1985 ou Flowerdew et Amrhein 1989) pour expliquer spécialement la migration. Nous nous limitons, ici, à traiter les questions méthodologiques liées aux effets d'agrégation.

2. MÉTHODOLOGIE

Les données sont décomposées selon l'âge et le sexe et selon la division de recensement. Par conséquent, l'âge et le sexe s'ajoutent au nombre d'unités spatiales (le nombre d'unités déclarantes dans l'analyse, c'est-à-dire 260, 130, 65 ou 10) et aux règles d'agrégation comme facteurs pouvant influencer la valeur des paramètres et des statistiques qui servent à analyser les tendances de la migration. Comme Amrhein et Flowerdew (1990) font une comparaison détaillée des différentes règles d'agrégation, nous ne ferons ici qu'un bref exposé des divers algorithmes. Il y a quatre algorithmes:

Algorithme de régression de Poisson:

Il s'agit d'un modèle de régression de Poisson conçu de manière à tenir compte des différences d'âge et de sexe. Il est exprimé sous la forme suivante:

$$\ln y_{ij}^{kl} = \beta_o + \beta_1 \ln P_i + \beta_2 \ln P_j + \beta_3 \ln d_{ij}, \qquad (1)$$

où P_i est la population du lieu d'origine, P_j est la population du lieu de destination, d_{ij} est la distance entre les divisions de recensement i et j, y_{ij}^{kl} est le nombre de personnes de sexe l qui appartiennent au groupe d'âge k, et ln est le logarithme naturel. À l'exclusion des territoires, on compte 260 régions qui correspondent aux divisions de recensement du Canada. La matrice de migration a été construite à partir d'un échantillon de déclarations de revenus pour 1986 (tirées de la base de données sur les déclarants de la Division des données régionales et administratives de Statistique Canada). Des cases de la matrice de flux de 1986 ont été groupées au besoin pour respecter les divisions de recensement de 1981.

Algorithme du voisin aléatoire (RanNei):

Cet algorithme est exprimé également sous la forme de l'équation (1) sauf qu'il s'applique à un plus petit nombre de régions, créées artificiellement. On crée ces régions synthétiques en choisissant tout d'abord au hasard une région de départ, puis en choisissant un nombre déterminé de régions adjacentes pour former une nouvelle région à partir des divisions de recensement originales. L'analyste détermine le nombre total de régions synthétiques qui doivent être créées (dans la présente étude, nous faisons des analyses en utilisant tour à tour 130, 65 et 10 régions, toutes créées à partir des 260 régions originales). Les problèmes d'îles auxquels donne lieu la définition de contiguïté sont résolus de façon heuristique.

Algorithme d'Openshaw modifié (ModOpen):

Comme le précédent, cet algorithme s'exprime par l'équation (1) et s'applique à un plus petit nombre de régions, créées artificiellement. Dans ce cas-ci, on crée les régions synthétiques en choisissant tout d'abord au hasard une région de départ, puis en choisissant un nombre aléatoire de régions adjacentes pour former une nouvelle région à partir des divisions de recensement originales. Le nombre de régions adjacentes pouvant servir à former une région synthétique varie de un à un maximum déterminé. Cet algorithme est une version modifiée de l'algorithme présenté dans Openshaw (1977).

Algorithme d'interaction maximum (MaxInt):

Cet algorithme est semblable à RanNei sauf que les régions adjacentes qui servent à former des régions synthétiques sont choisies selon des règles précises. En premier lieu, on choisit la région contiguë qui est le plus en interaction (au point de vue des migrants) avec la région de départ. On pourrait certes imaginer d'autres règles mais cela pourra être l'objet de recherches futures.

3. RÉSULTATS

Les algorithmes RanNei, ModOpen et MaxInt sont analysés tour à tour avec 130, 65 et 10 régions. Ces expériences sont désignées comme suit:

Modèle I: le modèle de régression de Poisson original avec ses 260 divisions de recensement.
Modèle II: le "modèle provincial" avec ses 10 agrégations correspondant à chacune des provinces.
Modèle III: RanNei avec 130 régions.
Modèle IV: MaxInt avec 130 régions.
Modèle V: ModOpen avec 130 régions.
Modèle VI: RanNei avec 65 régions.
Modèle VII: MaxInt avec 65 régions.
Modèle VIII: ModOpen avec 65 régions.

Cette analyse produit une diversité de résultats. Les six tableaux ci-dessous contiennent les résultats pour les modèles I et II ainsi que les résultats moyens obtenus à la suite de 100 simulations pour les modèles III à VIII. De plus, on analyse chacun des huit modèles pour cinq groupes d'âge (0-15, 16-24, 25-44, 45-64, 65 et plus) et pour les hommes et les femmes (soit en tout (2*5)[2 + (6 x 100)] = 6020 séries de résultats). Les groupes d'âge-sexe sont désignés de la façon suivante:

M1 et F1 désignent respectivement les hommes et les femmes de 0 à 15 ans.
M2 et F2 désignent respectivement les hommes et les femmes de 16 à 24 ans.
M3 et F3 désignent respectivement les hommes et les femmes de 25 à 44 ans.
M4 et F4 désignent respectivement les hommes et les femmes de 45 à 64 ans.
M5 et F5 désignent respectivement les hommes et les femmes de 65 ans et plus.

Les résultats ci-dessous portent sur deux statistiques sommaires et les quatre paramètres de l'équation (1). En ce qui concerne les deux statistiques sommaires et le coefficient de la distance, nous avons reproduit, pour des raisons de commodité, les résultats obtenus à l'aide de données agrégées dans Amrhein et Flowerdew (1990). L'analyse des résultats peut prendre au moins deux formes. La première est une comparaison préliminaire des résultats. On évalue le degré de disparité des résultats des divers modèles dans un cadre d'analyse sommaire défini par des anticipations analytiques et par une étude élémentaire de la signification des différences d'une certaine grandeur. Par ailleurs, compte tenu de la nature multidimensionnelle de l'étude [quatre algorithmes, nombre varié de régions (260, 130, 65, 10), deux sexes et cinq groupes d'âge], on peut aussi faire une analyse de la variance pour les résultats des modèles III à VIII. Nous présentons tout d'abord les tableaux des résultats des huit modèles, puis l'analyse préliminaire et enfin, les résultats de l'analyse de variance (obtenus à l'aide de la version 5.0 du SAS).

Tableau 1: Proportion de la variation expliquée

Groupe d'âge-sexe

Modèle	F1	M1	F2	M2	F3	M3	F4	M4	F5	M5	Total
I	.705	.705	.739	.727	.764	.766	.751	.747	.721	.705	.763
II	.821	.819	.863	.825	.794	.791	.737	.749	.703	.670	.820
III	.741	.708	.741	.730	.763	.764	.763	.759	.748	.703	.762
IV	.743	.712	.743	.732	.767	.767	.766	.762	.749	.711	.777
V	.727	.728	.759	.750	.780	.781	.778	.773	.763	.721	.761
VI	.688	.689	.726	.718	.744	.744	.752	.746	.756	.712	.744
VII	.689	.690	.727	.719	.745	.745	.753	.747	.757	.712	.762
VIII	.715	.716	.748	.742	.766	.767	.771	.766	.771	.729	.742

Tableau 2: R^2

Groupe d'âge-sexe

Modèle	F1	M1	F2	M2	F3	M3	F4	M4	F5	M5	Total
I	.377	.375	.516	.536	.451	.477	.445	.423	.472	.377	.456
II	.908	.907	.920	.874	.876	.879	.814	.825	.779	.726	.900
III	.642	.597	.692	.680	.663	.670	.675	.671	.668	.616	.675
IV	.700	.607	.701	.688	.674	.681	.686	.682	.670	.608	.641
V	.573	.573	.668	.660	.637	.647	.642	.637	.645	.589	.668
VI	.606	.606	.673	.659	.669	.669	.671	.668	.677	.621	.676
VII	.604	.604	.669	.653	.669	.668	.671	.666	.678	.624	.680
VIII	.612	.612	.687	.672	.674	.676	.679	.674	.695	.639	.673

Tableau 3: Constante

Groupe d'âge-sexe

Modèle	F1	M1	F2	M2	F3	M3	F4	M4	F5	M5
I	-10.7	-10.7	-12.2	-12.9	-12.5	-12.9	-12.7	-12.6	-12.4	-10.7
II	-16.5	-16.6	-17.9	-17.7	-16.7	-17.1	-19.7	-20.0	-21.9	-22.2
III	-10.9	- 9.6	-10.9	-11.6	-11.4	-11.9	-11.7	-11.6	-11.6	-10.7
IV	-10.9	- 9.6	-10.9	-11.6	-11.4	-11.9	-11.7	-11.6	-11.6	- 9.7
V	-10.6	-10.5	-11.8	-12.5	-12.0	-12.5	-12.5	-12.4	-12.5	-11.7
VI	- 9.3	- 9.3	-10.4	-11.1	-10.9	-11.4	-11.5	-11.4	-11.7	-10.8
VII	- 9.5	- 9.5	-10.6	-11.3	-11.0	-11.6	-11.6	-11.5	-11.7	-10.9
VIII	-10.0	-10.0	-11.1	-11.8	-11.4	-11.8	-12.1	-12.0	-12.3	-11.6

Tableau 4: Coefficient de la population du lieu d'origine

Groupe d'âge-sexe

Modèle	F1	M1	F2	M2	F3	M3	F4	M4	F5	M5
I	.747	.748	.748	.728	.864	.860	.864	.866	.766	.747
II	.921	.923	.952	.948	.952	.973	.969	.984	.905	.894
III	.697	.703	.697	.675	.817	.814	.836	.837	.744	.708
IV	.699	.706	.699	.676	.819	.816	.836	.838	.744	.705
V	.734	.735	.732	.714	.836	.833	.853	.854	.772	.741
VI	.661	.663	.657	.636	.768	.766	.808	.807	.733	.704
VII	.668	.670	.662	.641	.775	.773	.811	.811	.732	.703
VIII	.698	.699	.692	.677	.792	.789	.824	.824	.759	.731

Tableau 5: Coefficient de la population du lieu de destination

	Groupe d'âge-sexe									
Modèle	F1	M1	F2	M2	F3	M3	F4	M4	F5	M5
I	.764	.765	.864	.913	.821	.859	.761	.755	.808	.764
II	1.01	1.11	1.21	1.15	1.11	1.11	1.27	1.27	1.46	1.46
III	.844	.750	.844	.890	.805	.845	.747	.740	.805	.739
IV	.844	.751	.844	.889	.805	.846	.746	.740	.802	.750
V	.773	.774	.861	.905	.822	.858	.774	.767	.826	.768
VI	.748	.750	.838	.881	.798	.839	.747	.739	.820	.754
VII	.757	.758	.845	.889	.805	.847	.753	.745	.825	.759
VIII	.764	.765	.846	.888	.811	.846	.771	.764	.832	.776

Tableau 6: Coefficient de la distance

	Groupe d'âge-sexe										
Modèle	F1	M1	F2	M2	F3	M3	F4	M4	F5	M5	Total
I	-.914	-.913	-.903	-.841	-.894	-.880	-.977	-.982	-1.02	-.914	-.893
II	-.712	-.707	-.789	-.701	-.684	-.660	-.809	-.803	-.860	-.820	-.710
III	-.928	-.920	-.928	-.858	-.913	-.897	-1.02	-1.02	-1.07	-1.07	-.927
IV	-.928	-.921	-.928	-.859	-.916	-.899	-1.02	-1.02	-1.07	-.922	-.888
V	-.880	-.879	-.889	-.824	-.880	-.866	-.975	-.978	-1.03	-1.02	-.917
VI	-.845	-.844	-.875	-.810	-.850	-.834	-.957	-.958	-1.02	-1.02	-.869
VII	-.843	-.842	-.876	-.809	-.849	-.833	-.959	-.959	-1.03	-1.02	-.846
VIII	-.827	-.827	-.856	-.791	-.838	-.825	-.940	-.942	-.999	-.988	-.861

3.1 Analyse préliminaire

La principale conclusion que l'on peut tirer d'une analyse préliminaire des résultats des six tableaux est qu'il semble y avoir peu de différence entre les huit modèles (comparer les lignes dans les tableaux 1 à 6). Cette conclusion donne à penser que l'étude de la migration selon l'âge et le sexe au Canada n'a pas révélé de véritable effet d'agrégation, soit à cause des propriétés du modèle de régression de Poisson, des algorithmes ou de la structure du système urbain au Canada. Cette dernière constatation rejoint celles faites dans d'autres ouvrages (voir, par exemple, Simmons 1980). Néanmoins, la décomposition de la population selon l'âge et le sexe semble favoriser une plus grande variabilité entre les modèles comparativement à la population prise dans son ensemble. Si l'on considère le tableau 1 par exemple, l'efficacité générale du modèle de Poisson, qui est représentée par la proportion de la variation expliquée, varie effectivement plus au niveau des groupes d'âge-sexe qu'au niveau de la population en général. À cet égard, comparons plus particulièrement les lignes correspondant aux modèles VI et VII à celle qui correspond au modèle II -- deux premières colonnes et dernière colonne. Si l'on considère maintenant la statistique plus largement reconnue R^2 (coefficient de détermination), l'existence d'une plus grande variabilité au niveau des groupes d'âge-sexe ne fait aucun doute; on pourrait interpréter cette variabilité comme un effet d'agrégation signalé par la statistique. Néanmoins, il est nécessaire d'expliquer les différences de résultats observées entre les deux statistiques sommaires; elles pourraient être attribuables à la différence entre une statistique linéaire et une statistique non linéaire.

Les quatre paramètres des équations de régression de Poisson (tableaux 3 à 6) révèlent des tendances conformes à celles observées dans les études antérieures. Comparons par exemple, dans le tableau 6, les coefficients de la distance indiqués pour les divers groupes d'âge-sexe avec ceux figurant dans la dernière colonne pour chaque modèle. De façon générale, les résultats de chaque tableau portent à croire que le nombre de régions engendre une plus grande variabilité que l'algorithme d'agrégation. On peut s'en rendre compte en comparant dans chaque tableau les lignes qui correspondent aux modèles VI, VII et VIII (65 régions) à celles correspondant aux

modèles III, IV et V (130 régions), puis à celles correspondant aux modèles I (260 régions) et II (10 régions). Cette comparaison renforce l'idée d'effet d'agrégation (Openshaw 1984).

Comme nous venons de le voir, la comparaison des lignes de chaque tableau permet d'étudier l'effet conjugué de l'algorithme d'agrégation (RanNei, MaxInt et ModOpen) et du nombre d'agrégats dans chaque circonstance. Par contraste, la comparaison des colonnes de chaque tableau permet d'étudier l'effet de l'âge et du sexe sur la capacité du modèle de signaler un effet d'agrégation. Dans le tableau 1 par exemple, on peut voir l'effet du sexe et de l'âge par la variabilité des valeurs qui forment une ligne. Là encore, si on prend comme base de comparaison les résultats du modèle I pour l'ensemble de la population, la variabilité semble plus forte dans le cas des modèles à 65 régions (VI, VII, VIII) et du modèle provincial (modèle II) que dans le cas des modèles à 130 régions. On peut voir dans chaque ligne que les 0-15 ans dépendent, comme prévu, des 25-44 ans. De même, les valeurs observées pour les groupes d'âge moyen sont celles qui se rapprochent le plus des valeurs pour la population en général. Cela était aussi prévisible, compte tenu de la proportion des flux absolus dans la matrice de migration qui se rapportent à ce groupe d'âge. Toutefois, vu le très grand nombre d'effets (catégories), une analyse plus structurée s'impose; c'est pourquoi nous allons effectuer une analyse de variance avec SAS.

3.2 Analyse de variance

Nous faisons une analyse de variance pour isoler les effets de l'algorithme, de la taille du modèle, de l'âge et du sexe pour la sous-population analysée dans chacun des modèles. Il y a 100 observations pour chaque âge et chaque sexe pour les modèles III à VIII. Le modèle original à 260 régions (modèle I) et le modèle provincial (modèle II) ne font pas partie de cette analyse puisqu'il n'y a qu'une observation dans chaque cas. Les résultats présentés ci-dessous permettent donc d'analyser les différences entre les grappes de divisions de recensement générées aléatoirement à partir des 260 divisions originales et non de comparer les diverses données agrégées aux données originales. Néanmoins, l'observation d'un effet dû à la taille du modèle tendrait sûrement à confirmer l'existence d'un effet d'agrégation.

Le modèle de l'analyse de variance est construit avec trois algorithmes X deux tailles X deux sexes X 5 groupes d'âge, ou 3 X 2 X 2 X 5 = 60 cases. Chaque case contient 100 observations, de sorte que nous avons en tout 6 000 observations. Les résultats de l'analyse de variance figurent ci-dessous. Une des hypothèses requises pour cette analyse est que les statistiques sommaires et les valeurs des paramètres suivent une distribution normale dans la population. Les paramètres du modèle de régression de Poisson sont distribués approximativement selon une loi de chi carré avec un nombre de degrés de liberté égal à la différence entre le nombre d'observations et le nombre de paramètres dans le modèle. Comme la distribution chi carré tend vers la distribution normale lorsque le nombre d'observations augmente, nous supposons qu'avec 6 000 observations, l'hypothèse ci-dessus se vérifie à défaut de toute information visant à prouver le contraire.

Lorsqu'on analyse les résultats d'une analyse de variance, il est important de se rappeler qu'un écart observé qui n'est pas significatif avec un petit nombre d'observations peut le devenir si ce nombre atteint des proportions respectables. Ainsi, avec 6 000 observations, des écarts relativement faibles seront significatifs. De fait, comme nous le verrons plus bas, l'analyse de variance indique que presque toutes les variables sont significatives. Enfin, une analyse étendue de la série de données, que nous ne décrivons pas ici, a révélé l'existence de nombreux effets d'interaction significatifs entre les variables. Par conséquent, il faudra prendre bonne note de l'existence de ces effets comme des points mentionnés ci-dessus au moment de l'analyse des résultats.

Le tableau 7 contient les résultats du modèle d'analyse de variance qui vise à expliquer la proportion de la variation expliquée par l'équation (1). La variable dépendante est la proportion de la variation expliquée. Les variables indépendantes sont l'algorithme d'agrégation (MOD = RanNei, ModOpen, MaxInt), le sexe de la cohorte (SEX = M ou F), l'âge de la cohorte (AGE, où 1 = 0 à 15 ans, 2 = 16 à 24 ans, 3 = 25 à 44 ans, 4 = 45 à 64 ans et 5 = 65 ans et plus), et le nombre de régions utilisé (R130 = 130 régions et R65 = 65 régions).

Tableau 7:
Résultats de l'analyse de variance
Variable dépendante: DEV

Source	d.l.	Somme des carrés	Carré moyen	Valeur F	PR > F
Modèle	8	2.1983	0.3648	1345.85	0.00
Erreur	5 991	1.6239	0.0003	erreur-type	
Total corrigé	5 999	4.5422		(0.0165)	

Source	d.l.	Somme des carrés	Valeur F	PR > F
MOD	2	0.3821	704.80	0.00
SEX	1	0.2649	977.21	0.00
AGE	4	2.0329	1874.98	0.00
SIZE	1	0.2385	880.05	0.00

Compte tenu de la taille de l'échantillon (n = 6 000), il n'est pas étonnant que le modèle soit significatif et que chaque variable du modèle contribue de façon significative à expliquer la variation. Toutefois, les résultats de l'analyse de variance nous disent seulement qu'au moins une des soixante moyennes de case est différente des autres. Il est utile de savoir lesquelles de ces moyennes sont différentes; à cette fin, nous appliquons un test de comparaisons multiples pour chacune des variables. Le tableau 8 donne les résultats du test de Scheffé pour le modèle testé dans le tableau 7.

Tableau 8: Test de comparaisons multiples de Scheffé

Groupes de Scheffé	Moyenne	N	Modèle
A	0.7526	2 000	ModOpen
B	0.7368	2 000	MaxInt
C	0.7348	2 000	RanNei
A	0.7614	1 200	AGE 4
A	0.7610	1 200	3
B	0.7363	1 200	2
B	0.7361	1 200	5
C	0.7122	1 200	1
A	0.7477	3 000	R130
B	0.7351	3 000	R65
A	0.7481	3 000	F
B	0.7348	3 000	M

On constate avec intérêt que chacun des trois algorithmes d'agrégation est significatif et différent des deux autres, l'algorithme d'Openshaw modifié (ModOpen) étant celui qui a le plus grand pouvoir explicatif. Si on ne tient pas compte de l'observation selon laquelle .75 n'est pas vraiment différent de .73, on peut qualifier d'"effet d'agrégation" la différence significative de rendement entre les trois algorithmes. À tout le moins, cette différence laisse supposer que les modèles spatiaux sont sensibles au mode de division de l'espace. Cela n'est certes pas une constatation nouvelle, mais il est bon de la répéter parfois. Dans la même ligne de pensée, la différence attribuable au nombre de régions est aussi l'indice d'un effet d'agrégation. De même, chaque sexe et chaque groupe d'âge a un effet significatif et différent sur le pouvoir explicatif de l'équation (1). Il n'y a rien d'étonnant à ce que les adultes d'âge moyen et d'âge plus avancé (25 à 64 ans) soient le groupe qui explique le mieux les flux de migration puisque les données sont tirées de déclarations de revenus et que le groupe des 25-65 ans forme la masse des salariés. De plus, comme on peut le voir ci-dessous, les migrants en âge de travailler représentent la majeure partie du fichier de données.

Groupe d'âge	Hommes	Femmes	Total
Enfants (0 à 15)	164 817	156 537	322 354
Jeunes adultes (16 à 24)	120 661	128 392	249 053
Adultes d'âge moyen (25 à 44)	259 149	243 728	502 877
Adultes d'âge plus avancé (45 à 64)	63 937	64 069	128 006
Adultes retraités (65 +)	23 558	33 722	57 280
Total			1 259 57 0

Les tableaux 9 et 10 contiennent les résultats relatifs à la variable dépendante R^2 (coefficient de détermination). Ces résultats sont comparables à ceux des tableaux 7 et 8.

Tableau: 9
Résultats de l'analyse de variance
Variable dépendante: R^2

Source	d.l.	Somme des carrés	Carré moyen	Valeur F	PR > F
Modèle	8	3.7487	0.4686	116.89	0.00
Erreur	5 991	24.0165	0.0041	erreur-type	
Total corrigé	5 999	27.7652		0.0633	

Source	d.l.	Somme des carrés	Valeur F	PR > F
MOD	2	0.2770	34.54	0.0001
SEX	1	0.6022	150.22	0.0001
AGE	4	2.8642	178.62	0.0000
SIZE	1	0.0053	1.33	0.2489

Tableau 10: Test de comparaisons multiples de Scheffé

Groupes De Scheffé	Moyenne	N	Modèle
A	0.6603	2 000	MaxInt
A	0.6572	2 000	RanNei
B	0.6446	2 000	ModOpen
A	0.6752	1 200	AGE 2
A B	0.6686	1 200	4
B	0.6664	1 200	3
C	0.6441	1 200	5
D	0.6156	1 200	1
A	0.6549	3 000	R130
A	0.6531	3 000	R65
A	0.6640	3 000	F
B	0.6440	3 000	M

Les valeurs R^2 qui ont servi à l'analyse ont été calculées à l'aide des résultats de la régression de Poisson et non à partir des résultats d'une autre analyse de régression linéaire. Le premier point digne d'attention parmi ces résultats est le manque de signification de la variable SIZE. D'autres études (voir Openshaw 1984) laissent supposer qu'on peut obtenir diverses valeurs pour une statistique en modifiant les unités spatiales. Il ne faut donc pas s'étonner de ce que la variable R^2 ne soit pas significative étant donné la diversité des modèles qui sont testés ici. Le coefficient de détermination a une forte variabilité dans chacune des séries de 100 observations,

et il y a un intervalle de variation commun aux 60 séries. C'est ce que montre le tableau récapitulatif des tests de Scheffé. Les algorithmes RanNei et MaxInt donnent des résultats comparables, ce qui, encore une fois, est peu étonnant étant donné que les algorithmes se distinguent les uns des autres uniquement par la manière dont est choisie la région adjacente. Enfin, les groupes d'âge ne sont pas tous "significativement" différents les uns des autres en ce qui a trait au pouvoir explicatif. Cette constatation, de même que les autres tests du tableau 10, donne à penser que la statistique R^2 a un pouvoir de différenciation moins grand que celui de la statistique DEV. En ce qui regarde les groupes d'âge, il est surprenant de constater que le R^2 moyen le plus élevé correspond au groupe des jeunes adultes. Les deux groupes suivants sont respectivement les premier et deuxième dans le tableau 8 et ils ne sont pas significativement différents l'un de l'autre. Le groupe des jeunes adultes n'est pas non plus significativement différent des deux groupes d'adultes d'âge plus avancé. Enfin, comme dans le tableau 8, le groupe des adultes retraités et celui des enfants occupent respectivement le quatrième et le cinquième rang.

En dernier lieu, les tableaux 11 et 12 contiennent les résultats de l'analyse de variance et du test de Scheffé pour le coefficient de la distance de l'équation (1). Les résultats relatifs aux trois autres paramètres sont semblables et ne sont pas reproduits ici; on peut toutefois les obtenir en s'adressant à l'auteur.

Tableau 11:
Résultats de l'analyse de variance
Variable dépendante: COEF4
(coefficient de la distance)

Source	d.l.	Somme des carrés	Carré moyen	Valeur F	PR > F
Modèle	8	32.2980	4.0373	2655.64	0.00
Erreur	5 991	9.1079	0.0015	erreur-type	
Total corrigé	5 999	41.4059		0.0390	

Source	d.l.	Somme des carrés	Valeur F	PR > F
MOD	2	0.9422	309.90	0.00
SEX	1	0.7417	487.86	0.00
AGE	4	27.0859	4453.99	0.00
SIZE	1	3.5293	2321.49	0.00

Tableau 12: Test de comparaisons multiples de Scheffé

Groupe de Scheffé	Moyenne	N	Modèle
A	-0.9025	2 000	ModOpen
B	-0.9252	2 000	MaxInt
C	-0.9317	2 000	RanNei
A	-0.8586	1 200	AGE 2
B	-0.8665	1 200	3
C	-0.8736	1 200	1
D	-0.9786	1 200	4
E	-1.0217	1 200	5
A	-0.8955	3 000	R65
B	-0.9441	3 000	R130
A	-0.9087	3 000	M
B	-0.9309	3 000	F

Comme dans le cas des tableaux 7 et 8, les résultats des tableaux 11 et 12 nous disent que chacune des soixante moyennes de case est significative et que ces moyennes diffèrent l'une de l'autre pour les classes d'une même variable. Les valeurs observées sont négatives, le chiffre le plus élevé en valeur absolue étant celui qui correspond à la valeur moyenne la plus faible pour chaque modèle. Le paramètre de distance (COEF4) nous permet de faire des constatations intéressantes en ce qui a trait aux groupes d'âge. Rappelons-nous que les habitudes de mobilité des enfants dépendent entièrement de celles de leurs parents. Le groupe le plus sensible à la distance est celui des adultes retraités; le groupe le moins sensible est celui des 16-24 ans. Il est suivi du groupe des adultes d'âge moyen, puis des enfants. Les adultes d'âge plus avancé ne sont qu'un peu moins sensibles à la distance que les adultes retraités. Par ailleurs, avec deux fois plus de lieux d'origine et de lieux de destination que le modèle à 65 régions, le modèle à 130 régions comprend beaucoup plus de déplacements "locaux" et montre par le fait même que les migrants sont plus sensibles à la distance. Enfin, les femmes sont plus sensibles à la distance que les hommes.

Les résultats des tableaux 11 et 12 sont, parmi tous les résultats présentés jusqu'ici, ceux qui illustrent le mieux l'effet de l'agrégation d'unités spatiales sur les statistiques produites (ce qui témoigne peut-être de notre capacité de comprendre les sorties de modèles) et sur la description du comportement des sous-groupes d'une population. Qu'importe que les écarts soient suffisamment grands ou non pour que l'on puisse parler d'effet d'agrégation, les tableaux 11 et 12 donnent nettement à penser qu'il existe des écarts. Autrement dit, le niveau de décomposition géographique influe sur les résultats des modèles et cette influence varie selon les groupes d'âge-sexe. Bien que les conclusions de l'étude puissent être propres au Canada, la taille de l'échantillon est suffisamment grande pour que l'on puisse croire que les différences observées ne sont pas momentanées.

3.3 Groupes d'âge-sexe

Les tableaux 7 à 12 traitent séparément les effets de l'âge et du sexe. En combinant ces deux caractéristiques, nous serons en mesure de voir si les hommes et les femmes du même groupe d'âge ont des comportements différents. Les tableaux 13, 14 et 15 contiennent les résultats des tests de Scheffé pour les variables dépendantes DEV, R^2 et COEF4 respectivement. L'effet d'agrégation étudié ici est un effet d'agrégation démographique plutôt que géographique. En particulier, comparons les groupes de Scheffé formés dans le cas où l'âge et le sexe sont des variables distinctes (tableaux 8, 10, 12) aux groupes des tableaux ci-dessous. Les groupes d'âge-sexe sont définis comme auparavant.

Le tableau 13 donne à penser que les adultes d'âge moyen et les adultes d'âge plus avancé, hommes et femmes, sont les groupes qui expliquent le mieux la variation dans l'équation (1) alors que les enfants et les hommes retraités sont ceux qui l'expliquent le moins bien. Les groupes 3 et 4 sont très proches l'un de l'autre, comme dans le tableau 8. En ce qui a trait à R^2, le tableau 14 présente des résultats semblables à ceux du tableau 13, les trois groupes du bas étant les mêmes dans les deux cas. Ce qu'on note de particulier dans le tableau 14, c'est la grande capacité des femmes retraitées et des jeunes adultes d'expliquer les valeurs observées du coefficient de détermination, comparativement à la variable précédente. En revanche, on note un plus grand degré de chevauchement dans les groupes du tableau 14 que dans ceux du tableau 10. Ce phénomène n'a rien d'étonnant si l'on tient compte de l'effet d'interaction entre l'âge et le sexe et du degré de signification de ces variables, mentionné plus haut. Nous devons toutefois en déduire qu'il faut analyser ces résultats avec circonspection. Dans le tableau 13 par exemple, les cinq premiers groupes ont au moins une catégorie (F3) en commun. En conclusion, les effets de l'agrégation d'une population en fonction de l'âge et du sexe sont importants dans l'analyse des résultats d'un modèle.

Tableau 13: Test de comparaisons multiples de Scheffé pour la variable DEV lorsque les groupes d'âge-sexe représentent une seule variable

Groupe de Scheffé	Moyenne	N	Modèle
A	0.7640	600	F4
A B	0.7613	600	M3
A B C	0.7608	600	F3
C	0.7589	600	M4
C	0.7575	600	F5
D	0.7407	600	F3
E	0.7319	600	M2
F	0.7174	600	F1
F	0.7148	600	M5
G	0.7070	600	M1

Tableau 14: Test de comparaisons multiples de Scheffé pour la variable R^2 lorsque les groupes d'âge-sexe représentent une seule variable

Groupe de Scheffé	Moyenne	N	Modèle
A	0.6817	600	F2
A B	0.6719	600	F5
A B	0.6709	600	F4
A B	0.6687	600	M2
A B	0.6684	600	M3
B	0.6664	600	M4
B	0.6644	600	F3
C	0.6313	600	F1
D	0.6163	600	M5
E	0.6000	600	M1

Le tableau 15 contient les résultats du test de Scheffé pour le coefficient de la distance lorsque le groupe d'âge-sexe représente une seule variable. Une fois de plus, toutes les variables de la régression de Poisson utilisée dans l'analyse de variance sont significatives. Ces résultats se rapprochent sensiblement de ceux du tableau 12. Les adultes d'âge plus avancé et les adultes retraités, hommes et femmes, sont plus sensibles à la distance que les jeunes adultes et les adultes d'âge moyen. Les jeunes hommes et les hommes d'âge moyen constituent les deux groupes les moins sensibles à la distance. Fait intéressant, les enfants et les femmes d'âge moyen sont également sensibles à la distance mais ils le sont moins que les jeunes femmes. Le lien étroit qui existe entre les enfants et les femmes d'âge moyen est clair: les enfants se déplacent plus souvent avec la mère qu'avec le père. Cette constatation ne ressort pas dans le tableau 12 du fait que le groupe d'âge 3 comprend aussi bien des hommes que des femmes. La similitude des groupes M4 et F4 ne ressort pas non plus dans le tableau 12. Une question intéressante que l'on peut se poser est la suivante: pourquoi les hommes et les femmes retraités (M5 et F5) ont-ils des comportements différents lorsqu'ils sont considérés séparément? Cela s'explique peut-être par des différences de taux de survie entre les conjoints, c'est-à-dire que les femmes très âgées se comportent différemment et sont en plus grand nombre que les hommes très âgés.

Tableau 15: Test de comparaisons multiples de Scheffé pour la variable COEF4 lorsque les groupes d'âge-sexe représentent une seule variable

Groupes de Scheffé	Moyenne	N	Modèle
A	-0.8252	600	M2
B	-0.8588	600	M3
C	-0.8721	600	M1
C	-0.8742	600	F3
C	-0.8750	600	F1
D	-0.8920	600	F2
E	-0.9773	600	F4
E	-0.9799	600	M4
F	-0.0075	600	M5
G	-0.0360	600	F5

4. CONCLUSIONS

L'existence d'un effet d'agrégation dans la migration au Canada n'est pas encore parfaitement démontrée. Des analyses préliminaires permettent de croire qu'il existe de nettes différences entre les résultats relatifs à divers sous-ensembles de la population et à divers groupes de régions mais que ces différences ne sont pas appréciables en soi. Cette évaluation doit souligner du même coup les lacunes des données et les problèmes que soulève l'étude d'un phénomène de processus comme la migration. En même temps, les données constituent une base suffisamment large pour que les différences observées soient peu susceptibles de s'estomper avec de nouveaux échantillons. D'ailleurs, nous pouvons affirmer, données à l'appui, que les résultats du modèle de régression de Poisson sont sensibles à l'agrégation géographique comme à l'agrégation démographique. Ces résultats sont aussi sensibles à la manière dont sont générés les groupes de divisions de recensement originales. Ces observations donnent à penser que la généralisabilité des résultats de modèles d'interaction spatiale n'est peut-être pas aussi évidente qu'on peut le supposer. Il s'agit en fait de choisir le niveau de décomposition géographique qui convient au processus étudié. Les écarts attribuables à des niveaux d'agrégation différents ne seront peut-être pas appréciables en soi mais ils seront susceptibles d'être statistiquement significatifs. Dans chaque application, il est souhaitable de choisir le niveau de décomposition géographique approprié mais celui-ci ne correspond pas nécessairement toujours au niveau de décomposition le plus grand.

REMERCIEMENTS

L'auteur remercie le National Center for Geographic Information and Analysis pour son aide financière (subvention n° 150-6710A). Les données qui ont servi à cette étude ont été acquises grâce à une subvention de l'Institute for Market and Social Analysis, de Toronto (Ontario). L'utilisation de temps machine a été rendue possible grâce à une subvention à la recherche du Conseil de recherches de l'Université de Toronto et à une subvention versée par la province de l'Ontario au Ontario Centre for Large Scale Computation. Algorithm II a été programmé par Felipe Calderon, du département de géographie de l'Université de Toronto.

BIBLIOGRAPHIE

Amrhein, C., et Flowerdew, R. (1990). The effect of data aggregation on a Poisson regression model of Canadian migration. Département de Géographie, University of Toronto.

Clark, W.A.V. (1986). *Human Migration*, Beverly Hills, CA: Sage Publications.

Clark, W.A.V., et White, K. (1990). Modelling elderly mobility, *Environment and Planning A*, 22, 909-924.

Flowerdew, R., et Amrhein, C. (1989). Poisson regression models of Canadian census division migration flows, *Papers of the Regional Science Association*, 67, 89-102.

Greenwood, M. (1981). *Migration and Economic Growth in the United States*. New York: Academic Press.

Greenwood, M. (1985). Human migration: theory, models, and empirical studies, *Journal of the Regional Science Association*, 25, 521-544.

Haurin, D., et Haurin, R. (1990). Youth migration in the United States: an analysis of a deindustrializing region. Paper presented to the Annual Meeting of the Population Association of America, Toronto.

Ledent, J., et Liaw, K.-L. (1989). Provincial out-migration patterns of Canadian elderly: characterization and explanation, *Environment and Planning A*, 21, 1093-1112.

Mueller, C. (1982). *The Economics of Labor Migration*. New York: Academic Press.

Openshaw, S. (1977). Algorithm 3: a procedure to generate pseudo-random aggregations of N zones into M zones, where M is less than N, *Environment and Planning A*, 9, 1423-1428.

Openshaw, S. (1984). The modifiable areal unit problem CATMOG 38. Norwich, England: Geo Abstracts.

Shaw, R. (1975). Migration Theory and Fact. Philadelphia, PA: Regional Science Research Institute.

Shaw, R. (1985). *Intermetropolitan Migration in Canada: Changing Determinants Over Three Decades*. Toronto: NC Press.

Simmons, J. (1980). Changing migration patterns in Canada: 1966-1971 to 1971-1976, *The Canadian Journal of Regional Science* 3, 139-162.

CONFÉRENCIER SPÉCIAL INVITÉ

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

POSSIBILITÉS D'APPLICATION DES MODÈLES SPATIAUX DANS L'ESTIMATION DES ERREURS NON DUES À L'ÉCHANTILLONNAGE

P.P. Biemer[1]

Les méthodes classiques d'estimation des composantes de l'erreur quadratique moyenne totale -- par exemple, les mesures répétées, la vérification d'enregistrements et le recoupement de tâches (d'intervieweurs) -- sont des opérations coûteuses qui donnent lieu immanquablement à des problèmes de collecte de données. En outre, les hypothèses des modèles sur lesquelles reposent les estimations se vérifient rarement dans la pratique de sorte qu'en définitive, les estimations ne sont guère plus que des indicateurs approximatifs des composantes d'erreur. Pour ne pas avoir à estimer directement les composantes de l'EQM totale par des expériences spéciales sur le terrain, qui sont le plus souvent coûteuses et complexes, des chercheurs ont tenté de *modéliser* la variation expérimentale non contrôlée au lieu de recourir à la randomisation ou à des mesures répétées (qu'on appelle méthodes non expérimentales). À cette fin, ils ont eu recours à deux techniques particulières : l'estimation par régression et l'estimation de variable instrumentale. Le modélisateur peut désormais obtenir de meilleurs estimateurs de l'erreur non due à l'échantillonnage avec les modèles spatiaux qu'avec d'autres modèles. Après un rappel des ouvrages portant sur les méthodes d'estimation non expérimentales, nous analysons le principe des modèles spatiaux dans le contexte d'un recensement de la population.

Le principe fondamental des modèles spatiaux est le suivant. Soient y_{ij} la réponse au recensement fournie par l'unité j dans la tâche du recenseur i et τ_{ij} la vraie valeur se rattachant à l'unité en question. Supposons

$$y_{ij} = \tau_{ij} + b_i + e_{ij},$$

où $b_i \sim (0, \sigma_b^2)$ et $e_{ij} \sim (0, \sigma_e^2)$. Notre objectif est d'estimer σ_b^2 et σ_e^2 sans qu'il y ait "recoupement" des tâches des recenseurs, c.-à-d. en appliquant la méthode utilisée habituellement dans les recensements pour répartir les unités entre les recenseurs. À cette fin, nous redéfinissons la composante $\tau = [\tau_{ij}]$ sous la forme $X\beta$, où X est une matrice de constantes connues et β, un vecteur de paramètres à estimer. En outre, nous décrivons l'erreur du modèle $y - X\beta = \mu$ à l'aide d'un modèle spatial qui tire profit de la relation de dépendance spatiale qui existe entre les tâches de recenseurs de secteurs contigus. Nous abordons aussi dans l'article des questions relatives à l'estimabilité et à l'estimation. Enfin, nous décrivons les résultats d'une petite étude de simulation faite par le U.S. Bureau of the Census.

BIBLIOGRAPHIE

Biemer, P.P. (1986). A spatial modeling approach to the evaluation of Census nonsampling error, *Proceedings of the Census Annual Research Conference*, Washington (DC), 7-20.

[1] P.P. Biemer, Research Triangle Institute, C.P. 12194, Research Triangle Park (NC) 27709, É.-U.

ALLOCUTION DE CLÔTURE

Recueil du Symposium 91 de Statistique Canada
Questions spatiales liées aux statistiques
Novembre 1991

ALLOCUTION DE CLÔTURE

G.J. Brackstone[1]

Cela nous amène à la fin de notre Symposium '91. Je désirerais dire quelques mots à titre de conclusion. Pour ceux d'entre vous qui ne le savent pas, cela fait maintenant plusieurs années que nous, de Statistique Canada, organisons un symposium en méthodologie et quand le sujet du symposium de cette année a été proposé pour la première fois, certains d'entre nous se demandaient s'il s'agissait d'un sujet approprié pour un symposium en méthodologie. Auparavant, nous avions traité d'aspects plus traditionnels en méthodologie d'enquête et ce sujet était un peu hors des sentiers battus pour nous.

Je suis heureux de dire et d'admettre que nos inquiétudes n'étaient pas justifiées. Si l'on en juge par le nombre d'inscriptions à cette conférence et par la participation qu'elle a suscitée, cela a certainement été un symposium couronné de succès de notre point de vue et j'espère que vous partagez cette opinion.

En fait, plus de 400 personnes se sont inscrites à la conférence. Bien entendu, cela comprend beaucoup de gens de Statistique Canada, les employés de Statistique Canada représentent, effectivement, la majorité des participants. Mais 135 personnes de l'extérieur de Statistique Canada se sont inscrites et, en particulier, un bon nombre d'entre elles venaient des États-Unis et certaines d'outre-mer. Et, de ce point de vue seulement, le symposium a été un succès.

Je pense aussi que la gamme des sujets qui ont été traités est très impressionnante. Nous avons commencé mardi matin avec l'excellent exposé de Michael Goodchild, qui a donné le ton pour le symposium et qui a exposé systématiquement certains des problèmes et des défis. En un sens, nous sommes passés de cet exposé effectué par un géographe et qui présentait certains problèmes statistiques à l'exposé final de Paul Biemer, présenté par un statisticien, un méthodologiste d'enquête, qui étudie les utilisations possibles de méthodes géographiques dans un problème relatif à la méthodologie d'enquête. Entre ces deux limites, nous avons traité de la majorité des fonctions d'un bureau statistique qui entrent en jeu dans une enquête: la collecte, la constitution de la base de sondage, le traitement, l'estimation de l'erreur non due à l'échantillonnage et nous avons abordé une gamme étendue de domaines d'application: statistiques médicales, agriculture, environnement, pour n'en nommer que quelques-uns.

Je ne tenterai même pas de résumer ce que la conférence nous a permis d'apprendre. C'est à vous tous qu'il revient de résumer ce que vous avez appris et d'en tirer profit. Mais je pense qu'il serait approprié de faire quelques observations à propos de la géographie dans un organisme statistique. Le rôle de la géographie est bien établi dans le contexte d'un recensement de la population et il est bien reconnu comme un élément essentiel lors de la réalisation d'un recensement. Pour ce qui est d'autres activités d'enquête, son importance a été, traditionnellement, moins bien établie. À Statistique Canada, nous avons fait certains efforts, tout comme la Division de la géographie, en vue de démontrer l'importance et la valeur d'outils et de concepts géographiques pour la réalisation d'enquêtes autres que le recensement.

Selon moi, c'est, bien entendu, dans le domaine de l'analyse et, en particulier, de l'analyse géographique ou spatiale de données recueillies par des organismes statistiques qu'il reste encore le plus de progrès à réaliser. Un bon nombre des communications présentées lors des trois derniers jours ont porté sur ces aspects, et ont signalé certains des problèmes qui existent en matière d'analyse spatiale. J'espère qu'elles ont au moins soulevé un certain intérêt parmi les participants, tant de Statistique Canada que de l'extérieur, et qu'elles ont peut-être su éveiller l'intérêt pour des sujets qui pourraient faire l'objet d'études plus poussées après ce symposium.

---

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

Nous avons beaucoup entendu parlé de la qualité des données. Ce sujet était, en fait, le thème du symposium de l'année dernière. La qualité des données géographiques est un élément dont nous devons tenir compte quand nous utilisons ces données. On nous a beaucoup parlé des problèmes, à commencer par les erreurs écologiques, et diverses manifestations de ces erreurs ont été mentionnées au cours des derniers jours. Nous avons aussi beaucoup entendu parler de l'utilisation des données géographiques pour des programmes - l'utilisation de données afin de déterminer les avantages que des régions particulières tireront.

Nous avons certainement eu notre expérience avec ce problème à Statistique Canada. Je pense qu'on y a très bien fait allusion dans le cas du traitement fiscal dans les régions du Nord et je suis heureux d'entendre qu'on a trouvé une nouvelle façon de traiter ce problème parce que l'ancienne méthode nous créait des difficultés. Dans d'autres exemples, certaines régions ne voulaient pas être classées comme urbaines afin que les personnes qui y habitaient puissent avoir droit à des avantages. Parfois, nous avons d'autres régions du pays qui veulent être considérées comme urbaines parce que, dans leur cas, les avantages vont dans l'autre sens. Et nous avons aussi des cas où des représentants de régions locales ne veulent pas que nous fusionnions leurs régions avec d'autres régions ou veulent que nous les fusionnions avec d'autres régions, afin que les personnes qui y résident puissent profiter de programmes particuliers. Ce sont des exemples des problèmes que nous rencontrons avec l'utilisation de données géographiques pour des programmes.

Je désirerais, bien entendu, dire quelques mots pour remercier les personnes qui étaient responsables d'organiser cette rencontre. Tout d'abord, Mary March, qui a été la coordonnatrice globale de ce travail. Elle a été appuyée par une équipe organisatrice composée de L. Chatterton, C. Weiss, J. Yan et P. Tallon.

Bien entendu, de nombreuses autres personnes ont aidé l'équipe en coulisses pour l'élaboration du programme, pour la publicité, pour l'impression, pour l'inscription, pour la restauration et pour l'accueil. Je désirerais remercier les personnes responsables de la préparation de cette pièce, Gilbert Gauthier qui s'est occupé des effets audiovisuels afin de faire en sorte que la bonne diapositive soit montrée avec le bon orateur. Je désirerais aussi remercier nos interprètes qui ont travaillé ici pendant trois journées complètes à traduire tout ce qui a été dit dans l'autre langue officielle. Les promoteurs ont été mentionnés au début de cette conférence: Statistique Canada et le Laboratoire de recherche en statistique et probabilité, Carleton University et l'Université d'Ottawa et l'Association canadienne des géographes. Finalement, je dois remercier toutes les personnes qui ont présenté des communications et les présidents et présidentes dont la contribution a servi à faire de ce symposium un succès.

Comme nous l'avons fait pour les symposiums précédents, nous publierons les Actes du présent symposium. Nous poursuivrons notre série de symposiums l'année prochaine alors que nous quitterons la dimension spatiale. Nous retournerons à la dimension temporelle parce que le symposium de l'année prochaine portera sur la conception et l'analyse des enquêtes longitudinales. Il s'agit d'un domaine que nous, de Statistique Canada, n'avons pas beaucoup étudié jusqu'ici. Mais nous commençons à y travailler de plus en plus et nous comptons tirer profit de l'expérience de personnes qui ont participé à des enquêtes longitudinales dans d'autres organismes.

Le Symposium '91 est terminé. Je vous remercie tous pour votre participation et votre appui.