November 1991

SYMPOSIUM 91

Spatial Issues in Statistics

PROCEEDINGS

Statistics Canada Statistique Canada

Canada

SYMPOSIUM 91

Spatial Issues in Statistics

November 12-14, 1991

Ottawa, Ontario, Canada

PROCEEDINGS

July 1992

Symposium 91 Organizing Committee

Mary March
Caroline Weiss
Liane Chatterton
Joel Yan

Published by authority of the Minister responsible for Statistics Canada

PREFACE

Symposium 91 was the eighth in the series of international symposia held annually at Statistics Canada since 1984. Each year the symposium focuses on a particular theme. The 1991 theme was Spatial issues in statistics.

The 1991 symposium was attended by over 400 persons from several countries who met over three days in the Simon Goldberg Conference Centre in Ottawa to hear experts from numerous government agencies, universities, and private industry. During the symposium, participants heard 28 invited papers presented in 10 different sessions which addressed a wide spectrum of issues including: Geographic Perspectives on Data Modelling; Spatial Considerations in the Design of Surveys or Frames; Spatial Analysis of Health and Environmental Data; Spatial Developments in Data Processing; Geographic Innovations in Data Collection; Spatial Data Quality; Medical Geography; Spatial Analysis of Survey Data; Geographic Frameworks for Statistical Data; Data Analysis from a Geographic Perspective.

Aside from translation and formatting, the Proceedings of Symposium 91 contain a record of the papers as submitted by the authors. The order of presentation of the papers is the same as that of the symposium itself.

The Symposium 91 Committee would like to acknowledge the contributions of the many persons involved in the preparation of these Proceedings.

Naturally, thanks go to the presenters at Symposium 91 who took the time to put their ideas into words and submit them for inclusion in this volume. The efforts of many others were vital in the publication of these Proceedings. Processing of the manuscript was expertly handled by Christine Larabie, Carmen Lacroix and Judy Clarke. Proofreading was done by numerous methodologists and subject matter specialists: 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 and P. Whitridge. Christine Larabie co-ordinated the translation, proofreading and production of the manuscript. Neil Pecore assisted in co-ordination of the manuscript preparation.

Statistics Canada's ninth annual international symposium will be held November 2-4, 1992 in Ottawa. The topic will be "Design and Analysis of Longitudinal Surveys".

Symposium 91 Committee

July 1992

STATISTICS CANADA SYMPOSIUM SERIES

1984 - Analysis of Survey Data
1985 - Small Area Statistics
1986 - Missing Data in Surveys
1987 - Statistical Uses of Administrative Data
1988 - The Impact of High Technology on Survey Taking
1989 - Analysis of Data in Time
1990 - Measurement and Improvement of Data Quality
1991 - Spatial Issues in Statistics

STATISTICS CANADA INTERNATIONAL SYMPOSIUM SERIES
PROCEEDINGS ORDERING INFORMATION

To order additional copies of the proceedings of Symposium 91: Spatial Issues in Statistics, use the order form on this page. A limited number of copies of back issues from the 1987, 1988, 1989, and 1990 symposia are also available. To order, complete this form and send it (or a copy) to:

SYMPOSIUM 91 PROCEEDINGS
STATISTICS CANADA
FINANCIAL OPERATIONS DIVISION
R.H. COATS BUILDING, 6th FLOOR
TUNNEY'S PASTURE
OTTAWA, ONTARIO
K1A 0T6
CANADA

Please include payment with your order (cheque or money order, in Canadian funds or equivalent, payable to "The Receiver General for Canada - Symposium 91 Proceedings").

SYMPOSIUM PROCEEDINGS: ISSUES AVAILABLE

1987 -	Statistical Uses of Administrative Data - ENGLISH	____ @ $10 EACH
1987 -	Statistical Uses of Administrative Data - FRENCH	____ @ $10 EACH
1987 -	SET OF 1 ENGLISH AND 1 FRENCH	____ @ $12 PER SET
1988 -	The Impact of High Technology on Survey Taking - BILINGUAL	____ @ $10 EACH
1989 -	Analysis of Data in Time - BILINGUAL	____ @ $20 EACH
1990 -	Measurement and Improvement of Data Quality - ENGLISH	____ @ $15 EACH
1990 -	Measurement and Improvement of Data Quality - FRENCH	____ @ $15 EACH
1990 -	SET OF 1 ENGLISH AND 1 FRENCH	____ @ $25 PER SET
1991 -	Spatial Issues in Statistics - ENGLISH	____ @ $20 EACH
1991 -	Spatial Issues in Statistics - FRENCH	____ @ $20 EACH
1991 -	SET OF 1 ENGLISH AND 1 FRENCH	____ @ $35 PER SET

PLEASE ADD $2 PER VOLUME FOR SHIPPING $ ____

TOTAL AMOUNT OF ORDER $ ____
(Prices include GST; GST registration no.: R121491807)

PLEASE INCLUDE YOUR FULL MAILING ADDRESS WITH YOUR ORDER

NAME _____

ADDRESS _____

_____

CITY _____ PROV/STATE _____ COUNTRY _____

POSTAL CODE _____ TELEPHONE (____)_____ FAX _____

Please note: Each Symposium 91 registrant not employed by Statistics Canada receives one free copy of the Symposium 91 Proceedings.

SPATIAL ISSUES IN STATISTICS

TABLE OF CONTENTS[1]

[1] In cases of joint authorship, the name of the presenter is shown boldface.

NOTE: The original language of all Symposium 91 papers was English except for the following: Opening Remarks, I.P. Fellegi (bilingual); Session 4, J. Muzard (French), P. Daoust (French); Session 6, P. Ardilly (French).

OPENING REMARKS

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

OPENING REMARKS

I.P. Fellegi[1]

Welcome to Symposium '91. A methodology symposium has become an annual event at Statistics Canada. Since 1984, we have organized eight of these symposia - each one focusing on a different methodology topic. Topics covered in previous symposia were:

Analysis of Survey Data;
Small Area Statistics;
Missing Data in Surveys;
Statistical Uses of Administrative Data;
The Impact of High Technology on Survey Taking;
Analysis of Data in Time; and
Measurement and Improvement of Data Quality.

The topic chosen for this year's symposium is "Spatial Issues in Statistics". In choosing this topic, we were motivated by three considerations.

First, the needs of our clients. In a country with as much regional diversity as Canada, we always had to provide significant geographic breakdowns of our output. But there are now a whole range of types of uses for which an intensifier capability of spatial analysis is a prerequisite. Clearly, all environmental issues fall into this category: it is difficult even to talk about environmental statistics except within a spatial orientation. Environment related health risk analysis is another example. Indeed, health atlases represent a significant advance in exploratory data analysis. Modern marketing uses geographically referenced statistical information with ever increasing flexibility. And so on.

The second reason relates to technology. The last decade has seen extensive development in the area of Geographic Information Systems (or G.I.S.). Computer-supported systems are becoming increasingly powerful and, at the same time, more available. Large amounts of spatial data can now be captured, stored, edited, transformed, manipulated, searched, retrieved, analyzed and displayed. Also, the quality of data available through these systems has the potential to improve - both because of better technologies used in collecting data and because of the systems' editing capabilities. Most important, an increasing number of our clients have direct or indirect access to G.I.S.

The third reason that "Spatial Issues in Statistics" could be considered a timely topic is methodological. There is a burgeoning interest among researchers and practitioners, including those who work at Statistics Canada, in the domain of Spatial Statistics. This involves a special class of methods for handling and analyzing spatially distributed data - considering issues such as spatial auto-correlation and using concepts such as "spatial diffusion modelling", "fractals" and "directional statistics". With the coming of age of the G.I.S., there is a rediscovery of Spatial Statistics and with improved technology and the expected availability of expert systems, it is much easier to apply these sophisticated techniques.

The spatial dimension is increasingly an integral part of our work at Statistics Canada. In order to collect the information efficiently, we can use spatial information in planning and controlling collection operations. In order to be useful, the information we provide must be available for a variety of geographic distributions, with

---

[1] I.P. Fellegi, Chief Statistician, Statistics Canada, Tunney's Pasture, 26-A, R.H. Coats Building, Ottawa, Ontario, Canada K1A 0T6.

flexibility, good timeliness and in a user friendly mode of display. In order to improve the accuracy of our results, spatial data can be used in improving estimates and analyses.

These developments and preoccupations are typified by the Census, our largest undertaking. Planning and organization of the collection process alone - to ensure complete and efficient coverage of the population - requires the storage and use of a vast amount of spatial data: to delineate enumeration areas, to prepare maps, to calculate costs, to establish differential piece rates for our enumerators as a function of the type of area they have to enumerate, and to generate enumeration quality control tools such as the Address Register, used by the Census of Population, and the Township Plans, used by the Census of Agriculture. Over the years, processing of the collected data (including data entry, editing, imputation and weighting) has also made use of spatial information. Here, spatial information is used as part of the quality control of these processes and for sub-dividing the population into groups for weighting and imputation. Census data are subject to another geographic enhancement that greatly enhances its value. The data are linked with geographic location identifiers at fairly low levels through a process known as geo-coding which attaches Universal Transverse Mercator coordinates to enumeration area and block face centroid. Once processing is completed, the Census data are widely exploited, and this is in no small part due to the richness added by the geographic information that accompanies it. We have also developed the capacity to display Census data in many useful ways - as is exemplified in the Metropolitan Atlas Series which has proved to be a popular is offering in the Census line of products. A new product, that is currently pilot tested using 1986 Census data will be available as part of the 1991 Census Output: Census data, together with a digitized map of Canada broken down into very small building blocks, and linked by a sophisticated Geographic Information System.

In the past, our specialists in geographic information systems and geography, have mostly concentrated on developments to support the census. Their achievements have been numerous. Not the least impressive of them has been the ability to automatically delineate boundaries of more than 40% of EAs in large urban centres (using a technology called Computer-Assisted Distracting) and to produce high-quality computer-generated maps (using Computer-Assisted Mapping) of more than 50% of the enumeration areas for the 1991 Census.

In recent years, the spatial expertise that we have developed has begun to be applied more widely in the Bureau. Data from different sources are being linked through the use of geographic information such as Postal Codes, resulting significantly in enhanced data sets. Information collected by surveys or by other means is being displayed spatially. A prime example is our Environmental Information System containing a rich array of data from social, economic, and demographic surveys, as well as physical measures and geographic features. We have also developed the capacity to exploit data collected using remote-sensing technology - primarily data on agricultural land use - and use it to augment information we have traditionally obtained from surveys and administrative data.

We are also starting to use ideas from spatial statistics in analyzing data - particularly in areas such as health and environment. This has been partly driven by the fact that these issues are highly related to location. But technological and methodological advances are strong facilitating forces.

A number of spatial issues are to be discussed in the sessions that have been planned. Clearly, there is a lot of interest in the interface of spatial issues and statistics. The large number of people who have shown interest in this meeting and who have chosen to attend it is perhaps the ultimate "market test" that this topic has, indeed, come of age. We are pleased to welcome participants from Canada as well as visitors from France and the United States. There are representatives from government, including our provincial focal points, as well as from industry and the academic world. Hopefully, this will lead to a particularly useful cross-fertilization of ideas.

On behalf of all the organizers of the Symposium, I want to extend special thanks to Carleton University and the University of Ottawa Laboratory for Research in Statistics and Probability for once again co-sponsoring this event. I also want to welcome a new co-sponsor this year - the Canadian Association of Geographers - whose enthusiastic support is greatly appreciated. We thank all our co-sponsors for their practical assistance, advice, and financial and moral support.

Finally, a very warm welcome to you all. I sincerely hope and expect that you will profit by your attendance at this Symposium.

KEYNOTE ADDRESS

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

KEYNOTE ADDRESS: GEOGRAPHY AND STATISTICIANS

M.F. Goodchild[1]

ABSTRACT

Spatial data present unique problems to statistics because of their inherent properties, particularly of spatial heterogeneity and spatial dependence. The paper reviews the practical implications of these properties in such areas as modifiable reporting zones and ecological fallacies. It looks at the development of the spatial tradition in statistics, and the statistical tradition in geography. Current interest in geographic information systems offers an opportunity to bring the two groups closer together, and to address longstanding issues in a systematic and practical manner. New developments in exploratory spatial analysis and visualization add to the impact of digital technology and enhance the value of the spatial perspective.

KEY WORDS: Spatial data; Spatial statistics; Thematic mapping; GIS.

1. INTRODUCTION

Geography has always been a discipline of discovery. In Classical times it was a form of mathematics, concerned with measuring and projecting the earth and developing methods of navigation. From the 15th to the 19th Centuries it explored and mapped the Earth, and its practitioners were navigators and natural historians. Now that we know the source of the Nile, and new measurements of the height of Mt. Everest no longer capture the public imagination, geography has turned to the study of the processes which form and modify the physical landscape, to the dynamic and evolving relationships between humanity and the environment, and to the processes which shape the geographical diversity of human culture. Geography is infinitely complex: the more closely one looks at any of its aspects, human or physical, the more detail one sees and the more there is to explain and understand (Mandelbrot 1967).

Geography discovered statistics in its Quantitative Revolution which began in the 1950s and continued through the late 1960s. To most, this meant the application of the standard battery of statistical techniques, from student's t and F through principal components analysis and canonical correlation, to observed cases of some phenomenon. The fact that these cases were embedded in a spatio-temporal continuum was not given any great importance - space and time merely provided the sampling frame.

To a smaller group, it became increasingly apparent that geography was a special case. Some discovered spatial statistics, and used techniques such as point pattern analysis to make inferences about spatial processes (Getis and Boots 1978). Others explored spatial autocorrelation (Cliff and Ord 1981), random fields, regionalized variables or geostatistics (Isaaks and Srivastava 1989), and some made significant contributions. But these are difficult areas, and have remained largely outside the coverage of standard courses in statistics for geographers. Even today, there is very little spatial statistics in the average undergraduate course text, which focuses on the application of the standard battery of (nonspatial) statistical tests to geographical problems (see for example Barber 1988; Clark and Hosking 1986; for a notable exceptions, see Griffith and Amrhein 1991).

---

[1] M.F. Goodchild, Director, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA 93106, U.S.A.

Continuing developments that began in the 1980s are likely to change this situation, for the better. I will call these Geographic Information Systems (GIS), although the term is used in a much broader sense than the currently available set of GIS software packages (see for example Burrough 1986; Maguire, Goodchild and Rhind 1991). This paper is structured into four sections. The first identifies the areas of statistics that are of most relevance to geographical analysis and GIS. The second looks at some important issues in the application of statistics to geographical data in a GIS context. This is followed by a section that focuses on the use of statistical approaches to the problem of accuracy in spatial data, which is a key issue in the use of GIS technology. Finally, the last section looks at issues of implementation.

2. WHAT KINDS OF STATISTICS?

Perhaps the most striking property of geographical data from a statistical point of view is spatial dependence. Geographers might express this in the form: all places are related but nearby places are more related than distant places. It is almost impossible to conceive of a spatially independent variable, one whose value is unpredictable over the shortest distances. The variogram used in geostatistics shows the empirical variance between observations as a function of the distance between them, and is observed to rise monotonically to a "sill" at a distance termed the "range" and then fail to rise further, for most geographically distributed variables. This "range" is also known in geomorphology as the "grain" of the landscape, the distance beyond which the landscape fails to provide further surprises.

One way of understanding the effect of spatial dependence is as an inflation in the degrees of freedom of a test. An observation taken within a short distance of another one may not be a "new" observation, if the variable is strongly spatially dependent (Richardson 1990).

A second, again often ignored property is spatial heterogeneity (Anselin 1989). It is common for the coefficients of models to vary spatially, and it may be meaningless to try to estimate an "average" value for an arbitrarily defined study area. Rather, geographers may find more value in techniques such as adaptive spatial filtering, expansion methods (Casetti 1972) or geographical brushing (Monmonier 1989) that explicitly estimate the spatial variability of the model's coefficients. In this context, the conventional view that the study area provides a sample from some larger and normally unspecified population is highly inappropriate.

A third issue in the analysis of geographical data is the existence of hierarchical structures. Simple geographical objects may be members of larger, complex objects, and processes often operate across scales. In this sense, the view of a sample as a set of similar objects with associated attributes can be very limiting. From a database point of view, the "flat file" model which is implicit in many statistical tests places a severe restriction on our ability to understand spatial processes.

Finally, geographical space is inherently continuous and infinitely complex; a complete description would require an infinite number of records or "tuples". To describe geographical space in finite terms, some explicit discretization is required, through sampling, aggregation, generalization, modeling or approximation. Geographical data are thus almost always inaccurate, although the level of uncertainty associated with each item may sometimes be known. The objects that populate a spatial database - contour lines, sample points, patches on land cover maps, census tracts - may be no more than artifacts of this process of discretization, and their relationship to the underlying continuum may be difficult to characterize, or even unknown.

To summarize, geography needs statistics that explicitly recognize spatial dependence, and has great difficulty with the independence assumption made in so many statistical tests. Without them, the process of inference is likely to be misleading. It is interested in processes that are spatially non-stationary, and in techniques that can estimate the spatial variation of coefficients, since it is often unreasonable to assume stationarity within the arbitrary limits of a study area. It deals with processes that operate across scales, and with complex hierarchical relationships. And finally, many geographical observations represent arbitrary samplings or discretizations of continuous variation, rather than attributes of real objects.

These underlying problems often surface in concern over two specific issues: the Modifiable Areal Unit Problem and the Ecological Fallacy. Openshaw and Taylor (1979) provided perhaps the most dramatic illustration of the

MAUP, when they showed that it was possible to obtain correlations ranging from -0.99 to +0.99 between two spatial variables simply by reaggregating to different reporting zones. Without spatial dependence or spatial heterogeneity, reporting zones would be analogous to samples of an underlying population, and their effects would be no different from sampling effects. But Openshaw and Taylor showed clearly that reporting zones have a much more significant effect, as gerrymandering politicians have known for a long time.

One commits an Ecological Fallacy (Robinson 1950) when one assumes that all components of a geographical aggregate have the properties of the aggregate as a whole. In contrast to the MAUP, the EF would never occur if reporting zones were always perfectly homogeneous. Yet it is made constantly, and the assumption of homogeneity is what lies behind the use of cluster techniques (Weiss 1988) and a number of other methods in market research. It would be easy to compute and report measures of diversity for census reporting zones, and this could be done without significant infringement on confidentiality; the practice of only reporting means, averages and counts must sometimes have the effect of encouraging census users to commit the Ecological Fallacy.

3. APPLICATION ISSUES

Until the early 1980s, very little was available in the way of digital technology for supporting spatial statistics. The statistical packages such as SAS, SPSS or S offered only non-spatial methods. Spatial information could be processed by computer mapping packages, but only for the purpose of making a map. Limited mapping capabilities also existed in such packages as SAS, but the spatial data input to make maps could not be used to support spatial analysis.

Geographic Information Systems became widely available beginning in the early 1980s, and have diffused rapidly within universities, government agencies and the private sector. Unlike computer mapping packages, their primary purpose is to support the analysis of spatial data, although their capabilities to do so are often limited to simple geometric operations.

Packages to support spatial statistics also appeared during the 1980s, although they have not had the same degree of impact as GIS. Some, such as the widely distributed GEOEAS package for geostatistics developed by the U.S. Environmental Protection Agency, or Anselin's SPACESTAT (Anselin 1990), are stand alone packages with limited mapping capabilities. Griffith (1988) has developed procedures for spatial statistics using SAS and MINITAB, and others have linked GIS to standard statistical packages or standalone modules (Ding and Fotheringham 1991). All are oriented toward carrying out standard tests on well-defined sets of input data.

GIS is a technology of the 1980s, and its underlying interactive paradigm bears little relationship to the batch processing of the 1960s. As such it is an ideal basis for support of an exploratory approach to statistical analysis. In fact there has been increasing interest recently in the concept of Exploratory Spatial Analysis, by analogy to the Exploratory Data Analysis paradigm that invaded statistics in the late 1970s. It seems particularly inappropriate to adopt a lockstep confirmatory process given the difficulties imposed by the four characteristics of spatial data discussed earlier. Instead, a researcher working with geographical data should be encouraged to see the technology as a way of exploring a poorly formulated set of ideas. Visualization is important to this process, confidence limits may be more helpful than hypothesis tests, and randomization techniques may be more robust than the more traditional parametric tests.

3.1 Spatial dependence

We assume that spatial dependence is always present in geographical data, and must be recognized explicitly unless the spacing of samples is greater than the range. If a test of spatial dependence (Cliff and Ord 1981) shows it to be absent, this inference will always constitute a Type II statistical error - the null hypothesis of an absence of spatial dependence will have been accepted when in fact it is false. Degrees of freedom will always be inflated.

This position is orthogonal to tradition, and particularly in courses where spatial dependence is taught as an exception or problem, and where absence of spatial dependence among residuals is often used as diagnostic of

a fully-specified model. Unfortunately models of spatial dependence are difficult, in part because of the simultaneous nature of spatial dependence, which is in contrast to the one-directional nature of temporal dependence. Visualization may be the only way of communicating information about spatial dependence, since the parameters of models of spatial dependence have so little intuitive meaning.

Two families of techniques are available for dealing with spatial dependence, depending on the representation of space. The term "Spatial Statistics" normally refers to techniques that assume a space populated by well-defined objects that interact as wholes (Arbia 1989), even though they may actually be arbitrarily define objects serving as reporting zones for some underlying continuum. Measures of spatial dependence include the Moran and Geary measures of spatial autocorrelation (Cliff and Ord 1981), and focus on the level of dependence between neighboring objects. Autoregressive and moving average models are available to model processes between objects. Much currently available software makes use of standard packages (Griffith 1988). Finally, for these techniques the representation of space is reduced to a simple matrix of adjacencies, proximities or connectivities between objects, and results are invariant under any spatial transformations that preserve this matrix.

The term "Geostatistics" is used for techniques that model a continuous underlying distribution through point or block samples. Interpolation of this underlying surface is often the major objective. Variables are measured on continuous scales, and spatial dependence is described through the variogram or spectrum. Kriging is perhaps the best known technique of geostatistics (Isaaks and Srivastava 1989).

3.2 Other methods of spatial statistics

Besides these, spatial statistics includes a range of tests of various models of pattern. In the simplest case, it is often desirable to test whether a pattern of points could have been generated by a given stochastic process. To identify clusters of cases in epidemiology, for example, one needs tests that can distinguish between spatially constant and spatially varying rates of disease, and in the latter case provide estimates of the geographic distribution of rates. In other cases it may be desirable to distinguish between clusters resulting from spatially varying rates, and those resulting from contagious processes. Openshaw's Geographical Analysis Machine (Openshaw *et al.* 1987) is a stimulating example of the use of GIS technology to test for clusters.

Numerous descriptive indices have been defined for geographical patterns. Perhaps the best known of these is the centroid, defined as the mean of the x and y coordinates of a point set, invariant under rotation of the axes, and the point that minimizes the total of squared distances to the point set. Movement of the centroid is a useful indicator of changes in geographical distributions. Centroids also provide the only available information on geographic distributions of population at the lowest level of each national system of reporting zones - the EA in Canada and the ED or block in the U.S. In the U.K., some recent work by Bracken and Martin (1989) is providing estimates of continuous distributions at very large scales by disaggregating centroid counts using intelligently specified kernel functions. Centroids and other measures of geographic central tendency also have normative value as logical places from which to provide service to a dispersed population. More sophisticated and appropriate techniques for central facilities location are discussed in a practical context by Ghosh and Rushton (1987).

3.3 Exploratory spatial analysis

Exploratory Data Analysis (EDA) is already well accepted as a paradigm in statistical analysis of nonspatial data. It provides a set of techniques for adding to the scientist's natural intuition, and for developing hypotheses that can later be checked more objectively and deliberately. So it seems reasonable to assume that there is a similar potential role for Exploratory Spatial Analysis, or ESA. A small number of papers have already explored the potential of ESA. In fact preliminary indications are that ESA is significantly different from EDA, and may have substantially more potential.

We define ESA here in comparative terms, as a set of techniques designed to improve on or sharpen the scientist's ability to reach inferences by exploring data in its spatial context. The most valuable kinds of ESA would therefore be ones designed to aid the researcher in areas of spatial reasoning where human intuition is not good. Some of these are obvious, and have been used to justify GIS in the past: the difficulty of combining

evidence from different spatial sources, overcome in a GIS through the operation of overlay; or the difficulty of combining evidence of different forms - position (pattern), attributes, sound and text - when limited by the capabilities of a map display. The same argument for overlay can be made with respect to detecting change in a series of images: the eye is not good at coregistering and differencing, and thus a GIS can be a very effective tool at detecting change in a spatiotemporal time series. Nor is human intuition good at measuring spatial objects - a longstanding literature in cartography describes the problem of controlling the human eye's perception of the size of an object.

ESA tools can also improve on the eye's difficulty in integrating under a surface, to make estimates of total population within an arbitrary area from a contour map, for example. The reverse operation is also difficult - estimating a density surface from a map of dots. People are not good at reasoning between scales, or at testing patterns against statistical process models - it is easy to find numerous instances of intuition being misled into thinking that a random pattern of dots shows some systematic tendency. And finally, the eye is not good at separating regional from local trends, or detecting spatial outliers, particularly in multivariate data.

Haslett's SPIDER and similar systems (Haslett *et al.* 1991) make use of multiple screen windows to show data from different perspectives - map, time series, scatter plot, histogram - and then link them logically so that actions performed with the mouse in one window are reflected in the others. Pointing to an area on a map causes the corresponding point in a scatterplot to be highlighted. Animation of time series is also an effective but simple way to aid human intuition in gaining insight into spatial processes.

4. THE ACCURACY PROBLEM

We have already noted that almost all spatial data are inherently uncertain and inaccurate. At the same time there is a longstanding tradition in cartography to establish standards for the accuracy of features shown on maps, particularly positional accuracy. Many regulate the distance within which the true location of a point must lie 90% of the time - the Circular Map Accuracy Standard (CMAS) - and set it to a fraction of a millimeter on the finished map product. But for many types of geographical data it is impossible to measure CMAS because the objects being represented are artifacts of the process of discretization, and are not identifiable on the ground. CMAS is also concerned with point accuracy, and cannot be applied to complex line and area objects that are more than a collection of discrete points.

Substantial research has been devoted to developing error models for spatial databases. We define an error model as a stochastic process capable of replicating the distortions present in spatial data as a result of all sources of uncertainty. The variation between replications might be interpreted as the variation between alternative spatial database representations of the same geographical truth due to variation between observers, interpreters, measuring instruments or digitizers. The simplest such error model would be the map equivalent of the Gaussian distribution for scalar measurements - the distribution one would assume if one knew nothing whatever about the exact nature of the error process, other than that error is the additive combination of a large number of independent contributions.

Error modeling research is currently under way at a number of universities. All take the same basic approach: that uncertainty in the database can be described by an error model and associated parameters; that it is desirable to know the effects of this uncertainty in the form of confidence limits on the products from processing of the database; and that the mathematics of error analysis are complex, and therefore should be largely hidden from the user. At Newcastle, Openshaw and his group (Carver 1991) are working with a simple error band model that provides limits to the spatial deviation of every line in the database. At Utrecht, Burrough and his group (Heuvelink, Burrough and Stein 1989) use Monte Carlo simulation and Taylor series expansions to propagate error in the processing of rasters of continuous data. At Santa Barbara (Goodchild, Sun and Yang 1992), we have been using GRASS to simulate the propagation of error in operations on discrete multinomial land cover maps, based on a spatially autoregressive model of uncertainty.

5. IMPLEMENTATION ISSUES

If GIS provides the potential for more effective use of spatial statistics in the analysis of spatial data, then what developments are needed to bring this about? Progress to date has been limited, and there are still very few examples of effective spatial statistical research using GIS in the literature. At the same time we have a vast array of statistical software with very little spatial capabilities.

The GIS industry is large and flourishing - estimates of its size range from several hundred millions to several billions of dollars. But its market is largely in areas that require very little analysis - the management of vast collections of spatially distributed facilities by local governments, utility and natural resource companies and agencies. Thus recent developments in GIS software tend to have added much more to their capabilities as database systems than to their analytic power.

There seem to be three models for integration of GIS and techniques for statistical analysis. The first, termed here full integration, would require the addition of analytic capabilities to the GIS itself, and seems unlikely to occur without a significant reorientation of market priorities. The second, termed here loose coupling, is what we have at present. Data from GIS is communicated to analytic codes, either packages or standalone modules, through flat files of ASCII records. In this mode all of the higher level structures of spatial data - hierarchical relationships, adjacencies and complex geometries - are lost, and it is difficult (and may sometimes be impossible) to pass the results of analysis back to the GIS for mapping or storage. The third mode is termed close coupling, and results when the GIS and statistical package share common data models, thus preserving the full structure of the spatial data. Unfortunately at this time none of the readily available statistical packages recognize structures more complex than the simple flat file of records. On the other hand GIS tend to represent complex spatial structures through various implementations of the relational data model.

6. PRIORITIES

If the pace of current developments in GIS, spatial statistics and ESA is anything to go by, the future looks far from bleak. We are likely to see the emergence of very exciting packages for the analysis of social and economic data, transportation planning, emergency facility management, epidemiology, and site selection, that implement many of our existing methods of analysis in much more effective ways through GIS technology. Multimedia and hypermedia techniques open up entirely new perspectives on spatial data, and invite new forms of visualization and inference that go far beyond the familiar hard copy map. At the same time intuition can be very misleading, and statistical analysis in a spatial context is far from simple and straightforward.

Given the potential, there is a real problem knowing where and how to begin. Spatial statistics and spatial analysis comprise a wide range of methods and there are few organizing principles available to simplify the set. We can distinguish to some extent between analysis and modeling, and between exploratory and confirmatory paradigms, but the most effective distinction from a GIS perspective lies in the data model - the set of objects that is created from the process of discretizing continuous geographic variation.

Spatial statistics relies on remarkably few data models, fewer in fact than the set currently supported by available GIS. Perhaps the simplest is the undifferentiated set of points, used for example in epidemiology and point pattern analysis. Another is the table of attributes plus matrix of proximities, used in the measurement of spatial dependence and in autoregressive and moving average models. More complex are models based on attributes for two sets of objects, plus a rectangular table of interactions or proximities, used for example in modeling flows and spatial interactions. Another is the set of point samples from a continuous distribution, and another is the raster of cells - both of these are common in the study of physical and environmental processes.

GIS has been successful in part because it has managed to integrate many of these data models into a common framework. But while the diversity of models may be a major benefit of GIS, it is a significant barrier to the development of spatial statistics and their integration with GIS. Most successful efforts at software integration rely on the existence of a simple, common data model, and the statistical packages have been built in this fashion around the simple flat file. GIS's main contribution to spatial statistics may turn out to be its insistence on an explicit recognition of data models as the latter's primary organizing principle.

ACKNOWLEDGMENT

The National Center for Geographic Information and Analysis is supported by the National Science Foundation, grant SES 88-10917.

REFERENCES

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: Department of Geography, University of California.

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., and 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*, Brussels 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., and Hosking, P.L. (1986). *Statistical Methods for Geographers*, New York: Wiley.

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

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

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

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

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

Haslett, J., Bradley, R., Craig, P., Unwin, A., and others (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., and Stein, A. (1989). Propagation of errors in spatial modelling with GIS. *International Journal of Geographical Information Systems*, 3, 303-322.

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

Maguire, D.J., Goodchild, M.F., and 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., and 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., and Taylor, P.J. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In N. Wrigley (ed.) *Statistical Applications in the Spatial Sciences*, London: Pion, 127-144.

Richardson, S. (1990). Some remarks on the testing of association between spatial processes, *Spatial statistics: Past, Present and Future*, (ed.) 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

Geographic Perspectives on Data Modelling

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

STATE SPACE COMPOSITE ESTIMATION FOR SMALL AREAS

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

ABSTRACT

In estimation for small areas it is common to borrow strength from other small areas since the direct survey estimates often have large sampling variability. Often a composite estimator is formed which is a linear combination of the direct survey estimator with a synthetic or regression estimator. Recent developments have been aimed at also borrowing strength over time for repeated surveys using time series models for the small area parameters and the survey errors. We propose a state space composite estimator for which the Kalman filter for fitting time series models is generalized so that it does not require modelling of the survey errors. Instead, an additional covariance term which is due to serially correlated survey errors appears in the filtering equation. The jackknife resampling method is proposed as a method for non-parametric estimation of this covariance term. Application to the Canadian Labour Force Survey is discussed.

KEY WORDS: Repeated surveys; Structural models; Spatial and temporal smoothing; Jackknifed Kalman filter.

1. INTRODUCTION

We consider a periodic survey in which estimation of domain totals for the most recent survey occasion is of interest. The domains would be unplanned minor domains as defined by Purcell and Kish (1980). The problem of small area estimation arises because the realized sample size in a domain of interest may be small or even zero. Consequently, the direct estimator based on the sample would be unreliable or may not even be defined. There are several approaches to this problem; however, the main idea behind all of them is augment the current sample data with some extra information. This could consist of an implicit or explicit model that would allow us to make use of some covariate information. Data from other small areas or from previous survey occasions might also be used, again *via* a model. These approaches generally lead to increased stability of the small area estimates at the cost of some bias.

If data from other, similar small areas within a larger area are used then one possible approach is synthetic estimation in which a regression model is fitted using all the data and the small area estimate is just the estimated model expectation. If the model is not very good this could introduce severe bias in the small area estimators. Composite estimators which are convex combinations of design based and synthetic estimators attempt to balance the potential bias of the synthetic estimators against the instability of the design based estimators. This could be done systematically, based on an extension of the model, or in an ad hoc manner. If information about the same domain from earlier time points is used, estimates of change based on common samples could be used to adjust estimates from the previous survey occasion and the new estimator would be a convex combination of this with the design based estimator. If the method is applied recursively then data from all previous survey occasions could be used. There are other variations as well in which the parameters (θ_{kt}'s, k denoting the small area and t the time) are assumed random over both k and t under a superpopulation model. For estimating a given θ_{kt} strength is borrowed from other areas and from other time points by assuming that the θ_{kt}'s are connected over k and t though a suitable mixed linear model. An

---

[1] A.C. Singh and H.J. Mantel, Social Survey Methods Division, 16-A, R.H. Coats Building, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

important advantage of this approach based on random θ_{kt}'s is that it does not require common (or overlapping) sample units over time.

In this paper, we shall first review briefly in sections 2 and 3 various methods of small area estimation with emphasis on composite estimation. The term "composite" will be used in a general sense meaning that the estimator is a convex linear combination of the current and some synthetic estimator based on other areas or time points or both. The term "spatial smoothing" will be used to denote adjustments to the current estimate based on modelling data from other areas for the same time while the term "temporal smoothing" will be used for adjustment based on modelling data from earlier time points about the same area. Section 2 corresponds to the case of nonrandom θ_{kt}'s and section 3 to the case of random θ_{kt}'s. In section 3 a state space modelling approach to spatial and temporal smoothing for small area estimation is also reviewed. Here a state space framework is employed for connecting random θ_{kt}'s. In this context, we propose in section 4 an estimator termed the state space composite estimator. The proposed estimator requires modification of the Kalman filter to handle the case of correlated survey errors due to overlapping sample units over time. For this purpose, we also propose the jackknife resampling technique for estimation of the relevant covariance terms in the modified filtering equation. Section 5 contains some directions for future work, including a proposed simulation study using Canadian Labour Force Survey data.

2. REVIEW OF SMALL AREA ESTIMATION (θ_{kt}'s nonrandom)

2.1 Spatial smoothing for nonrandom θ_{kt}'s

For a given time t, a regression model is used to describe the θ_{kt}'s over k (k=1, ...K) within a larger area as functions of known covariates and some unknown parameters. The direct small area estimators (*e.g.* expansion or post-stratified) are used to estimate the parameters. The synthetic estimator, which is just the regression function evaluated at the known covariate values and the estimated parameter values, could be heavily biased unless the model is carefully chosen. The main reference for synthetic estimation is Gonzalez (1973).

In practice, a composite estimator which is a linear combination of synthetic and direct small area estimators may be preferred. The aim would be to balance the potential bias of the synthetic estimator against the instability of the direct estimator. The sample size dependent estimator of Drew, Singh, and Choudhry (1982) provides an example of composite estimation.

2.2 Temporal smoothing for nonrandom θ_{kt}'s

For a fixed small area k, it may be possible to relate the direct small area estimators over time t. In the case of overlapping repeated surveys, we can adjust the previous estimates by estimates of change based on matched units. This, then, provides extra information to improve the current estimator of θ_{kT} at time T. The idea was used by Jessen (1942), and Patterson (1950), among others (see Binder and Hidiroglou 1988 for a review). A general multivariate approach developed by Gurney and Daly (1965) consists of using the concept of "elementary" estimates. Under a multivariate linear model and assuming that the error covariance matrix can be specified, we can get MVLUE of $\{\theta_{kt}: t=1, \ldots T\}$. In practice, the specification and the task of inversion of the covariance matrix may lead to instability of the resulting estimates. The computations can be considerably simplified by inducing a special serial dependence structure when the θ_{kt}'s are assumed random, see section 3. A compromise to the MVLUE approach was also suggested by Gurney and Daly (1965), following Hansen, Hurwitz and Madow (1953), in which a composite estimator based on the current estimator and the previous time point estimator (adjusted for change) is computed recursively.

2.3 Spatial and temporal smoothing for nonrandom θ_{kt}'s

The general multivariate approach of Gurney and Daly (1965) mentioned above can also be used when both k and t vary. However, the problem of reliable specification and inversion of the covariance matrix will be further compounded.

Purcell and Kish (1980) propose another type of spatial-temporal smoothing termed SPREE (structure preserving estimation) which is suitable for frequency data. Reliable historical data, perhaps from a census, is used to construct a multi-dimensional table of counts or proportions in which one of the dimensions would correspond to small area, another to any factors for which we desire current counts within small areas, and other dimensions to associated factors. This is called the association structure. Current information is used to obtain reliable estimates of some of the marginal tables; this is called the allocation structure. Iterative proportional fitting would be used to reconcile the association structure to the allocation structure and then the associated factors would be collapsed to produce a two dimensional table with small areas and the factor of interest.

3. REVIEW OF SMALL AREA ESTIMATION (θ_{kt}'s random)

3.1 Spatial smoothing for random θ_{kt}'s

The main advantage of treating the θ_{kt}'s as random is that this provides a framework for a rational compromise between the potential bias of synthetic estimators and the instability of the direct estimators. Some of the main references are Fay and Herriot (1979), Särndal (1984), Battese, Harter, and Fuller (1988), Särndal and Hidiroglou (1989), Pfeffermann and Barnard (1991), Datta and Ghosh (1991), and Ghosh and Rao (1991). We will follow the basic framework as in Fay and Herriot, but instead of using a Bayesian viewpoint, we shall use the approach of best linear unbiased predictors (BLUP) for mixed linear models. It will be assumed that the covariate information, if any, is available at the small area level. If it is available at the sampling unit level, then perhaps more efficient estimators could be developed. These are, however, not considered in this paper.

For a fixed time t, the Fay-Herriot set-up is given by

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

$$\text{i.e.} \quad \underset{\sim}{y}_t = F_t \underset{\sim}{\beta}_t + \underset{\sim}{a}_t + \underset{\sim}{\varepsilon}_t \tag{3.1b}$$

where $\underset{\sim}{y}_t$ are the direct estimates, $\underset{\sim}{a}_t \sim (\underset{\sim}{0}, W_t)$, $\underset{\sim}{\varepsilon}_t \sim (\underset{\sim}{0}, V_t)$ and $\underset{\sim}{\varepsilon}_t$, $\underset{\sim}{a}_t$ are uncorrelated. Then it follows from the appendix that the BLUP of $\underset{\sim}{\theta}_t$ is

$$\hat{\underset{\sim}{\theta}}_t = F_t \hat{\underset{\sim}{\beta}}_t + \hat{\underset{\sim}{a}}_t \tag{3.2a}$$

where

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

It should be noted that $\hat{\underset{\sim}{\theta}}_t$ in (3.2a) is also a composite estimator; that is, it can be written as a convex combination of the direct estimator $\underset{\sim}{y}_t$ with the synthetic estimator $F_t \hat{\underset{\sim}{\beta}}_t$.

3.2 Temporal smoothing for random θ_{kt}'s

Some of the main references are Blight and Scott (1973), Scott and Smith (1974), Jones (1980), Bell and Hillmer (1987), Binder and Dick (1989), Choudhry and Rao (1989), Pfeffermann and Burck (1990), and Pfeffermann (1991). Many of the papers cited above do not directly address the problem of small area estimation, however, the underlying idea is essentially the same in dealing with the estimation of population means from repeated surveys using time series models. In this paper, we consider only structural time series modelling and the associated state space framework. For a given area k, modelling of the time series of direct small area estimators $\{y_{kt} : t = 1, \ldots, T\}$ is simply a special case of modelling the multivariate time series $\{\underset{\sim}{y}_t : t = 1, \ldots, T\}$ which is considered in the next sub-section. Therefore, we will not go into further details for this case.

3.3 Spatial and temporal smoothing for random θ_{kt}'s

This is the main topic which we wish to consider in this paper. We would like to borrow strength over both t and k using structural time series models. A recent paper by Pfeffermann and Burck (1990) provides an excellent formulation of a general class of models for small area estimation which includes spatial and temporal smoothing. In this paper we review an important sub-class which generalizes Fay-Herriot estimation.

Consider the Fay-Herriot set-up given by (3.1) above. For structural time series modelling, serial dependence is induced by letting $\underset{\sim}{\beta}_t$ and $\underset{\sim}{a}_t$ evolve over time. Letting $\underset{\sim}{\alpha}_t^T = (\underset{\sim}{\beta}_t^T \ \underset{\sim}{a}_t^T)$ and $H_t = (F_t \ I)$, the general model that we wish to consider has the form

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

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

where the $\underset{\sim}{\zeta}_t$'s are independent between time points and independent of all the $\underset{\sim}{\varepsilon}_t$'s. The usual state space model assumes that the $\underset{\sim}{\varepsilon}_t$'s are independent between time points; however, for repeated surveys the case where they may be correlated over time is more realistic. Thus, we consider two cases in the following subsections. We also assume that the matrices H_t, G_t, V_t and Γ_t are known. In practice, V_t and Γ_t would generally need to be estimated.

3.3.1 Uncorrelated survey errors $\underset{\sim}{\varepsilon}_t$'s

The case of independent $\underset{\sim}{\varepsilon}_t$'s (*i.e.* when samples for different survey occasions are selected independently) is considered by Singh, Mantel & Thomas (1991) where some empirical results are also presented. When the $\underset{\sim}{\varepsilon}_t$'s are uncorrelated the Kalman filter (KF) is a computationally feasible method for obtaining the BLUP of $\underset{\sim}{\alpha}_T$ based on the data $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_t$, see *e.g.* Duncan and Horn (1972). The KF is a recursive procedure that optimally combines the data $\underset{\sim}{y}_t$ with $\hat{\underset{\sim}{\alpha}}_{t-1}$ where $\hat{\underset{\sim}{\alpha}}_{t-1}$ is the BLUP of $\underset{\sim}{\alpha}_{t-1}$ based on $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_{t-1}$. To be more precise,

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

where $\hat{\underset{\sim}{\alpha}}_{t|t-1} = G_t \hat{\underset{\sim}{\alpha}}_{t-1}$ and $P_{t|t-1}$ is the MSE of $\hat{\underset{\sim}{\alpha}}_{t|t-1}$ as an estimate of $\underset{\sim}{\alpha}_t$. There are corresponding updating equations for P_t, the MSE of $\hat{\underset{\sim}{\alpha}}_t$ as an estimate of $\underset{\sim}{\alpha}_t$, and for $P_{t|t-1}$. From $\hat{\underset{\sim}{\alpha}}_t$ one can easily get the BLUP of the small area parameters $\underset{\sim}{\theta}_t$. Note that $\hat{\underset{\sim}{\alpha}}_t$ in (3.4) is also a composite estimator which could be written as a convex combination of the direct estimator $\underset{\sim}{y}_t$ with $H_t \hat{\underset{\sim}{\alpha}}_{t|t-1}$.

To estimate parameters of the covariance matrix Γ_t one can use maximum likelihood assuming normality of the error terms, see *e.g.* Pfeffermann and Burck (1991) for the use of the Newton-Raphson algorithm to obtain maximum likelihood estimates. Alternatively, the EM algorithm can be used, see Shumway and Stoffer (1982). However, for known G_t the method of moments can also be used which does not require normality of the error terms. This is similar to the method used in Fay & Herriot (1979) and is used in Singh, Mantel & Thomas (1991). When estimated values of the model parameters are substituted in the BLUP, it becomes the EBLUP (empirical best linear unbiased predictor). The MSE of the EBLUP would be somewhat larger than that of the BLUP and a corresponding adjustment to the estimated MSE can be made under certain conditions along the lines of Prasad & Rao (1990).

3.3.2 Correlated survey errors $\underset{\sim}{\varepsilon}_t$'s

This is the more realistic case. One way to handle this case is to assume that the survey errors can be modelled by an ARMA process, see *e.g.* Binder and Dick (1989), Pfeffermann and Burck (1990), and Pfeffermann (1991). The state vector is suitably enlarged so that the new measurement errors become uncorrelated. Now the usual KF can be run to get the desired BLUPs. The above method of handling correlated survey errors can be referred to as the parametric approach. It would be useful to have an alternative approach which is nonparametric in nature for times when it is difficult to formulate an appropriate model for the $\underset{\sim}{\varepsilon}_t$'s. This

approach might also be used to check the robustness of estimates using the parametric approach. In the next section we propose such an alternative based on a jackknifed version of the KF.

4. STATE SPACE COMPOSITE ESTIMATION BY THE JACKKNIFED KALMAN FILTER

We now describe a method of composite estimation for the correlated survey errors case which does not require time series modelling of the survey error process. When the $\underset{\sim}{\varepsilon}_t$'s are correlated over time the KF updating equation (3.4) can be modified to

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

where C_t is the covariance of $\hat{\underset{\sim}{\alpha}}_{t|t-1} - \underset{\sim}{\alpha}_t$ with $\underset{\sim}{\varepsilon}_t$. Here $\hat{\underset{\sim}{\alpha}}_t$ is the composite estimator which is the optimal combination of $\underset{\sim}{y}_t$ with $\hat{\underset{\sim}{\alpha}}_{t|t-1}$. The resulting $\hat{\underset{\sim}{\alpha}}_t$ would not generally be the BLUP of $\underset{\sim}{\alpha}_t$; however, it still offers a computationally feasible way to take account of all of the data and should lead to a fairly efficient estimator of $\underset{\sim}{\alpha}_T$. This type of estimation, which can be referred to as state space composite estimation, has some similarity to the type of composite estimation which was introduced by Hansen, Hurwitz and Madow (1953) and which was shown to be quite efficient in empirical studies by Gurney and Daly (1965) and Wolter (1979).

The proposed method for correlated survey errors is based on the observation that the relevant covariances can be estimated unbiasedly simply by design based methods. To see this, letting $\underset{\sim}{\alpha}_t^* = (\underset{\sim}{\alpha}_0, \ldots, \underset{\sim}{\alpha}_t)$ we note that

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

The second term here is identically zero since $E\left[\underset{\sim}{y}_t - H_t\underset{\sim}{\alpha}_t \mid \underset{\sim}{\alpha}_t^*\right] = 0$ and the first factor is fixed for given $\underset{\sim}{\alpha}_t^*$. So, it is sufficient to get an estimate of $\mathrm{Cov}(\hat{\underset{\sim}{\alpha}}_{t|t-1}, \underset{\sim}{y}_t \mid \underset{\sim}{\alpha}_t^*)$, which can be obtained by a resampling method such as the jackknife suitably defined for complex surveys. For multistage surveys application of the jackknife is usually implemented by dropping primary sampling units (PSUs) from the sample and observing how estimates are affected. In this repeated survey context dropping a PSU from the sample means dropping the same or corresponding PSU from each survey occasion. In particular, this requires a matching or identification of PSUs over time. For example, in a rotating panel survey PSUs which are in sample at both times t and $t-1$ could be matched, but when a PSU rotates out of the sample it is necessary to match it to a replacement PSU. Another implication is that the number of PSUs in the sample must be constant over survey occasions. The same practical restriction would also apply to other resampling methods.

The iterative procedure then, for the jackknifed Kalman filter (JKF), would be to calculate for each PSU in the sample how dropping that PSU would affect $\underset{\sim}{y}_t$. This would be combined with previous calculations of how dropping each PSU would affect $\hat{\underset{\sim}{\alpha}}_{t|t-1}$ to estimate the matrix C_t *via* the jackknife and to derive how dropping each unit would affect $\hat{\underset{\sim}{\alpha}}_{t+1|t}$.

Estimation of variance parameters for JKF can be done as in Singh, Mantel & Thomas (1991) by the method of moments for known G_t. However, extra information about $\mathrm{Cov}(\underset{\sim}{\varepsilon}_t, \underset{\sim}{\varepsilon}_{t-1})$ would be required for this purpose. This can also be estimated by the jackknife because PSUs are assumed to be connected over time. However, it may also be possible to use a design based estimate of $\mathrm{Cov}(\underset{\sim}{\varepsilon}_t, \underset{\sim}{\varepsilon}_{t-1})$.

5. FUTURE WORK

The jackknifed Kalman filter proposed here will be investigated in a realistic situation by a simulation study to compare it to other methods described in this paper, particularly the modelling approach.

A good example of a potential application of the jackknifed Kalman filter is the Canadian Labour Force Survey (LFS). The sample design for this survey is described in Statistics Canada (1990). To summarize, each province is divided into economic regions. Each economic region is covered by three frames, a self representing frame, a non self representing frame, and a special area frame consisting of remote areas, institutions and military bases.

The self representing frame includes urban areas large enough to have an expected sample size of at least fifty households, and each such urban area would be included in the sample. It is further sub-divided into regular and apartment frames. The primary sampling units for the regular frame are clusters consisting of groups of blocks or block faces containing an appropriate number of dwellings. For the apartment frame the primary sampling units are buildings. Dwellings within PSUs are selected systematically. Each month roughly one sixth of the dwellings in the sample are rotated out and replaced so that no household will be in the sample for more than six consecutive months. When all of the dwellings within a PSU have been exhausted the PSU itself rotates out of the sample and is replaced.

The non self representing frame is sub-divided into strata which are classified as being rural, urban, or mixed. PSUs are defined within strata on the basis of geographic contiguity and other factors and each PSU selected to the sample is further sub-divided into clusters or groups from which a systematic sample of dwellings is selected. A rotation pattern similar to that for self representing areas is employed.

For the application of the jackknife to the LFS we would delete first stage sampling units. The number of first stage sampling units within strata would generally be constant over time and, in most cases, when a unit rotates out of the sample there would be a well defined replacement unit with which it could be matched. Special procedures would be needed to handle the situations when this failed to hold true. Two such situations might be when there was a redesign of the survey, as happens roughly every ten years, and when there was an adjustment to the sample size.

Pfeffermann and Bleuer (1992) have already done much work developing a particular type of state space model for the rotating panel survey data obtained from the LFS and it would be useful to have a comparison of the two approaches, that is, the modelling and the jackknife approaches, in that context. A simulation study based on LFS data could be carried out. It would first be necessary to construct a longitudinal pseudo-population in which individuals and households are present for an extended period of time. LFS sampling would then be simulated in this pseudo-population. For simplicity and feasibility we could restrict the simulation study to one province and exclude special areas.

Further work can be done also to develop the jackknifed Kalman filter. Pfeffermann and Burck (1990) consider constraining small area estimates so that specified control totals are equal to direct survey estimates of the corresponding population totals. Constraints of this type help to robustify the procedures against model failure as well as forcing the small area estimates to be consistent with published estimates for larger areas. Such modifications could also be easily applied to the jackknifed Kalman filter.

The methods we have described use auxiliary information, if it is available, at the domain level. If auxiliary information is actually available at the unit level then there may be considerable loss of information in aggregating it to the domain level. It is therefore of interest to develop methods for small area estimation that would make use of auxiliary information at the unit level. Another issue related to auxiliary data is accounting for possible measurement errors. This issue would be pertinent, for example, if covariate information were obtained from a list frame and could be out of date or of dubious quality.

ACKNOWLEDGEMENT

We would like to thank Danny Pfeffermann and Jon Rao for useful discussions. The first author's research was supported in part by a grant from the Natural Sciences and Engineering Research Council of Canada held at Carleton University, Ottawa. Thanks are also due to Christine Larabie for assistance in processing the manuscript.

APPENDIX: Spatial smoothing for random $\underset{\sim}{\theta}_t$

Consider the random effects model

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

where $\underset{\sim}{\varepsilon}$ and $\underset{\sim}{a}$ are uncorrelated, $\text{rank}(X) = r$, $r<k$.

Lemma 1 (Rao 1973, p. 234, Pfeffermann 1984). If the mean of $\underset{\sim}{a}$ is not known then the BLUP of $\underset{\sim}{\theta}$ is the same as the BLUE of $\underset{\sim}{\theta}$ when it is regarded as being fixed. Moreover, the MSE also remains the same.

Now suppose $\underset{\sim}{a}$ has mean $\underset{\sim}{0}$. Then the mean of $\underset{\sim}{\theta}$ lies in an r dimensional subspace of $\mathbf{R}^k$. The BLUP of $\underset{\sim}{\theta}$ will be different from the BLUE of $\underset{\sim}{\theta}$ when it is regarded as being fixed. It follows from Rao (1973, p. 267) and also from Harville (1976) that the BLUP of $\underset{\sim}{a}$ based on $\underset{\sim}{y}$ is obtained by treating $E(\underset{\sim}{a} \mid \underset{\sim}{y})$ as a linear regression function and then substituting BLUEs for all the unknown parameters in the linear function. Thus we have

Lemma 2 ($\underset{\sim}{\beta}$ known). $$\hat{\underset{\sim}{a}} = \text{BLUP of } \underset{\sim}{a} = E(\underset{\sim}{a}) + \text{Cov}(\underset{\sim}{a}, \underset{\sim}{y})\text{Var}(\underset{\sim}{y})^{-1}(\underset{\sim}{y} - E(\underset{\sim}{y}))$$
$$= \underset{\sim}{0} + WU^{-1}(\underset{\sim}{y} - X\underset{\sim}{\beta})$$

where $U = V + W$, and

Lemma 3 ($\underset{\sim}{\beta}$ unknown). $$\hat{\underset{\sim}{a}} = \text{BLUP of } \underset{\sim}{a} = WU^{-1}(\underset{\sim}{y} - X\hat{\underset{\sim}{\beta}})$$

where $\hat{\underset{\sim}{\beta}} = (X'U^{-1}X)^{-1}X'U^{-1}\underset{\sim}{y}$ is the BLUE of $\underset{\sim}{\beta}$.

REFERENCES

Battese, G.E., Harter, R.M., and 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., and 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., and Dick, J.P. (1989). Modelling and estimation for repeated surveys, *Survey Methodology*, 15, 1, 29-45.

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

Choudhry, G.H., and Rao, J.N.K. (1989). Small area estimation using models that combine time series and cross-sectional data, *Proceedings of the Statistics Canada Symposium on Analysis of Data in Time*, (Eds. A.C. Singh and P. Whitridge), 67-74.

Datta, G.S., and 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., and Choudhry, G.H. (1982). Evaluation of small area estimation techniques for the Canadian Labour Force Survey, *Survey Methodology*, 8, 17-47.

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

Pfeffermann, D., and Burck, L. (1990). Robust small area estimation combining time series and cross-sectional data, *Survey Methodology*, 16, 2, 217-237.

Prasad, N.G.N., and 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., and 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., and Hidiroglou, M.A. (1989). Small domain estimation: a conditional analysis, *Journal of the American Statistical Association*, 84, 255-275.

Scott, A.J., and 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., and 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., and 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.

Statistics Canada (1990). *Methodology of the Canadian Labour Force Survey*, (authors: Singh, M.P., Drew, J.D., Gambino, J.G., and Mayda, F.) cat. no. 71-526.

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

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

THE EFFECT OF ALTERNATIVE GROUPINGS ON LOCAL AREA ESTIMATES OF UNDERCOUNT

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

ABSTRACT

The 1990 Post-Enumeration Survey (PES) was designed to produce Census tabulations of states and local areas corrected for undercount or overcount of the population. The budgeted PES sample was too small even to produce direct estimates for many states, much less for smaller areas. Therefore, the PES was designed to produce synthetic estimates. Estimates were formed for poststrata, which involved grouping areas across states. These estimates were smoothed using a statistical model. Alternate groupings of the data for estimation and modelling yielded important differences in the estimated undercount for local areas.

KEY WORDS: Census; Small area estimation; Model selection; Synthetic estimation.

1. INTRODUCTION

The 1990 Post-Enumeration Survey (PES) was designed to produce Census tabulations of states and local areas corrected for the undercount or overcount of population. It consisted of two parts. The first was a sample of the population, known as the P sample. The proportion of the P sample that was included in the Census estimated the proportion of the total population that was included in the Census. The proportion of the sample that is included in the Census is determined by matching the people in the P sample to the Census records. One needs a sample of the Census enumerations to estimate the number of erroneous enumerations. This sample is known as the E sample and is the other part of the PES. The E sample records are checked against the Census itself to determine the extent of duplication. They are also reinterviewed to determine the extent of fictitious enumerations, inclusions by the Census of people born after the Census reference day, *etc.* The PES used the dual-system model to estimate the true population. See Hogan (1991) for a description of the PES and the dual system estimator.

The initial set of estimates were needed by May 17, 1991. From these estimates, corrected Census tabulations were produced by July 15, 1991. The Secretary of Commerce ultimately decided not to correct the actual census tabulations using the results of the PES. Since that time, we have continued the research into alternative estimates of the undercount. Part of the current research has involved redoing some of the original matching and coding work. The results of such research are not reported here. Instead this paper focuses on only one aspect: the effect of forming the estimates using alternative geographic groupings.

The next two sections describe the original PES design and estimates, including the pre-specified smoothing work. Section 4 describes two alternative approaches to the smoothing. Section 5 compares the results of the approaches with each other as they affect estimates of undercount for states, counties and places.

---

[1] H. Hogan and C. Isaki, Statistical Research Division, Bureau of the Census, Washington, DC 20233. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

2. SAMPLING AND ESTIMATION

The PES was designed to support synthetic estimation of population counts for local areas. The population was grouped into poststrata. PES estimates of the true population for each of these groups were obtained. The ratios of the PES estimate of the true population to the Census counts by poststrata were called the adjustment factors, and were the basis for synthetic estimation.

Poststrata were formed to include persons as similar as possible with respect to their probability of being counted in the Census. The variables used to define the poststrata were Census division, size and type of place, race/origin, age, sex and tenure (owner/renter). See Table 1. Twelve separate poststrata by age-sex were created for American Indians living on reservations. A full cross-classification by all variables would produce many poststrata representing very small populations, so many cells are combined. For example, there is no tenure category outside central cities, nor separate category for Blacks living in rural New England. In sum, we constructed 116 poststrata groups each with 12 age-sex categories or 1,392 separate poststrata.

Table 1. Variables Used in Poststratification

Race/Origin:		Black, Non-Black Hispanic, Asian and Pacific Islanders, and all other
Age:		0-9, 10-19, 20-29, 30-44, 45-64, 65+
Sex:		Male, Female
Census Division:		New England, Middle Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain, Pacific
Place/Size:	*	Central City of Major Primary Metropolitan Statistical Areas (PMSA's)
	*	Central City of Large Metropolitan Statistical Areas (MSA's) (with at least one city with population of 250,000 or more)
	*	Central City of Small MSA
	*	PMSA or Large MSA: Not Central City
	*	Small MSA: Not Central City
	*	Non-MSA incorporated places with population of 10,000 or more
	*	All other
Tenure:		Owner vs. Renter

The primary sampling unit for the 1990 PES was the block cluster composed of either a block or a collection of blocks. A sample of approximately 5,300 block clusters was chosen. The same blocks were used for the P sample and the E sample. The E sample consisted of all Census enumerations, correct or incorrect, in the sample blocks. The P sample consisted of all people living in housing units and non-institutional group quarters in the sample blocks at the time of the PES interview, about 172,000 units. Dual system estimates were made for each of the 1,392 poststrata.

3. SMOOTHING AND CARRYING DOWN

It was anticipated that many of the 1,392 poststrata adjustment factors would have coefficients of variation too high to be useful for adjustment. A regression approach was adopted to reduce the variance. A regression was used to predict the adjustment factor, for each poststratum. The regression-predicted factor was then combined with the observed factor to form the smoothed factor The model thus attempted to "borrow strength" from many strata, somewhat in the spirit of an empirical Bayes estimation approach or equivalently, a variance components model.

Let

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

where

Y	=	Vector of observed adjustment factors by poststratum
X	=	Matrix of explanatory variables
β	=	Vector of regression parameters
w	=	Model error, assumed $N(0, \sigma^2 I)$
e	=	Sampling error, assumed $N(0, V)$, where V is the sampling error covariance matrix.

The error terms, w and e, were assumed independent, and V was assumed to be known. The observations were the adjustment factors for the 1,392 poststrata. The model was fit separately for the four Census regions, and for the American Indian Reservation poststrata.

The variables used to form the poststrata were used as predictors, together with a few measures of Census taking difficulty. There were indicators for race (Black, Asian), Hispanic Origin, age category, tenure, Census division, and place/size category. Interactions were allowed between race and place/size, between age-sex-race, and between age-sex-tenure. These variables were expressed as indicators, or if categories were combined, as proportions. Several variables expressed degrees of census-taking difficulty. The proportion of people enumerated on questionnaires returned by mail measured public cooperation with the Census. The proportion of Census whole-person-substitution measured the extent to which the Census relied on imputation. Another variable indicated the proportion of enumeration conducted using traditional door-to-door enumeration, a method used primarily in remote rural areas.

In determining the explanatory variables that were to comprise X, indicators for race, age, tenure were forced to enter the model, while the other variables were selected based on their predictive power. The explanatory variables were selected using a best-subsets regression (Furnival and Wilson 1974). This approach was chosen over more subjective approaches to meet the requirement of prespecification. For any subset of carrier variables, X, an iterative procedure was used to estimate β and σ^2. That is, given an estimate of σ^2, we can compute $\hat{\Sigma} = (V + \hat{\sigma}^2 I)$ and the generalized least-squares estimate:

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

Then, through a maximum-likelihood estimation procedure, we obtain a new estimate of σ^2. The process must iterate until the estimates converge. The smoothed adjustment factors $\hat{y}$ are computed as

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

Were there no covariances in V, this would be equivalent to adding back to the regression estimate a part of the residual, with the part being proportional to the model variance and inversely proportional to the sampling variance. Since covariances are involved, the actual smoothed factor can lie outside the interval between the observed and regression adjustment factor. As a final step, the smoothed factors were ratio-adjusted so that for each regional estimate of total, the smoothed undercount equalled the directly-estimated undercount (using Y only).

Experience from earlier tests and theoretical considerations suggested that the sample estimated variances would be higher for large or very small estimated adjustment factors. If the sample estimated variances were related only to the true adjustment factors, this would have been appropriately accounted for in the generalized least-squares fitting and smoothing. However, it was likely that the sampling errors of the estimated variances would be related to the sampling errors in the estimated adjustment factors. This could have resulted in under- or over-weighting of certain factors. For this reason, and also in an effort to "borrow strength" to improve the sampling error variance estimate, the variances were pre-smoothed using the model:

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

where

v_i	=	True variance of the raw adjustment factor
n_i	=	P sample number of people in the i^{th} poststratum
CV_i	=	Coefficient of variation of the P-sample person weights

W_i	=	A crude regression approximation to the adjustment factor, constrained to be at least 1.00
AI_{1i}	=	Age indicator for ages 0 to 19
AI_{2i}	=	Age indicator for ages 20 to 44
Min_i	=	Variable indicating the proportion of minority in the i^{th} poststratum.

The term "W_i" is included to account for the correlation between the true variance and the true adjustment factor. It is estimated using the same carrier variables and best-subsets regression program as the final estimates but without iteration and, of course, using the sample estimated variances.

The variance model was fit by region using least square with weights inversely proportional to the square root of n_i. Coefficients with t -statistics less than two were set to zero and the model refit. This model seemed to work fine in pulling up the low variances. However, for points with high sample variances, it predicted much lower variances. Using these model variances in the factor regression would have given these extreme points great weight. To lessen this problem, any point with a Studentized residual greater than four (4) was omitted from the modelling and the sample estimated variance was used. Two iterations were used to identify outliers. The original correlations were used with the pre-smoothed variances to compute covariances. A further description of the smoothing process is found in Isaki *et al.* (1991).

The estimated undercount were distributed geographically below the poststratum level by multiplying the poststrata adjustment factors and Census counts for each poststrata in each area. The Census counts for groups excluded from the PES frame, *e.g.* the institutional population, remain unchanged.

4. ALTERNATIVE APPROACHES TO SMOOTHING

Our experience in fitting the regional models to the full 1,380 poststrata (excluding the 12 Indian strata), lead us to experiment with fewer poststratum. Specifically the 115 non-Indian poststrata groups were retained, but the twelve age sex groups within were reduced to six:

* 0 - 9 Males and Females
*10 - 19 Males and Females
*20 - 44 Males
*20 - 44 Females
*45 - 64 Males and Females
*65 and Over Males and Females.

These groupings combined age and sex categories where the true undercount were relatively similar.

Using fewer poststrata had several advantages. A larger sample size for each poststrata would lead to better estimates of the variance/covariance matrix. Also, with fewer poststrata, we could fit the model nationally, thus borrowing strength across region. We felt that this was a particular advantage for some minority groups, which had relatively few observations (poststrata) in some of the regions. For example, the Asian and Pacific Islander strata in the Northeast; Hispanic poststrata in the Northeast and Midwest and Black poststrata in the West.

The first alternative model that was fit used all 690 (115 x 6) poststrata. We refer to this as the national model. The explanatory variables available for modeling were substantially the same as in each Regional model. One difference in fitting the National model is that we did not need to pre-specify any "must" variables. In addition, we created several explicit regional variables to reflect differences that were implicit in the regional fitting.

The development of the national model and the smoothed adjustment factors followed along the lines described earlier for any region except in two respects. First, as mentioned above, a best subsets regression procedure was used to select the explanatory variables for regression with *only* the intercept variable forced into the model. A comparison of the estimated σ^2 using the best subset versus using all available explanatory variables indicated that our estimated σ^2 did not suffer from a severe bias. Second, instead of ratio adjusting the smoothed factors to a single U.S. total, we ratio adjusted smoothed factors to their respective U.S. minority and nonminority totals.

We first smoothed the estimated sampling variances using the model described in section 3. This process identified 10 raw variances as outliers that were not smoothed. All 10 outliers exhibited positive studentized residuals exceeding 4.0. We stopped the outlier identification process after two iterations Application of the best subsets procedure resulted in $\hat{\sigma}^2$ = .000285 and a total of 14 explanatory variables selected. Upon smoothing the adjustment factors the ratio adjustments for U.S. minority and nonminority were 1.00692 and 1.00017, respectively. In addition, six raw adjustment factors when compared to the regression predictor exceeded three times the standard error of the predictor and were trimmed to the boundary. When compared to the regional models the most striking differences occur in the $\hat{\sigma}^2$ and the number of explanatory variables in regression. The average $\hat{\sigma}^2$ for the four regional models was about .00055 with an average 20 explanatory variables. With the expected lower sampling variances and the resulting $\hat{\sigma}^2$ for the National model, a marked reduction of the standard errors of the smoothed factors was obtained.

An alternative approach was to divide the 115 poststrata groups into the 66 non-minority poststrata and the 49 minority poststrata, and to fit separate models to each. Implicit in this model is the assumption that minorities act more like other minorities, than they act like their non-minority neighbors. We call this the split model.

Again, we followed the same procedures in fitting the minority and nonminority adjustment factors separately. We ratio adjusted the smoothed adjustment factors to their respective U.S. totals obtained via the raw adjustment factors. When applying the combined set of smoothed adjustment factors and estimating their covariance an approximation that minority and nonminority factors were uncorrelated was made. This approximation was not felt to be too severe as less than 5% of the raw correlations were found to exceed .25 in absolute value. In minority modelling, smoothing the sample variances resulted in five outliers. Applying best subsets regression 11 explanatory variables were selected for the regression model with $\hat{\sigma}^2$ = .000558. The outlier check on the raw adjustment factors relative to the regression resulted in four trimmed factors. Comparison of the standard errors of the minority factors with the National model indicated that the standard errors of the minority factors under the split models were smaller. The ratio adjustment factor was 1.00658 and nearly identical.

For nonminority modelling we identified six variances as outliers in variance smoothing. The best subset regression yielded eleven explanatory variables with $\hat{\sigma}^2$ + .000118 which was about one-half the $\hat{\sigma}^2$ of the National model. Two raw adjustment factors were trimmed. When comparing the standard errors of the smoothed factors with those from the National we observed that the standard errors for the National tended to be higher. The ratio adjustment factor was 1.00036 and again nearly the same as used in the National model. The differences in the $\hat{\sigma}^2$ under the current model suggested that a better error structure to be used in the National model is one that assumes separate $\hat{\sigma}^2$ for minority and nonminority adjustment factors. Such an investigation is in progress.

Table 2 gives the undercount results for the 116 poststrata groups of these four approaches: the Divisional (unsmoothed) estimates, the Regional (1,380) estimates, the National (690) estimate and the Split minority/non-minority estimates. The advantages of the two alternative approaches can be seen by looking at a few of the poststrata groups. Looking first at non-minorities the Regional estimate for New England Central City was an unexpected 1.16 percent *overcount*. Both the alternatives change this to virtually no net error. Similarly, the Regional model measured almost a one percent overcount for "Other Areas" in the East North Central. The National model changes this into no net error, while the separate non-minority model predicts a small undercount.

The relative advantages of these approaches are greater when we turn our attention to the minority poststrata. Within any region, there were fewer minority than non-minority poststrata, and these were estimated with higher variance. Further, since the minority poststrata were further divided into Black, Non-Black Hispanic and sometimes Asian, there was often little "strength" to borrow, at least from members of the same race/ethnic group.

For example in the Northeast Region, there were only two separate Hispanic poststrata groups with an additional three groups where Hispanics were combined with Blacks. The two separate groups were New York City and Other Large MSA central cities. Their direct (Division) estimates were 4.0 percent and 9.91 percent. Both were estimated with high standard errors: 3.81 and 6.07, respectively. The Regional Smoothing reduced these estimated undercounts to 1.73 and 2.01, *i.e.* below the national average. The National and the Minority models

were able to borrow strength from Hispanics in other regions and produced estimates of 5.5 and 5.2 (National) and 6.5 and 6.6 (Minority). The new models were able to avoid an even worse anomaly in the Midwest. The "direct" estimates of undercount for Hispanics in large MSA's (including Chicago and Detroit) in the East North Central was 0.38. The regional estimate was a 1.6 percent *overcount*. While this outcome is possible, most people would find the alternative estimates of 4.0 and 3.5 more probable.

The West region provides a final example. The "Direct" (Divisional) estimate for Blacks in Non Central city of MSA's was 14.3 percent, among the highest measured. The regional smoothing raises this estimate to 16.4 percent. The alternative smooths lowered the estimates to a more acceptable 6.6 and 8.8 percent.

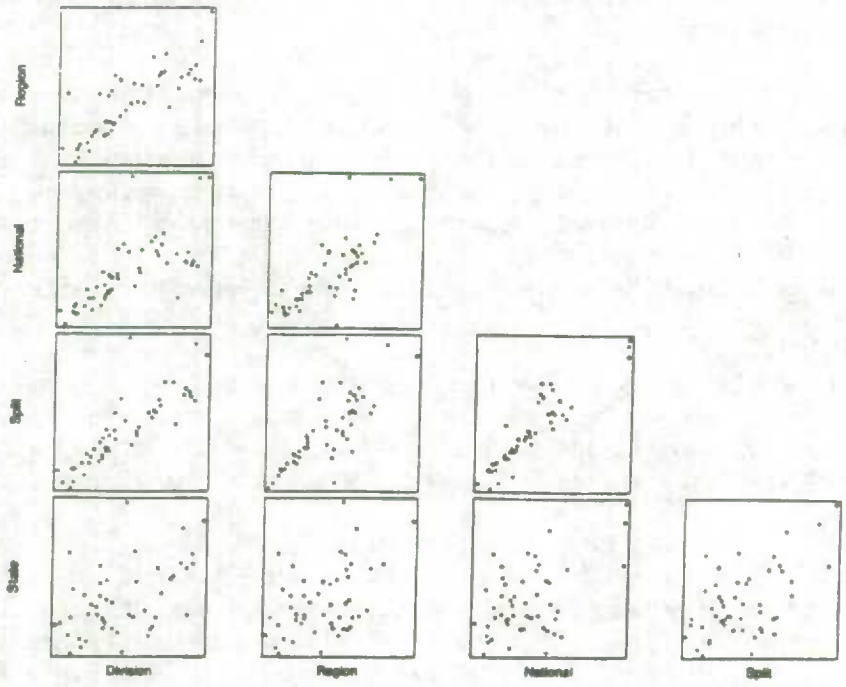
However, many of the anomalies occur precisely because the poststrata represent numerically small groups. The real question is how much difference do they make when aggregated to levels of political geography.

5. EFFECT ON STATE, COUNTY AND PLACE ESTIMATES

The PES sample was allocated to support the "Regional" set of synthetic estimates described in Section 4. An early design decision was not to allocate the sample so as to produce direct state estimates with acceptable variances. However, one can produce a set of direct estimates using only the sample in each state. We refer to such an approach as the state model. These state model estimates have large sampling variances that make them inappropriate for most uses. However, it is useful to compare them with the indirectly derived state estimates using the four sets of adjustment factors provided in Table 2.

Figure 1 plots the percent undercount for States by each of the five approaches. Several results stand out. The Direct State approach shows little correlation with the four other models. This, of course, may reflect true differences in undercount between states that are not picked up when aggregating across state. We believe that the difference is more likely attributable to the sampling variance of these estimates. Indeed, the differences between the Direct State estimates and the Divisional estimates were only beyond sampling error for three states Montana, Idaho and Washington State. Since these three states were all grouped with California as part of the West, it is quite possible that the data do exhibit some differences in real undercount.

Figure 1: Percent Undercount of States by Five Models



The Division and Region models do show greater agreement. Still, many points lie off the diagonal. The unsmoothed estimates were highest in the South Atlantic. The undercount for the East South Central and West South Central were much lower. The Regional Smoothing brought all three divisions together, with the West South Central now being slightly higher. That is, smoothing lowered the undercount for the South Atlantic by 0.9 percentage points and raised the undercount by 1.2 and 0.8 in the East and West South Central Divisions. The smoothed estimates were 427,000 lower than the unsmoothed for the South Atlantic, and 191,000 and 222,000 higher for the East South Central and West South Central, respectively.

These same divisions show up again as we compare the National model and the split model. The states generally fall on the diagonal. However, a cluster of points in the middle lie clearly off the diagonal, with the estimate from the Minority/Non-minority split model clearly being higher than for the National model. These points turn out to be the eight states from the South Atlantic Division. Again, the indicator for South Atlantic entered the Non-minority regression, but did not enter in either the South Regional model or for the National model. The points where the split model estimates a lower undercount than the National model do not show such a strong regional clustering: Connecticut, Massachusetts, Rhode Island fall into this group, but so do Oregon and Washington. The three points clumped together with high undercounts are California, Hawaii, and New Mexico. The two estimates for Hawaii are in almost perfect agreement (3.57 and 3.59) with each other. Each reflects a dramatic change from the production Regional estimate of 2.47 percent.

Figure 2 plots the percent undercount estimates for each of the four models in Table 2 for the 458 counties with population above 10,000. Direct county estimates were not made, so there are only four models to compare. Clearly, the results from the four models are highly correlated, but with important exceptions. The Division model shows less agreement with the other three than they do with each other. Again, investigating the off diagonal points shows that many of the differences between models can be attributed to counties in the South Atlantic Division.

Figure 2: Percent Undercount of Counties by Four Models

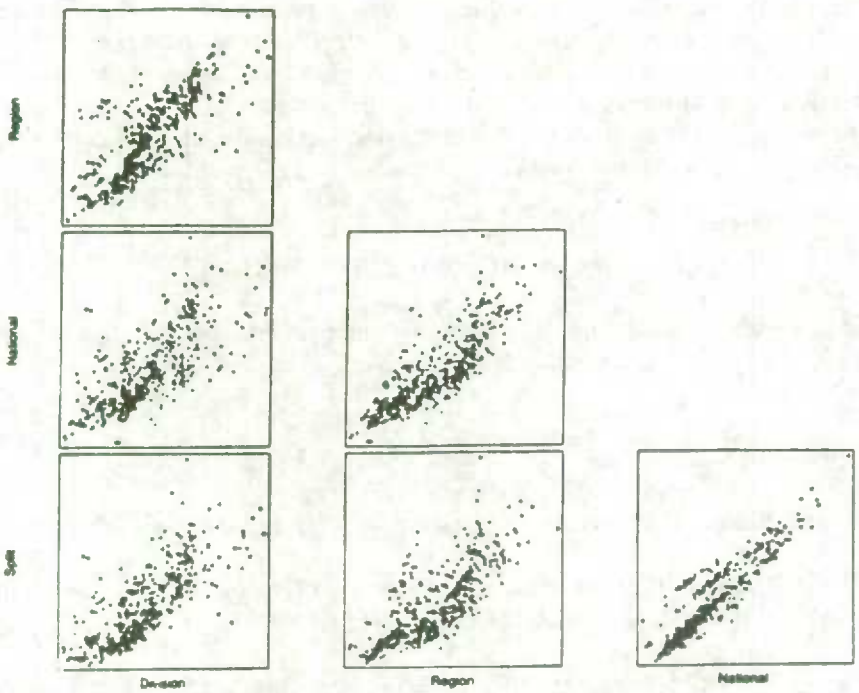
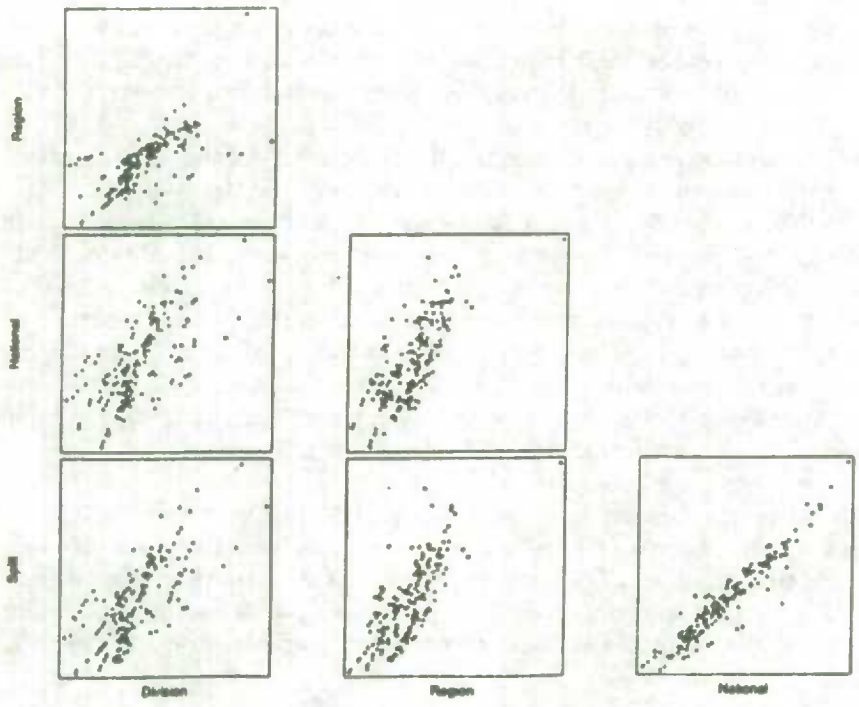


Figure 3 plots undercount estimates for the 195 places with population above 10,000. The extreme outlier in the Regional model is Ingelwood, California. This city had an estimated undercount of 11.15 percent. Both alternative models reduce the estimated undercount considerably: to 6.94 for the National Model and to 6.05 for the Minority/Non-minority Split.

Figure 3: Percent Undercount of Places by Four Models



In this paper, we have investigated the effect of modeling the undercount for alternative groupings of poststrata. Overall, the results are similar. All are, of course, based on the same data set. However, for some states, counties, and places, the different approaches imply quite different undercounts. The National model and the Split Minority/Non-minority models each seem to offer significant advantages over the regional model. Both remove many of the troublesome anomalies that are present in the Regional model. The choice between the National model and the Split model is less clear. The ability of the Split model to detect divisional differences within race/ethnic group when they are present seems to be an important advantage. Still, at this point, neither model has clearly been shown to be superior.

6. ACKNOWLEDGEMENTS

We wish to thank Elizabeth Huang and Jule Tsay for producing the smoothed factors in all of the models used and Fred Navarro for producing the estimated undercount.

REFERENCES

Furnival, G.M., and 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., and 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., and Huang, E.T. (1988). On estimating census undercount in small areas. *Journal of Official Statistics,* 4:2, 2, 95-112.

Table 2: 1990 Post-Enumeration Survey Dual System Estimates
Percent Undercount by Post Stratum Group

	Division	Regional	National	Split	Division			Region			National			Split		
		Non-Minority			Black	Hisp.	Asian	Black	Hisp.	Asian	Black	Hisp.	Asian	Black	Hisp.	Asian
North East																
New England																
Central Cities	-1.74	-1.16	0.26	-0.07	5.69			4.25			5.76			3.95		
Non Central City MSA	0.61	0.19	0.80	0.57												
Other Places 10,000+	0.54	0.59	1.22	1.03	5.88 *			5.39 *			4.60 *			5.75 *		
Other areas	1.68	1.79	1.79	2.00												
Middle Atlantic																
New York City CC's																
Renter	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
Owner	-2.64	-0.23	-0.33	-0.76	-2.86			-0.15			1.07			2.27		
Other Large MSA Central city																
Renter	-6.41	-0.37	2.26	2.26	10.78	9.91		7.74	2.01		8.02	5.23		9.36	6.59	
Owner	-2.93	-0.19	-0.45	-0.65	2.66			-0.03			1.68			3.12		
Central cities of Small MSA	2.05	0.07	1.02	0.67	17.92			9.34			6.55			8.38		
Non Central City in NYC PMSA	5.03	0.42	1.12	0.52	5.63			6.73			4.67			5.34		
Non Central City in Large MSA	-0.80	0.36	0.70	0.48												
Non Central City in Small MSA	-0.78	-0.09	0.41	0.29	5.88 *			5.39 *			4.60 *			5.75 *		
Other Places 10,000+	1.36	0.41	1.39	0.81												
Other areas	0.43	0.70	0.98	0.87												
South																
South Atlantic																
Large MSA Central city																
Renter	11.49	5.00	3.78	4.05	10.46			9.33			7.64			7.95		
Owner	1.09	1.72	-0.17	0.84	1.68	2.77		0.95	4.92		1.55	3.87		2.13	3.90	
Central cities of Small MSA	2.84	2.74	1.86	2.18	4.93			4.00			5.34			5.67		
Non Central City in Large MSA	0.93	0.44	0.52	1.83	4.17	13.79		1.97	5.13		4.83	3.55		5.30		
Non Central City in Small MSA	3.50	2.80	1.38	1.97	0.27			3.59			4.14			4.52	3.69	
Other Places 10,000+	1.23	1.51	0.79	1.77	-1.71			1.60			3.06			3.66		
Other areas	3.25	2.71	1.23	1.79	5.68			2.64			2.09			3.36		

* Indicates cell combined across division.

	Division	Regional	National	Split	Division			Region			National			Split		
	Non-Minority				Black	Hisp.	Asian	Black	Hisp.	Asian	Black	Hisp.	Asian	Black	Hisp.	Asian
East South Central																
Large MSA Central city																
Renter	2.17	4.80	2.81	2.60	6.46			5.81			5.22			6.11		
Owner	3.19	2.56	0.45	0.06												
Central cities of Small MSA	0.90	2.58	0.80	0.62												
Non Central City in MSA	1.42	2.31	1.11	0.83	4.82			2.26			2.89			3.85		
Other Places 10,000+	-6.02	1.84	1.01	0.64												
Other areas	-0.95	1.65	0.09	0.16												
West South Central																
Houston, Dallas, Ft Worth CC's																
Renter	6.24	4.60	3.31	2.82												
Owner	0.56	1.49	0.34	0.18	8.09	8.96		6.64	7.11		5.60	5.03		5.98	5.44	
Other Large MSA Central city																
Renter	1.34	3.23	3.06	2.76												
Owner	-1.16	0.69	-0.10	0.04	4.54	3.18		4.82	3.76		4.86	3.27		5.31	3.70	
Central cities of Small MSA	-3.16	2.48	1.06	0.83												
Non Central City in MSA	2.07	2.28	1.21	0.99												
Other Places 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	
Other areas	1.72	1.96	1.09	0.97												
Midwest																
East North Central																
Chicago Detroit CC's																
Renter	2.76	5.17	3.32	2.31	6.76	0.38		5.77	-1.61		7.64	4.04		7.02	3.52	
Owner	-0.05	1.12	0.34	-0.25	0.42			1.98			1.79			2.16		
Other Large MSA Central city																
Renter	1.56	1.04	1.95	2.30	4.03			4.49			7.64			7.33		
Owner	-1.24	-0.15	0.84	0.41	7.09			0.64			1.92			2.77		
Central Cities of Small MSA	1.76	2.09	1.35	0.71	4.61			5.44			4.59					
Non Central City in Large MSA	0.84	0.59	0.64	0.62										5.06		
Non Central City in Small MSA	0.96	0.64	0.55	0.62	3.99	*		4.66	*		4.71	*		5.03	*	
Other Places 10,000+	0.42	0.20	0.78	0.62												
Other areas	-1.64	-0.99	0.01	0.65												

	Division	Regional	National	Split	Division			Region			National			Split		
	Non-Minority				Black	Hisp.	Asian	Black	Hisp.	Asian	Black	Hisp.	Asian	Black	Hisp.	Asian
West North Central																
Large MSA Central city																
Renter	5.20	2.47	3.05	2.83	5.47			5.44								
Owner	-0.53	-0.33	-0.05	-0.26							5.27			5.45		
Central cities of Small MSA	1.82	1.90	1.37	0.81	4.85			7.23			4.59					
Non Central City in Large MSA	1.09	0.71	0.92	0.85										5.67		
Non Central City in Small MSA	0.22	1.64	0.75	0.95	3.99	*		4.66	*		4.71	*		5.03	*	
Other Places 10,000+	0.83	0.75	1.00	0.73												
Other areas	0.78	0.31	1.05	1.45												
West																
Mountain																
Large MSA Central city																
Renter	4.65	5.03	3.69	2.86	1.48			4.61			4.08			4.29		
Owner	1.24	0.98	0.23	0.04												
Central cities of Small MSA	2.88	1.52	1.57	1.08												
Non Central City in MSA	0.60	0.75	0.58	0.79	7.39	*		7.80	*		4.65	*		4.68	*	
Other Places 10,000+	1.22	1.45	0.99	0.81												
Other areas	3.00	3.22	2.46	2.80												
Pacific																
Renter																
Los Angeles/Long Beach CC's	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
Other Large MSA Central city	3.73	3.72	4.31	3.62												
Central cities of Small MSA																
Owner																
Los Angeles/Long Beach CC's	-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
Other Large MSA Central city	1.39	1.39	0.89	0.40												
Central cities of Small MSA																
Central cities of Small MSA	0.56	0.95	1.38	0.95												
Non Central City in Large MSA	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
Non Central City in Small MSA	2.90	3.15	2.26	1.30												
Other Places 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
Other areas	3.15	1.92	2.56	2.51												
Reservation Indian							12.72			12.72			12.72			12.72

SESSION 2

Spatial Considerations in the Design of Surveys or Frames

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

CONSTRUCTION OF SPATIALLY ARTICULATED LIST FRAMES FOR HOUSEHOLD SURVEYS

A. Saalfeld[1]

ABSTRACT

This paper presents and compares some new and some old methods for ordering spatial entities to reflect spatial proximity. It then describes how those ordered entities may constitute list frames for use in systematic sampling. The new methods derive from methods for ordering vertices or edges in a connected acyclic graph, using an Eulerian tour to assign a cyclic order. The new ordering methods are coordinate system independent and rotation invariant. The new ordering methods do not depend on prior bucketing of space as do some standard existing methods. The new orderings depend instead on the spatial distribution of the input data and on the metric of the underlying space.

The new orderings are called *tree-orders*. They can be constructed in linear time from topo-logical data structures. They are fully characterized by a useful proximity-preserving property called *branch-recursion*. The paper describes how *tree-ordering* techniques can be applied to find orderings for fundamental spatial elements of list frames for use in multistage sampling:

- Ordering planar point data such as housing units.
- Ordering area data such as blocks or regions.
- Ordering line segment data such as street segments.
- Ordering pseudo-line segment data such as block faces.
- Ordering any of the above spatial entities to respect geographic hierarchies.

For each of the spatial entities listed above, the orderings produced by extending the tree-ordering methods exhibit important proximity-preserving properties. The paper describes sampling applications of the new orderings of the diverse spatial objects.

KEY WORDS: Spatial data; Linear orders; Trees; Proximity.

1. INTRODUCTION

It is useful, and sometimes necessary, to order data. The traditional series computer (in contrast to some modern parallel machines) uses a linear paradigm. It requires data to be read in some order and places those data in consecutive cells of memory that may be recovered in groups or blocks at a time. At its most elementary level, database management is the art of organizing or ordering data so that they may be accessed and utilized most efficiently for some particular set of operations of interest. This paper presents a new way of ordering data that will permit a collection of important operations to be carried out efficiently and effectively. Those operations include both the general computer data management operations and some important survey sampling operations such as cluster sampling and systematic sampling. In this section we review and summarize some definitions and basic concepts needed to describe our ordering techniques.

---

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

1.1 Ordering and Lists

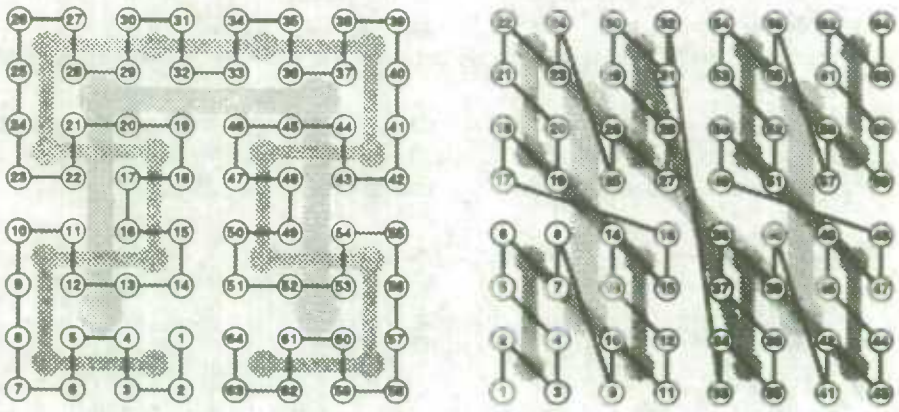
Throughout this paper, *ordering* or *order*, without any qualifying adjectives, will refer to a total order or linear order to a finite set of n elements. Such an order is nothing more than a sequencing of the n elements, a one-to-one association of the elements, a one-to-one association of the elements of the set with the integers 1 through n, or a listing of the n elements. A set of n elements that have been ordered will be called a *list* or an *ordered list*.

In a list of n elements, the $(i + 1)$st element is the *successor* to the i th element; and every element except the nth or last element has a unique successor. Similarly, every element except the first element has a unique *predecessor*. We may build a *cyclic list* or a *cyclic order* from a list by naming the first element to be the successor to the last element (and the last element to be the predecessor to the first). Cyclic lists are often useful because they have no distinguished elements that require special case handling. For example, with a cyclic list, one may begin *anywhere* in the list and exhaustively enumerate elements by taking successors until one returns to the chosen starting element.

1.1.1 Limitations of Linear Ordering Schemes

It is impossible to capture all of the complexity of a two-dimensional structure in a one-dimensional representation. There is, for example, no continuous bijective mapping from a line to a plane which has a continuous inverse. Similarly, there can be no distance preserving mapping from a full rectangular lattice of points to a set of points on the line. We can, however, find bijective mappings from the integers 1 through 4^m onto points on a 2^m by 2^m rectangular lattice which preserve adjacency. (One such mapping is called a Hilbert curve, illustrated in figure 1). Of course, the inverse mapping cannot ever preserve *all* adjacencies since each interior lattice point has 4 neighbors and each interior integer has 2 neighbors on the line. The applications section 3 will point out some of the other properties we can or cannot always maintain with other types of spatial data when we employ a *linear* ordering.

Figure 1: Quadrant-Recursive Orderings: Hilbert Curve (left) and Peano Key



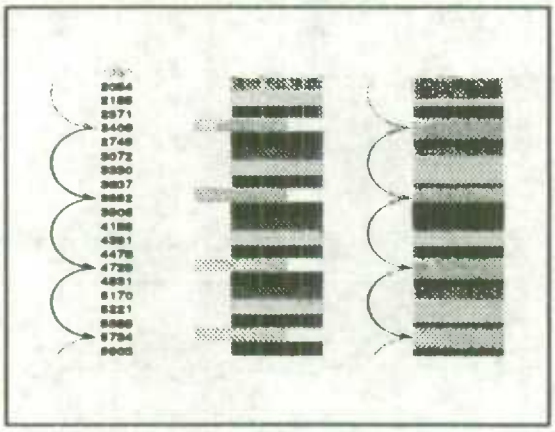
1.2 Spatial Queries

Points in two-dimensional and higher dimensional space are often assigned an order or primary key to facilitate their storage in and retrieval from databases. Space-filling curves, such as the Peano key and Hilbert curve (Faloutsos and Rong 1989) and (Faloutsos and Roseman 1989), have proved quite useful for range queries and nearest neighbor searches. These curves are instances of a large class of orderings called *quadrant-recursive* orderings (Mark 1990). The defining property of quadrant-recursive orderings is that, in any recursive decomposition of a rectangular region into subquadrants, the points of any subquadrant always appear consecutively in the quadrant-recursive ordering. Points within any subquadrant are enumerated exhaustively before exiting the subquadrant. We will see in section 2.2.4 that quadrant-recursive orderings are a spacial case of a more general class of orderings called *branch-recursive*.

1.3 Systematic Sampling

Systematic sampling traditionally refers to selection of a subset from a list, where the subset is formed by selecting elements at regular intervals (called the *skip interval*) (Kish 1965). Elements may be *weighted* to adjust their probability of selection (see figure 2).

Figure 2: Systematic Sampling from Lists



If all weights are 1, then a skip interval of k produces a $1/k$ sample. We may think of the sampled elements as having the induced order of their sequenced systematic selection achieved by skipping through the list.

If points in the plane are assigned any quadrant-recursive order, then a systematic selection procedure will sample every subquadrant, no matter what its size, to within one unit of the overall sampling fraction. This representative coverage property was noted and utilized for Peano key ordering by Wolter and Harter (Wolter *et al.* 1989).

1.4 Graphs and Maps

The linework of any map has an underlying structure of a *graph*[2]. We will use the usual combinatorial definitions of graph theory found in the standard text by Harary (Harary 1969). A *graph* $G = (V, E)$ consists of a finite non-empty set V of *vertices* together with a set E of unordered pairs of vertices called *edges*. A vertex v and an edge $\{u, w\}$ are *incident* if and only if $v = u$ or $v = w$. The *degree* of a vertex is the number of edges incident to the vertex. A *walk* of the graph G consists of a sequence $(v_1v_2v_3...v_k)$ of vertices v_i, not necessarily all distinct, such that for each $j = 1,2,...,(k - 1)$, $\{v_j, v_{j+1}\}$ is an edge of G. A *tour* is a walk $(v_1v_2v_3...v_k)$ such that $v_1 = v_k$. A *path* is a walk with no edges repeated. A *cycle* is a path $(v_1v_2v_3...v_k)$ with $k \geq 3$ such that $v_1 = v_k$. A *tree* is a graph with no cycles. A tree as we have defined it is sometimes called a *free tree* to differentiate it from a *rooted tree*, which possesses a distinguished vertex called the *root*.

1.5 Trees

We describe some properties of trees that make them easier to work with than graphs in general. We will show in sections 3.3 and 3.4 how problems of ordering graph components can be converted to problems of ordering tree components for a derived tree. Computer scientists have developed a number of ways of ordering the vertices of rooted trees embedded in the plane (Aho 1985). We will be looking at new orderings for free trees.

---

[2] For some applications it may be preferable and even necessary to regard the linework of a map as a *pseudo-graph*, a structure which allows multiple edges between two vertices. For the applications which we are examining here, the distinction is unimportant.

1.5.1 Combinatorial Properties of Trees

We list some important properties of trees.

Property 1 *Every tree with n vertices has exactly $(n - 1)$ edges.*

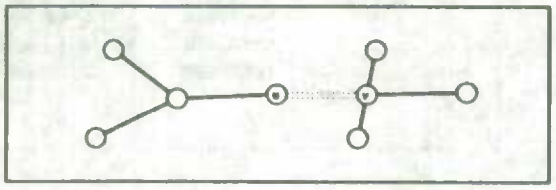
Property 2 *A connected graph having n vertices and $(n - 1)$ edges is a tree.*

Property 3 *Adding a new edge to a tree (between existing vertices) always creates a cycle.*

Property 4 *Removing an edge always disconnects a tree.*

Property 5 *There is a unique path between any two vertices of a tree.*

Figure 3: Edge Removal Creates Two Branches



We say that each edge determines *two branches* (see figure 3) which are the disconnected component subtrees resulting from that edge's removal. Always one of the branches determined by the edge $\{u, v\}$ contains u and the other branch always contains v.

1.5.2 Planar Embeddings of Trees

Not every graph can be drawn in the plane with non-intersecting line-segment edges, but a tree can always be represented or realized as a straight-line drawing in the plane. Moreover, suppose that for each vertex in a tree, we arbitrarily assign a cyclic order to the edges incident to that vertex. Then there is always a drawing in the plane of that tree with straight-line-segment edges such that the clockwise order of the edges incident to any vertex is the arbitrarily assigned cyclic order of the edges about the vertex.

2. EULERIAN TOURS AND TREE-ORDERINGS

2.1 Eulerian Tours

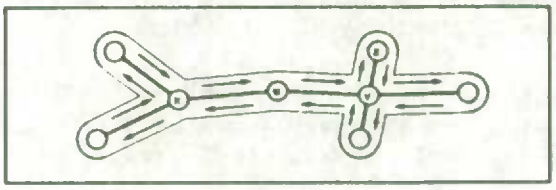
An Eulerian tour of a tree is a special well-balanced tour that traverses every edge exactly twice, once in each direction. We give two equivalent descriptions of an Eulerian tour of a tree. *Each description depends on our having assigned a cyclic order to the incident edges of each vertex.*

2.1.1 Geometric Version

Draw the tree so that the assigned cyclic order of edges at each vertex is the clockwise order. Start a tour at any vertex x. Depart along any incident edge $\{x, u\}$ toward u. Upon arriving at u, depart along the edge $\{u, v\}$ that is next to $\{x, u\}$ in the clockwise order around u. Upon arriving at v, depart along the edge $\{v, z\}$ that is next to $\{u, v\}$ in the clockwise order around v, *etc.*, until finally you return to x along the edge that precedes $\{x, u\}$ in the clockwise order (see figure 4).

The tour that we have described will traverse each edge twice, once in each direction and visit every vertex a number of times equal to its degree.

Figure 4: Geometric Depiction of Eulerian Tour



This tour can also be visualized as follows: imagine that the tree itself is the top view of a wall. Walk next to the wall with your right hand continuously touching the wall. You will eventually return to your starting point, at which time you will have touched both sides of every wall. Thus, had someone else been walking on top of the wall and keeping up with you, that person would have walked every edge of the tree exactly twice.

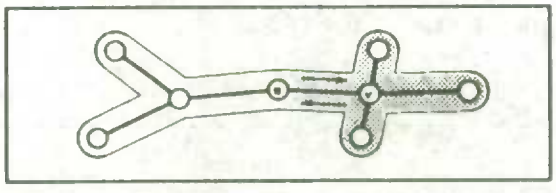
2.1.2 Combinatorial Version

An Eulerian tour of a tree on n vertices is a walk $(v_1 v_2 v_3 \ldots v_{2n-1})$, where the v_i's are vertices, clearly not all distinct, satisfying

1. $v_1 = v_{2n-1}$.
2. For $i = 1, 2, \ldots, 2n - 3$, $\{v_{i+1}, v_{i+2}\}$ is the successor to $\{v_i, v_{i+1}\}$ in the cyclic ordering of edges about the vertex v_{i+1}.
3. $\{v_1, v_2\}$ is the successor to $\{v_{2n-2}, v_1\}$ in the cyclic ordering of edges about the vertex v_1.
4. For every edge $\{u, v\}$ of the tree, the sequence uv and the sequence vu each appear exactly once in the walk $(v_1 v_2 \ldots v_{2n-1})$.
5. Each vertex except v_1 appears as often as its degree.
6. The vertex v_1 appears one more time than its degree.

Notice further that if uv appears before vu in the Eulerian tour, then the subwalk between uv and vu that starts and ends at v completely consumes every edge and vertex in the v-branch determined by edge $\{u, v\}$. Moreover, this subwalk touches nothing but the v-branch of the tree (see figure 5).

Figure 5: A Subwalk Consumes an Entire Branch



Similarly, the remainder of the tour (the part before uv and after vu) completely consumes every edge and vertex of the u-branch resulting from the removal of the edge $\{u, v\}$. Clearly, no vertex or edge that appears in the v-branch can ever appear in the u-branch.

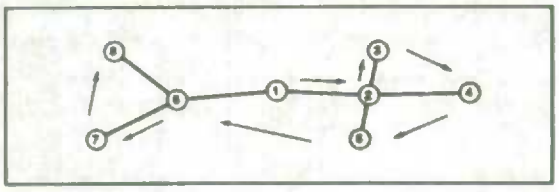
2.1.3 Eulerian Tree-Orderings

During the course of an Eulerian tour, all of the vertices and edges are visited at least once. Suppose we wish to assign the integers 1 through n to our n vertices. A procedure that visits the vertices in Eulerian tour order, assigning either the next available number or no number to every *visit* of each vertex, in such a way that exactly one of the visits of each vertex receives an order number, will be called an *Eulerian tree-ordering* (ETO) of vertices.

A procedure that visits the edges in Eulerian tour order, assigning either the next available number of no number to every *visit* of each edge, in such a way that exactly one of the two visits of each edge receives an order number, will be called an *Eulerian tree-ordering* (ETO) of edges.

Garey and Johnson (Garey 1979) and others (Preparata 1985), (Edelsbrunner 1987) describe one such Eulerian tree-ordering of vertices of a Euclidean minimum spanning tree (EMST) obtained by starting anywhere on the Eulerian tour and assigning the next available number to the *first* visit to each and every vertex (see figure 6).

Figure 6: First-Visit Eulerian Tree-Ordering



Their ordering, or for that matter, any other Eulerian tree-ordering of a EMST will always approximate a Euclidean Travelling Salesman Tour to within a factor of 2, since the Eulerian tour itself is never more than twice the length of the Euclidean Travelling Salesman Tour.

2We now describe a simple procedure for generating all other Eulerian tree-orderings of the vertices of a tree.

2.2 Tree-Ordering Vertices

Suppose that we are given a tree *and an Eulerian tour* $(v_1 v_2 \ldots v_{2n-1})$ for that tree (equivalently we are given an embedding of the tree in the plane and a starting vertex and edge). Then to order the vertices we will proceed a2s follows.

2.2.1 Setup: Weighting Vertex Visits

For $i = 1, 2, 3, \ldots, (2n - 1)$, regard each v_i that appears in tour $(v_1 v_2 \ldots v_{2n-1})$ as a vertex visit.

For $i = 1, 2, 3, \ldots, (2n - 1)$, assign a non-negative weight w_i to the i th visit so that the sum of weights for all visits to any fixed vertex v is one:

$$\text{For every } v \in V, \sum_{\{i | v_i = v\}} w_i = 1.$$

We call any such weight assignment a *unit-sum weight assignment.*

An important instance of a unit-sum weight assignment assigns the same weight to all visits to the same vertex. Because each vertex v_i is visited $\deg(v_i)$ times[3], that uniform weight is exactly given by:

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

$$\text{and} \qquad w_{2n-1} = 0.$$

---

[3] Because the Eulerian tour is cyclic, we like to count v_1 and v_{2n-1} as the same visit. We should only assign the appropriate weight to one or the other. We have chosen to assign the uniform weight to v_1.

2.2.2 Building the Sampling Interval

As we walk the Eulerian tour, we begin accumulating weights, (exactly as is done to build a weighted list for systematic sampling).

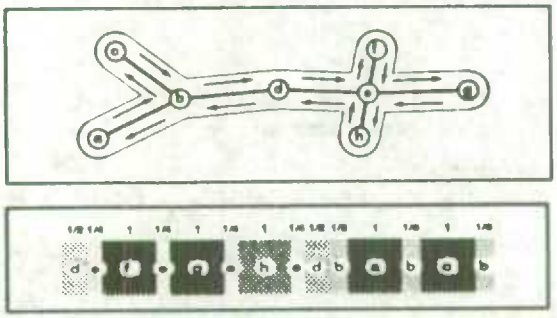
$$\text{Let} \quad W_o = 0,$$

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

2.2.3 An Illustration: Uniform Weighting

We illustrate the accumulating of weights for the uniform weighting scheme for (*defegehedbabcbd*), the walk drawn in figure 7.

Figure 7: A Tour and its Accumulated Weights

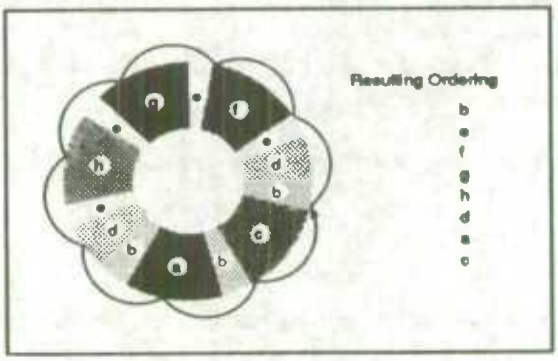


The total accumulated weights are exactly n and the total weight corresponding to each vertex of the tree is exactly one. We can assign numbers to the vertices by skipping through the weighted interval with skip interval equal to 1. This is the same as assigning an order number to a vertex each time a vertex visit takes us up to or past the next whole integer:

If $\lfloor W_{j-1} \rfloor < \lfloor W_j \rfloor$, then assign vertex v_j the number $\lfloor W_j \rfloor$.

Because some vertices appear in several places in our accumulated weighted interval, one may suspect that our numbering scheme may assign more than one order number to a vertex. This cannot happen.

Figure 8: Cyclic Ordering with Uniform Weighting



Because the Eulerian tour is cyclic, we can make our cumulative interval cyclic and our resulting ordering cyclic as well by removing the dependence on the starting point of the tour when building our cumulative vertex-visit weight interval, as shown in figure 8. The following theorem, proved by (Saalfeld 1991), assures that our numbering procedure gives and ordering of vertices.

Theorem 2.1 *While making an Eulerian tour of a tree, build a separate (cyclic) interval of total length* n *units by assigning a non-negative weight to each vertex visit in any way so that the total weight for all visits to any individual vertex is one. Then every vertex gets hit exactly once by skipping one unit at a time through the cyclic* n*-interval.*

2.2.4 Branch-Recursion

Throughout this section we will regard the orderings generated by our ordering procedure as cyclic by making the first vertex successor to final vertex.

Corollary 2.2 *The collection of vertices of* **any branch** *of the tree always constitute a complete interval (i.e. appear consecutively) for any cyclic Eulerian tree-ordering.*

We will say that any cyclic vertex ordering that keeps vertices of a branch together for all branches is a *branch-recursive ordering.* Corollary 2.2 states that every Eulerian tree-ordering is branch recursive. The converse is also true.

Theorem 2.3 *Every branch-recursive cyclic ordering of the vertices of a tree is an Eulerian tree-ordering for some Eulerian tour of the tree and some unit-sum weight assignment to the vertex visits of that Eulerian tour.*

Branch-recursion constitutes a very strong proximity preservation property, where proximity is measured by the link-distance in the tree or graph. Branches of a tree may correspond to data clusters in cases where we have built minimum spanning trees. All quadrant-recursive orderings of a point set in the plane may be realized as orderings induced on the leaf subsets of branch-recursive orderings (*i.e.* Eulerian tree-orderings) of the quadtree of those points.

2.3 Tree-Ordering Edges

Our methods for ordering vertices are equally valid for edge ordering. Theorem 2.1 for vertices has an exact counterpart for edges:

Theorem 2.4 *While making an eulerian tour of a tree, build a separate (cyclic) interval of total length* $(n - 1)$ *units by assigning a non-negative weight to each edge visit in any way so that the total weight for all visits to any individual edge is one. Then every edge gets hit exactly once by skipping one unit at a time through the cyclic* $(n - 1)$*-interval.*

2.3.1 Uniform Edge Weighting

A *uniform* weighting scheme for edges instead of vertices would have each edge getting weight exactly ½ (since every edge is visited twice in the Eulerian Tour). But giving every edge weight ½ amounts to nothing more than skipping every other edge in our selection procedure. So we have the following corollaries to theorem 2.4:

Corollary 2.5 *While making an Eulerian tour of a tree, number every other edge visited. Then every edge gets exactly one number assigned to it.*

Corollary 2.6 *Edges which are consecutively numbered using a uniform weighting scheme are never more than link distance 2 apart.*

2.3.2 Branch-Recursion

As with vertices, every Eulerian tree-order of edges in branch-recursive in the same sense:

Corollary 2.7 *The collection of edges of* **any branch** *of the tree always constitute a complete interval (i.e. appear consecutively) for any cyclic Eulerian tree-ordering of edges.*

And, conversely,

Theorem 2.8 *Every branch-recursive cyclic ordering of the edges of a tree is an Eulerian tree-ordering of edges for some Eulerian tour of the tree and some unit-sum weight assignment to the edge visits of that Eulerian tour.*

3. BUILDING LIST FRAMES

In this section we will adapt our tree-ordering techniques to order spatial objects. Our approach in every case will be to convert the ordering problem to a tree-ordering problem, then solve the tree-ordering problem by a uniform-weighting of vertices or edges, as appropriate.

3.1 Ordering Point Data: Households

If we assign coordinates to households, regarding them as points in a plane, we may order those points as follows: We know how to order vertices of a tree. So we may convert the points into vertices by building a tree (adding edges); and one natural tree to build is a Euclidean minimum spanning tree (EMST). The EMST is unique if the points are in general position or if no two interpoint distances are equal. So the steps needed to convert the problem of ordering points in space to one of ordering tree vertices are:

ALGORITHM ORDER_PLANAR_POINTS

1. **Build euclidean minimum spanning tree.**
2. **Walk eulerian tour, tree-ordering vertices.**

We can build a Euclidean minimum spanning tree in time $O(n \log n)$ (Aho 1985), sorting the edges at each vertex in clockwise order as they are inserted. the planar embedding of the tree gives us the geometric version of the Eulerian tour for free (*i.e.* the usual clockwise ordering of edges around a vertex). We can then walk the Eulerian tour and order the vertices in $O(n)$ additional time.

3.1.1 A Cluster Sampling Application

Cluster sampling is a survey sampling strategy of selecting small groups (clusters) of neighboring points instead of selecting individual points randomly distributed. Within-cluster correlation may reduce the efficiency of such a strategy from a pure sampling viewpoint, but that consideration is often outweighed by the economic impact of reduced travel costs for interviewers.

A serious limitation to successfully selecting clusters from lists, however, is the fact that proximity in the list does not guarantee proximity on the ground. Selection of points from a list that has been ordered by performing our uniform-weight tree-ordering algorithm on a EMST of the points will guarantee very strong proximity correspondence. The following theorem holds:

Theorem 3.1 *Order points in the plane by building their EMST and applying the uniform-weight vertex tree-ordering algorithm. Then two consecutive points in the order have a maximum link distance of six and an average link distance of less than two.*

Proof: The degree of any vertex in a EMST is less than or equal to six. Thus the uniform-weight tree-ordering algorithm accumulates a weight of a least 1/6 with each vertex visit. Moreover, in any tree, the average degree is $\frac{2n-2}{n}$.

3.2 Ordering Points in Higher dimensional Spaces

To apply the methods of section 3.1 to points in higher dimensions, we must first address two issues: (1) building an EMST in higher dimensions, and (2) defining an Eulerian tour in higher dimensions.

There are straightforward $O(n^2)$ time algorithms for building a EMST in higher dimensions (Aho 1974). Some exact algorithms are known with complexity slightly sub-quadratic (Yao 1982).

Building an Eulerian tour in higher dimensions is not so straightforward. It requires establishing a cyclic order of edges about every vertex. One possibility is to project the edges onto some two-dimensional subspace, then order the projection of the edges clockwise on that plane. Another more canonical approach, suggest by Herbert Edelsbrunner (Edelsbrunner 1990), is to map the edge configuration about a vertex onto points on the surface of a sphere of dimension one less than the space of the EMST, then apply the ordering scheme to those points on the sphere recursively (*i.e.* build their EMST and order them in a space of smaller dimension).

In any case, if all we require is some ordering of the edges around each vertex, we can find one in $O(n \log n)$ time. We summarize the steps needed to convert the problem to a tree-ordering problem.

ALGORITHM ORDER_POINTS_IN_N_SPACE

1. **Build EMST.**
2. **Cyclically order edges at each vertex.**
3. **Walk Eulerina tour, tree-ordering vertices.**

3.2.1 A Sample Stratification Application

Sample stratification is a partitioning of the universe into groups which are similar across several characteristics. The characteristics should be in some sense comparable (dealing with relative incomparability is sometimes known as the Scaling Problem). Stratification is often accomplished by treating the observations as n-tuples of the n characteristics (*i.e.* as points in n-space) and finding a hyperplane or collection of hyperplanes that optimize separation of the points across the half-spaces or n-cells created. A more straightforward approach to stratification (and one that would be computationally much simpler) might be to partition a EMST of the points into branches of greatest separation. With branch-recursive ordering methods, this operation boils down to list splitting! We at the Bureau of the Census will be comparing results of using tree-ordering methods to the standard more complex stratification algorithms.

3.3 Ordering Vertices of any Graph

If we are only concerned with ordering the vertices of a graph, we may think of the graph as a tree with too many edges. So we throw away the least useful edges until we have whittled the graph down to a tree. If the edges have costs associated with them, we may wish to minimize the cost of the resulting tree, for example. We know exactly how many edges to throw away. We will discard an edge as long as it does not disconnect the graph and we still have $(n - 1)$ edges left. We summarize the steps needed to convert the problem to a tree-ordering problem.

ALGORITHM ORDER_GRAPH_VERTICES

1. **Build a (minimum) spanning tree.**
2. **Cyclically order edges at each vertex.**
3. **Walk Eulerina tour, tree-ordering vertices.**

3.4 Ordering Edges of any Graph

In section 3.3, we regarded our graph as having too many edges; and we threw some away. To order the edges of our graphs, we regard our graph as having too few vertices to be a tree; and we add vertices by splitting the vertices of the graph and creating more vertices with the same number of edges (see figure 9). Once again we use our knowledge of the edge/vertex relationship in a tree to know when to stop splitting vertices. We summarize the steps needed to convert the problem to a tree-ordering problem.

ALGORITHM ORDER_GRAPH_EDGES

1. **Split vertices.**
2. **Cyclically order edges at each vertex.**
3. **Walk Eulerina tour, tree-ordering edges.**

We must next order the tree edges about each split vertex. Then the tree edges may be assigned a cyclic order based on selecting alternate hits from an Eulerian tour of the corresponding edges of the derived tree.

Since we can certainly split vertices in $O(n \log n)$ time using a modified depth-first search, and also order edges about each split vertex in some reasonable fashion in the same time complexity, we can accomplish the following ordering for the edges of any connected graph efficiently:

Corollary 3.2 *One may find a cyclic ordering for the edges of any connected graph in $O(n \log n)$ time so that any two edges which are consecutive in the cyclic ordering never have link distance greater than two in the graph.*

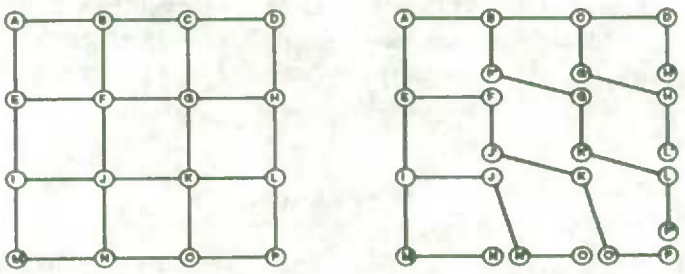
3.5 Ordering Line Segments in Two-Dimensional Networks

This is just the graph-edge ordering problem, but with fewer decisions to make because the Eulerian tour is given by the geometry. The word *network* will also imply that the topological information of the graph permits linear-time generation of the ordering. The steps for converting a connected-network edge-ordering problem to a tree-edge-ordering problem are:

ALGORITHM ORDER_SEGMENTS_IN_2-D_NETWORK

1. **Split vertices.**
2. **Walk Eulerina tour, tree-ordering edges.**

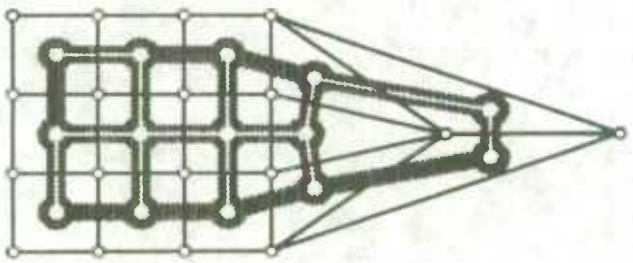
Figure 9: Vertex Splitting In a 2-D network



3.6 Ordering Regions in the Plane

Every graph or pseudograph in the plane has a planar graph dual that is itself a pseudograph. Every region of the plane corresponds to a vertex in the new pseudograph; and two vertices in the new pseudograph are adjacent (share an edge) if and only if the regions shared a face or common side. This dual is called the *adjacency pseudograph*; and to reduce a pseudograph to a tree on the same vertex set, the procedure is the same as with a graph - you throw away edges.

Figure 10: Plane graph, its Dual Graph and a Spanning Tree of the Dual



We summarize the steps needed to convert the problem to a tree-ordering problem.

ALGORITHM ORDER_REGIONS

1. **Build adjacency pseudograph.**
2. **Find MST.**
3. **Walk Eulerian tour, tree-ordering vertices.**

3.6.1 Application to Block Numbering

Consider the problem of numbering regions of a map in such a way that consecutively numbered regions are adjacent. It is well known that not every arrangement of blocks can be so numbered. In fact, when formulated as a problem in the adjacency graph, block numbering is nothing more or less than the problem of finding a Hamiltonian path for the adjacency graph (*i.e.* a path that passes through each vertex exactly once). Even the problem of merely deciding whether such a path exists for an arbitrary planar graph is NP-complete.

By throwing away edges so as to minimize the maximum degree of vertices in the resulting pruned tree, one may guarantee that the link distance between blocks numbered consecutively is no greater than the maximum degree of the resulting pruned tree.

3.6.2 Multistage Sampling

Sampling is often done in stages. Regions may be selected; and then individual households within selected regions may be subsampled. Region clustering, the capability of selecting groups of nearby regions, is important to reduce travel and other operational costs of surveys. *Non-compact regions clustering* involves the selection of nearby, but non-adjacent regions. Non-compact clustering is an attempt to gain the benefits of reduced travel costs without the negative impact of high correlation. Ordering regions by tree-ordering a pruned version of their adjacency graph will provide a reliable means of forming non-compact region clusters. Clustering of points within predetermined regions may also be achieved by assigning very high costs to potential edges that cross region boundaries when building a minimum cost spanning tree.

REFERENCES

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

Aho, A., Hopcroft, J., and 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). Personal communication.

Faloutsos, C., and 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.

Faloutos, C., and 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., and 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*, April, 22, 2, 145-157.

Preparata, F., and 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., and Harter, R. (1989). Sample Maintenance Based on Peano Keys, *Proceedings of the 1989 Symposium on Analysis of Data in Time*, Statistics Canada, Ottawa, Canada, (Eds. A. Singh and 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*, November, 11, 4, 721-736.

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

AUTOMATING THE DEVELOPMENT OF AREA SAMPLING FRAMES USING DIGITAL DATA DISPLAYED ON A GRAPHICS WORKSTATION

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

ABSTRACT

The National Agricultural Statistics Service uses area sampling frames to conduct surveys to collect agricultural data in the United States. These area frames are created by assigning land area to land use categories, and then delineating parcels of land for sampling purposes. Currently, paper-based products are used such as satellite imagery, high altitude photography and U.S. Geological Survey map products. This fall, a sampling frame will be created using the Computer Assisted Stratification and Sampling (CASS) system which will enable us to automate this process by displaying digital satellite data and digital map data on a graphics workstation. The new system will enable the cartographer to create a more accurate sampling frame by accessing more recent data and will enable us to complete the process much more quickly.

KEY WORDS: Stratification; Area frame sampling; Digital satellite imagery; Digital line graph.

1. INTRODUCTION

The National Agricultural Statistics Service (NASS) has been developing, using and analyzing area sampling frames since 1954 as a vehicle for conducting surveys to gather information regarding crop acreage, cost of production, farm expenditures, grain yield and production, livestock inventories and other agricultural items. An area frame for a land area such as a state or county consists of a collection or listing of all parcels of land for the area of interest. These land parcels can be defined based on factors such as ownership or based simply on easily identifiable boundaries as is done by NASS. Area frames provide complete coverage with all land areas being represented in a probability survey with a known chance of selection.

The existing procedure used to develop area frames is slow, labor intensive and expensive. The development of an area frame for a single state may require 11,000 hours and cost over $150,000.

This paper will briefly describe the currently used materials and procedures in developing an area frame. This will be followed by a description of a research project which is now going operational to develop area frames using digital inputs. For more information on area frame development, consult the bibliography.

2. PAPER-BASED AREA SAMPLING FRAMES

2.1 Materials Used in the Current Procedure

Area frames are currently developed on a state by state basis. The materials used in the stratification process include:

---

[1] The authors are respectively, head of Technology Research Section within the Survey Research Branch and Group Leader in charge of area frame development in the Area Frame Section within the Survey Sampling Branch of the National Agricultural Statistics Service, 14th and Independence, S.W., Room 4168-south, Washington, D.C., 20250, U.S.A.

Satellite Imagery: Satellite imagery is derived from digital data collected by scanners aboard satellites. Presently, the imagery product from the LANDSAT satellite is used. A scanner mounted on the satellite collects the reflected and emitted energy from the ground. Two types of scanners are used: a multispectral scanner (MSS) and a thematic mapper (TM). TM is the preferred product for stratification. TM is more costly due to its better resolution. The paper TM product is scaled at 1:250000.

National Aerial Photography Program (NAPP): NAPP is the product of a consortium of federal agencies, each of whom need and use aerial photography. Contact prints are used which are nine inches square and are scaled at 1:40000. NAPP is a primary stratification tool. Nearly all of the U.S. has been photographed through the NAPP program.

Topographic Quadrangle Map (Quad): These maps are produced by the U.S. Geological Survey (USGS) and the preferred scale is 1:24000 (7.5 minute series - 2.6 inches to a mile) which makes them useful for urban and ag-urban stratification and sampling.

Bureau of Land Management (BLM) Map: These maps, scaled at 1:100000, show the distribution of the federal and state land. They are useful in western states for delineating the range strata and for locating the boundaries of Indian reservations.

U.S. Geological Survey 1:100000 Map: These maps are of high quality and provide NASS with an accurate map base on which to stratify and digitize (to be defined later).

These materials, as they are currently being used, are all paper-based products.

2.2 Stratification

The process of land-use stratification is the delineation of land areas into land-use categories on photography and a corresponding map base utilizing satellite coverage as an aid. Table 1 displays the set of land-use categories which were used in the development of Missouri's area frame in 1987. The purpose of stratification is to reduce the sampling variability by creating homogeneous groups of sampling units. Although certain parts of the process are highly subjective in nature, precision work is required of the personnel stratifying the land (called stratifiers) to ensure that overlaps and omissions of land area do not occur and land is correctly stratified.

Table 1. Land-Use Strata Codes and Definitions

STRATUM CODE	DEFINITION	TARGET SIZE	
		miles^2	kilometers^2
11	General Cropland, 75% or more cultivated.	6-8	15.5-20.7
12	General Cropland, 50-74% cultivated.	6-8	15.5-20.7
20	General Cropland, 15-49% cultivated.	6-8	15.5-20.7
31	Ag-Urban, less than 15% cultivated, more than 100 dwellings per square mile, residential mixed with agriculture.	1-2	2.6-5.2
32	Residential/Commercial, no cultivation, more than 100 dwellings per square mile.	.5-1	1.3-2.6
40	Range and Pasture, less than 15% cultivated.	12-16	31.1-41.4
50	Non-agricultural, variable size segments.		
62	Water.		

Perhaps the most important concept conveyed during the initial training of personnel is the idea of using quality boundaries. A quality boundary is a permanent or, at least, long-lasting geographic feature which is easily found and identifiable by an interviewer. If an interviewer cannot accurately locate a segment in a timely manner, there is the potential for nonsampling errors to be introduced into the survey data. If the field interviewer,

unknowingly, does not collect data associated with all of the land inside the sampled area or collects data for an area outside of that selected, then survey results will be biased.

The objective of using permanent boundaries and the objective of obtaining homogeneous sampling units within a stratum often conflict in the actual practice of area frame stratification. Concessions may have to be made in marginal situations. Given that the area frame is to be used over a period of 15-20 years and represents a major investment, the best and most permanent boundaries must be used. Roads and rivers make good strata boundaries, while intermittent streams and field edges do not and should rarely be used. The following is a list of the most frequently-used geographic features which represent strata boundaries. The list is ranked by quality from highest to lowest:

- Paved highways.
- Secondary all-weather roads.
- Local farm to market roads.
- Railroads.
- Permanent rivers and streams.

The stratification is performed a county at a time for administrative purposes. Each stratifier is assigned a county and will work on that county until its completion. Stratification generally begins with determining the urban and ag-urban strata for the county. The agricultural areas are then stratified using the TM satellite imagery. The imagery is used primarily to ascertain where the cultivated areas and the non-cultivated areas are present in a county. The imagery is very useful in stratification because it is so timely. Although aerial photography may be one to five years old, the Landsat imagery usually covers the most recent growing season, providing a very recent look at the area. Using the Landsat imagery for locating crops and pasture and the photography for boundaries, the stratifier must make subjective decisions on placing areas in their respective strata.

Quality assurance is a major concern during the stratification phase. Throughout the process, checking and rechecking is performed to ensure a high quality product and to obtain the benefits of a second opinion by someone with a more experienced "eye".

After the stratification on the photography has been reviewed and approved, the strata boundaries will then be transferred to the map base (also called the frame sheets or frame maps). The map will later be digitized (electronically measured) to determine the areas of the primary sampling units (to be defined later). Once this transfer is completed, the transfer is examined by the checker and the next phase of stratification is begun...construction of primary sampling units.

2.3 Construction of Primary Sampling Units

The next step in the development of the area frame is to further subdivide the strata into primary sampling units (PSUs). The desired size of the PSU varies by strata, but contains, on the average, six to eight final sampling units or segments. The minimum PSU size is generally one segment. The use of primary sampling units introduces economic saving into area frame sampling. An entire frame need not be divided into segments in order to select a sample. The strata blocks are broken down into PSUs and a sample of PSUs is randomly selected. Only the randomly selected PSUs will be further subdivided into segments - saving a tremendous amount in labor costs. In delineating PSUs, the main focus is not homogeneity of land-use - that will have already been accomplished with the land-use stratification. The main concern is to achieve the desired size with good boundaries while trying to maintain that each PSU is a smaller representation of the strata as a whole.

The frame maps are reviewed by a statistician for completeness as a final check. The polygons created by drawing the PSUs are examined to make sure they form a closed polygon. The numbering system is checked for strata identification accuracy and sequential accuracy. The frame maps are further checked to ensure that omissions and overlaps do not exist. Once these checks have been accomplished, the frame maps are ready for the next step in the process...measuring the PSUs.

2.4 Digitization

The conversion of map points into two-dimensional X-Y coordinates is called digitization. Digitization involves electronically measuring the area of the PSUs on the frame maps. The PSUs need to be measured to determine the number of segments per PSU for sampling purposes. Electronically recording the PSU areas allows:

- measuring the PSU accurately,
- quality assurance,
- retaining a digital backup copy of the frame map in the unlikely event that a frame map is lost.

NASS utilizes analog to digital conversion tablets (4' x 5' digitizing tables) to establish a coordinate system overlaying the frame. A reference point, known as the origin (0,0), is established for X-Y coordinates on the map. X-Y coordinates, tagged with the appropriate identification, uniquely describe the borders of a PSU and therefore create a polygon for each PSU. The digitizing software records the X-Y coordinates in a file. Using the map scale, the area of each polygon (PSU) in the county is calculated in terms of square miles and stored in a file for that county.

The PSU areas for each county are summed and compared against the official county size. The same procedure is done for the state area. County areas are allowed to vary 3.0 percent from the published area. The accumulated state area is only allowed to vary 0.5 percent from the published area. The county area is allowed more variance because of the smaller area involved and because primary sampling units are allowed to cross county boundaries. Since the stratification is never allowed to cross state boundary lines, only a small amount of error is allowed.

The PSU areas are then accumulated for each stratum at the state level. The area of the PSU divided by the target segment size for the stratum is equal to the total number of ultimate sampling units (segments) in that PSU (rounded to the nearest integer). Summing the number of segments will yield the total number of segments in the stratum. This information will be used in determining the number of samples to be chosen for the entire state.

2.5 Sample Selection

After the total number of sample segments to be used in a state has been determined, a program is run to select the PSUs which will be further broken down into sample segments. Each segment has a specific target size depending on the stratum it is associated with such that each individual segment closely resembles the full PSU (as much as possible) with the best physical boundaries available.The selected PSUs are located on the frame map and the PSU boundaries are transferred to photography. The selected PSU is then broken down along identifiable boundaries into the required number of segments. The segments are manually numbered and a random number is chosen between one and the number of segments in the PSU. The segment corresponding to the random number is the selected segment. A photo enlargement of the selected segment will eventually be sent to the interviewer for enumeration.

3. DIGITAL-BASED AREA SAMPLING FRAMES

3.1 Research Background

NASS has been involved in a cooperative agreement with the National Aeronautics and Space Administration (NASA), the U.S. Space Agency, to develop area frames using digital inputs. The project with NASA began three years ago. Although the initial research agreement with NASA will expire in the fall on 1991, NASA will continue to provide software support through a cooperative agreement with the Ecosystem Science and Technology Branch (ECOSAT), a group located at Ames Research Center, Moffett Field, California.

NASS and ECOSAT have a history of cooperation in remote sensing research. The two agencies have worked together on a number of projects since the late 1970's. EDITOR, a software package for large area crop acreage estimation utilizing Landsat MSS data and associated ground data, was written by NASS and ECOSAT

personnel. Eighty percent of PEDITOR, a portable version of EDITOR, that is, a system suitable for implementation on a variety of hardware devices, was written at ECOSAT. PEDITOR is the primary software tool for remote sensing operational work in NASS. ECOSAT assembled a prototype microprocessor-based workstation, called MIDAS, for NASS and assisted with an experiment to determine the performance characteristics of the system when generating area estimates in the NASS operational environment. ECOSAT created display software for NASS that is compatible with PEDITOR and a precursor for the system required for area frame development.

The development of the area frame software has been an iterative process. The software supplied by ECOSAT to NASS has been constantly evaluated for completeness and ease of use by the Area Frame Section of NASS. This evaluation by the students and supervisors who work on the system, has led to a close working relationship with the people at NASA. Suggestions for improvement and modification are addressed on a very timely basis.

A new area frame system called the Computer Assisted Stratification and Sampling (CASS) system, based on PEDITOR concepts and integrated into the PEDITOR software will strengthen both the research and operational remote sensing programs at NASS and NASA and the development of area frames. One particular advantage to this approach is the ability to use digital information relating to land use from previous years' surveys as an aid to the development or updating of area frames. Other benefits will be discussed later.

3.2 The CASS System

CASS incorporates digital inputs in the form of thematic mapper (TM) LANDSAT data and USGS Digital Line Graph (DLG) data. The TM data (1:100000 scale) serves as a base to delineate land use according to our stratification scheme. For boundary identification, DLG data at a 1:100000 scale is overlaid onto the TM image using a graphics plane.

Displaying satellite data - Three bands of the seven band Landsat image are loaded into the system for display at 30 meter resolution. The image can be zoomed on the screen to various levels. This satellite image provides the most recent look at the area to be stratified. This digital product parallels the use of the paper color print of the satellite data and the black and white high altitude photography.

Color mapping - The display window uses 8 bits for each of the three bands yielding a sharp and colorful image. A color map is one which assigns brightness levels to each of the bands. A dynamic mapping function allows the user to control the brightness and contrast of each of the 3 planes. These settings can then be preserved in a file for later use. Several other functions allow the user to look at a single band, look at a histogram of each band, and do a linear, piecewise, or equiprobable mapping. The optimum mapping is the one which best distinguishes the cultivation and boundaries. This is checked by locating previous survey fields on the screen, and noting the colors.

Display and registration of DLG data - The DLG data are used as a reference. The stratifier must be sure that they are using good boundaries for the PSU and segment delineation steps. Current digital data includes transportation and hydrography. Political and administrative boundaries will be available soon. The original DLG files exist in a unit called a "panel" (7.5 or 15 minutes in size). A CASS function reads these files from tape and creates disk files. These disk files can then be added together and placed in one file to cover a larger area, such as a county. If a county crosses UTM (Universal Transverse Mercator) zones, the data can be converted so that a single county file can be utilized. (Panels from different zones cannot be added together.) Another CASS function displays the data file to the screen. In order to precisely overlay the TM image with the DLG, another program is used to "rubber-sheet" (or register) the DLG data to the backdrop of satellite data by selecting several matching points in each data set and running a least squares regression to fit the rest of the data. These points and the regression are also saved in a file and used each time this particular DLG file is displayed. Lastly, this registration file allows the user to determine latitude and longitude coordinates of a given point.

PSU delineation - In each county, polygons will be drawn off and tagged with the appropriate PSU number, which consists of a stratum number and a sequence number. This is done by determining what land use stratum a piece of land belongs to by interpreting the color TM display. At the same time, a PSU within some specific size range is delineated, using physical boundaries identified by the DLG and/or TM data. In CASS, this is done

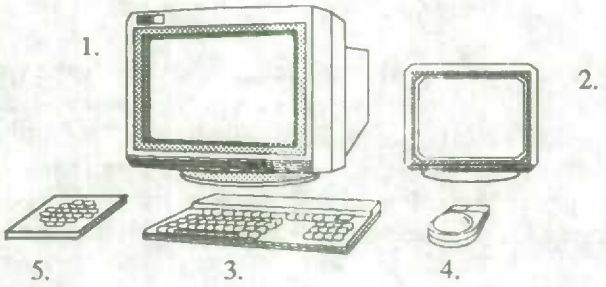
by keying in the PSU number, and then utilizing the mouse to pick points along the desired boundaries. When the PSU is closed, the area is immediately calculated and displayed. This allows the stratifier to determine if the PSU is within the target size for that stratum. If the polygon is too small or too large, the polygons can be combined, split or reshaped. When a county is completed, the polygons are saved to a stratification file to be later reviewed by an experienced stratifier. The user has the ability to check for overlapping polygons and for holes (or missing land areas). At any time, the user may list PSUs that have been created so far, to check numbering and areas.

PSU breakdown into segments - After the entire state has been stratified and the total area for each stratum has been calculated, a separate program is run to draw the sample of PSUs (first stage of sampling) which will be further broken down into final sampling units or segments. Not all PSUs are broken down - only those which were chosen using the sample select program. The user displays the stratification file (saved in the previous step) and types in the PSU number which is to be sampled. The software then erases all but the sample PSU from the screen. Many of the same functions which were involved in delineating the PSUs during the stratification phase are used to delineate the PSU into equal size segments. Similar quality control checks are done. When the PSU has been broken down into segments, a segment can be selected randomly using the segment selection command. This is then the second stage of sampling.

3.3 The CASS Workstation

The CASS Workstation (see Figure 1) includes several pieces of equipment. The display terminal (1) displays the color Landsat image, whereas the menu screen (2) displays the software. The keyboard (3) is used to enter commands onto the menu screen, where the mouse (4) is used to interact with the display terminal. The button box (5) is also used in connection with the display terminal to handle the overlay planes (change the color of the display or turn the display on and off) and to zoom the image.

Figure 1.
CASS Workstation



At this time, a UNIX-based Hewlett-Packard (HP) workstation is being used to handle the processing and storage requirements of the massive amount of data. The HP workstations possess the minimum capabilities for area frame development, that is, three image planes, four overlay planes and a 1024 x 1280 display coordinate system. The configuration allows for three bands of satellite data in the image planes while utilizing the graphics planes for displaying DLG, PSUs, and the command menu. A recent modification allows the DLG to be displayed in more than one overlay plane. One reason for this is to separately color the transportation and hydrography data.

3.4 The Initial Test

The purpose of the initial test was to gain basic experience with the software, compare CASS to the current method, and determine the speed of frame construction. Digital data covering three north-central counties in Missouri (Linn, Livingston and Macon) were used. These counties were chosen partly because the Area Frame Section had developed a new area frame for Missouri in 1987 for use in 1988 and because of the availability of digital data. A few results are listed.

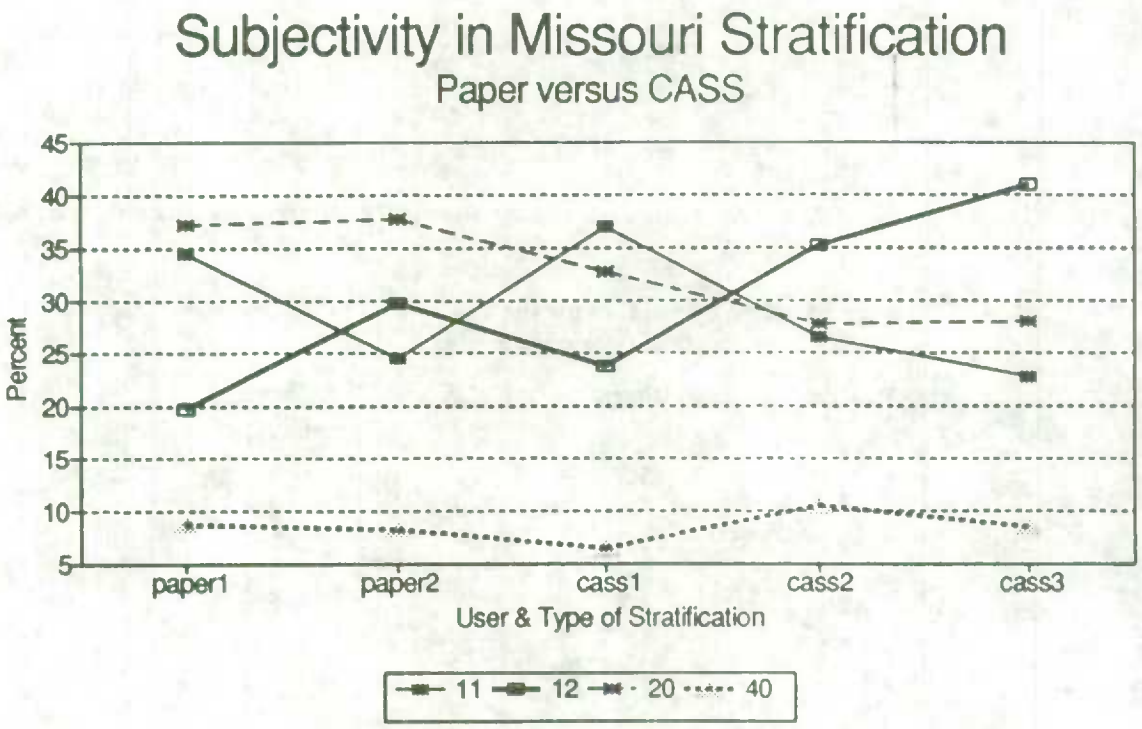
1) This test proved that stratification using CASS was possible.

2) CASS proved faster, as 2.5 - 3 weeks were required for the current stratification, and 2 - 3 days for CASS.

3) Several software enhancements were made. It was about this time that we decided to add sample select modules.

4) One important note is the subjective nature of the work. In this test, 3 people did the same counties in CASS and 2 people did the paper stratification. (See Figure 2.)

Figure 2.



5) It was also about this time that we switched from a SUN display to a Hewlett-Packard (HP) display. The screen size increased from roughly 512 to 1,024 pixels. The number of overlays increased from 2 to 4, and the HP enabled the full TM scene to be loaded into memory rather than just one screen's worth.

3.5 The Expanded Test

The next phase of research was to represent a pseudo-operational environment. A larger area in Michigan was selected (21 counties) and system requirements were to be considered. Reasons for selecting this area was first, that the state had just received a new frame in 1989 (implemented in 1990), and second, that the Remote Sensing Section of NASS had recently done some work in the dry bean area of Michigan in regards to supervised classification (therefore data was available). In this test, only one person did each county. Because of the recency of the paper stratification, a potential bias exists in the CASS stratification.

3.6 Analysis

This analysis is based on results from the expanded test using the 21 counties in Michigan. Evaluations are both quantitative and qualitative in nature.

Table 2 compares the total area stratified in the Michigan study using the CASS system versus the current manual mode. The percentage difference (given as CASS-operational/operational) is also shown. Some difference is shown for stratum 12 and stratum 31 with a larger difference shown for stratum 40. Comments from stratifiers generally favored the CASS system for the stratum 40 difference as they felt they were better able the pick out the woodland areas.

Table 2. Comparison of Total Area by Stratum for the 21 Counties in the Michigan Study. Area is in square miles.

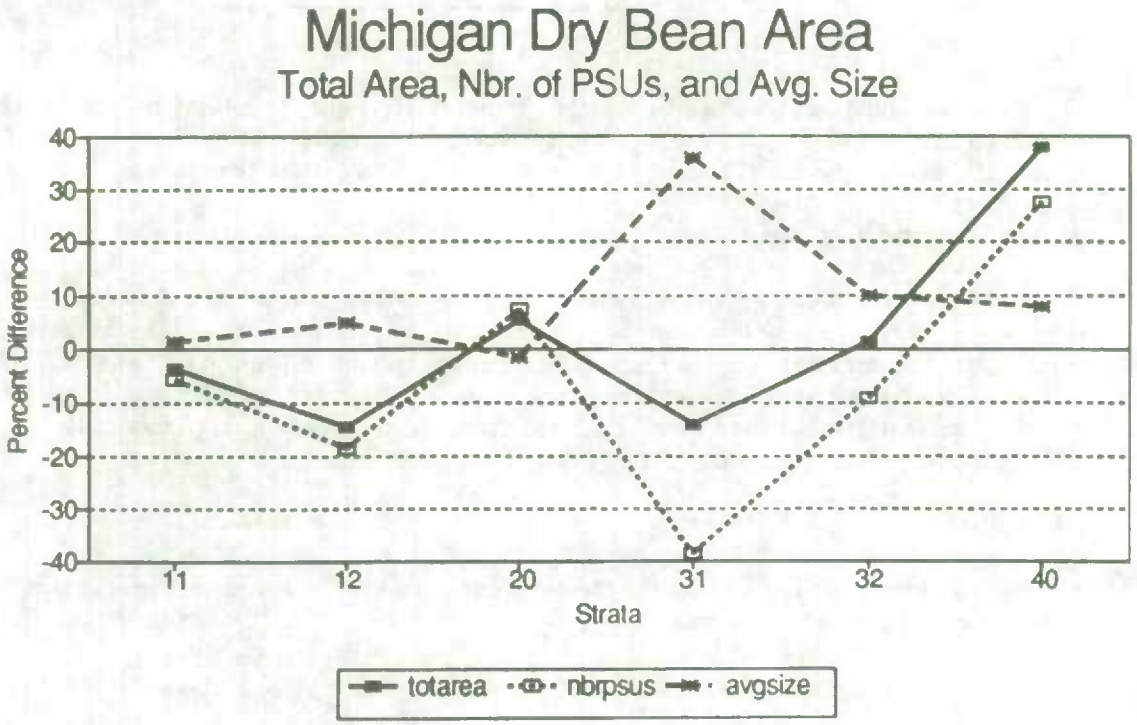
Stratum	Operational	CASS	% Difference
11	5,427.4	5,225.8	-3.7
12	2,547.8	2,175.6	-14.6
20	3,264.0	3,448.0	5.6
31	648.0	555.3	-14.3
32	168.4	170.7	1.4
40	1,347.3	1,856.2	37.8

Table 3 describes the average PSU size by stratum as a comparison between the operational mode and the CASS system.

Table 3. Average PSU Size by Stratum for the 21 Counties in the Michigan Study. Area is in square miles.

Stratum	Operational	CASS	% Difference
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.



The quantitative data are interesting to review (see Figure 3) but the qualitative aspect of the study focuses on more informative issues. When the Michigan test was completed, the users compared the paper and CASS frames of 5 counties in Michigan. The issues raised, relate to the digital display, stratification, and sample selection.

A few improvements to the digital display will help the user do a better job the next time.

a) TM
- It was determined that the display of the satellite image could be improved. Initially, the zoom levels were set at 1, 2, 4, 8, and 16. At the end of the Michigan test, zoom levels were changed to 1, 2, 3, 4 and 5.
- Initially, each stratifier created a color map for their county. This created a problem during the review, as each person's color map was different. It was then decided to have one color map per LANDSAT scene in the future. Also, the stratifiers were not aware of some of the mapping techniques, which might have helped in the LANDSAT interpretation.

b) DLG
- The digital line graph data was entered into one overlay plane. Only at the end of the test, was a modification made to display DLG in different planes. In this way, roads can be displayed in one plane (in one color), and water in another plane (in another color).
- Also, the line display was improved. Because of an error, lines did not retain the dots and dashes when the image was redisplayed or zoomed. This has been fixed.
- Also, the points marked by the user will appear larger at the zoom levels two or greater.
- Finally, the lines all "look" the same within a given overlay plane. Therefore, the stratifiers initially used illegal boundaries (such as 3z3z swamp boundaries) which "looked" the same as legal boundaries. A program was then written to delete all illegal boundaries from the DLG files. However, work is currently underway to better display the DLG so the user can "recognize" the specific type of boundary.

A few comments about the stratification are mentioned below.

a) EXPERIENCE
- The color interpretation was a new experience, and it will take time to become proficient at this.
- Also, on paper, entire strata are delineated and then split into PSUs. On CASS, strata are determined but delineated at the PSU level.
- Looking back, several people commented that they might have done better with the experience they have now.

b) STRATIFICATION
- In general, the stratifiers felt that the stratification of cultivated areas was better in CASS. As to urban areas, the paper did better, due to increased resolution of the high altitude photography. On paper, houses can be seen somewhat, but on CASS the only guide to cities were the closeness of roads. A few ideas to handle this are covered in the "Future Considerations" section. Lastly, the water was more detailed on the paper maps, but by displaying one band, the user can clearly see the water, and the DLG displays streams and rivers.

c) BOUNDARIES/CLOUDS
- As to the boundaries, the users felt the current method was somewhat better. This is due in part to the resolution of the data.
- In two cases, clouds were a problem. In one case, a cloud totally obscured a piece of land which resulted in its misclassification (should have been a stratum 11). In the second case, a fair sized island was missed due to a wispy cloud which totally obscured it. If the cloud is wispy, a different band combination may allow us to "see" through it.

An analysis of the sample selection portion of CASS is currently being done. A comparison will be done using the segment size, the homogeneity of the PSUs, and the type of boundaries used. The important outcome was the distinction (as to the software) between the stratification and sample selection processes. The first attempts

at sample selection proved to be fairly time consuming due to display difficulties. This can be improved by working with smaller TM data files (rather than using a full LANDSAT scene), by automating the identification of the PSUs location, and by working at different zoom levels. The speed can also be improved in the long run through the graphical user interface (see the "Future Considerations" section), and in the short term by modifying the commands available with the mouse, and those which must be keyed in on the menu terminal. Lastly, the resolution of the TM image can be improved by the bands which are selected, and through increased resolution from future satellites. One important note is that boundaries are more vital to this process since the polygons are much smaller.

3.7 Benefits of Using CASS

In any project, such as CASS, several components of quality can be addressed. These are accuracy, resources, timeliness, and relevance. (Statistical Policy Working Paper 20, 1991).

As to accuracy, several benefits come to mind.

a) The satellite data provides more recent data than the high altitude photography (which is important when a frame is used for 15 years), and provides better resolution (1:100000 rather than 1:250000) than the paper-based satellite imagery. This relates to better agricultural stratification.

b) The tedious, error prone process of transferring from a satellite image print, to high altitude photography, to a 1:100000 USGS map, to a digital file can be eliminated.

c) PSUs can more easily be revised in CASS, than on paper.

d) Sample segments are chosen by the software using a random number generator instead of looking through a set of random number sheets.

e) Several concerns are boundaries, urban stratification, and cloud cover. The DLG is only as good as the date it was created or revised. Boundaries on LANDSAT are harder to identify than boundaries on the photos. In cases where the black and white photography is more recent than the 1:100000 maps, changes in boundaries could be determined. Also, houses cannot be seen as well using the current Landsat data. Ways to improve on these are discussed in the "Future Considerations" section.

Resources and Timeliness go together. With CASS, we could do more states with the same number of people, or do the same number of states with fewer people. NASS prefers the latter, as personnel can be reallocated to other areas where there is need.

a) A reduction of staff can occur, as the labor intensive stratification process is automated on CASS.

b) A reduction of staff can occur, as the digitization process (to calculate areas and serve as a backup) is no longer required.

c) In general, CASS takes less time than the current process. A county takes about 2 weeks to do on paper, but only 2-3 days to do the same process on CASS.

d) The ability to immediately determine the PSU size should help reduce extra labor involved in the sample selection process.

e) Cost is an important resource. The hardware costs will be large initially, and then decrease substantially. The material costs will increase due to the purchase of digital TM data rather than TM imagery. Salary costs should be much less due to the reduced staff and the shorter time involved. Details on this are not yet available.

Relevance refers to the use of frames and sample segments by the end-user.

a) The digital aspect of the frame, will allow a frame to be updated rather than having to start from scratch (which is currently done on paper due to the age of the paper, and the impossibility of erasing color pencil).

b) The locations of sample segments can be identified (latitude and longitude) and read into a Geographic Information System (GIS) database.

c) Allows easier exploration of specialized area frames. The Remote Sensing Section in NASS can provide crop-classified satellite imagery to assist in the development of specialized area frames.

3.8 Future Considerations

There are several ways in which we hope to improve CASS. This involves the stratification process, the boundary determination, and other possibilities.

The stratification process may be improved in several ways.

a) As of yet, only bands 2, 3 and 4 have been used for display purposes. By accessing other bands, we may see additional color detail, and even be able to see through clouds.

b) We are looking into the use of unsupervised and supervised classification of satellite data and hope to use it as an aid to stratification.

c) We are looking into using Census Bureau TIGER data and the corresponding Public Law File to locate the boundaries of a Census block on CASS and look up the number of households, to better determine the urban land classification. Topographic Quad maps may also be used to determine urban classification and the appropriate boundaries.

d) A gentleman from the U.S. Department of Agriculture (USDA) is looking into a coordinated purchase of Landsat data by all agencies within the U.S. Department of Agriculture (USDA). By sharing the data, all agencies should benefit, and CASS might benefit from possibly having more than one Landsat date per scene.

The boundary selection process is vital and may be improved in several ways.

a) The use of other bands (especially band 5) should provide more detail.

b) We are investigating the use of filters to improve the ability to detect "edges" or boundaries.

c) When Landsat 6 becomes available, we will need to utilize the panchromatic band (which gives 15 meter resolution) with the spectral bands (which will still have 30 meter resolution).

d) The availability of 1:100,000 DLG for administrative and political boundary data should be helpful. Also, "area" records on the DLG file are not yet utilized, could provide useful information.

A few other improvements are given.

a) The NASA programmers are working on a Graphical User Interface for CASS, so that we can get rid of the "menu" terminal, by displaying menus within "windows".

b) The polygon files created within CASS, have successfully been read into a GIS using ARC/INFO. An exchange of information the other way, may be possible.

c) Latitude and longitude coordinates can be determined within CASS. This will help NASS to order photography more easily and accurately, and fill other potential needs.

REFERENCES

Ciancio, N.J., Rockwell, D.A., and 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., and Nealon, J. (1987). Area Frame Design for Agricultural Surveys. U.S. Department of Agriculture, National Agricultural Statistics Service.

Fecso, R., and 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., and 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., and 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., and 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

Spatial Analysis of Health and Environmental Data

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

SPATIAL AUTOCORRELATION: TROUBLE OR NEW PARADIGM?

P. Legendre[1]

Spatial autocorrelation may be loosely defined as the property of random variables that are observed to take values, at pairs of locations a certain distance apart, that are more similar (positive autocorrelation) or less similar (negative autocorrelation) than expected for randomly associated pairs of observations. Autocorrelation is a very general property of ecological variables and, indeed, of all variables observed along time series (temporal autocorrelation) or across geographic space (spatial autocorrelation). Most natural ecological phenomena display geographical patchiness, and it is found at all spatial scales - from microns to continental and ocean-wide scales. Picturing the spatial variation of the variable(s) under study in the form of a map shows the structure to be smoothly continuous, or else marked by sharp discontinuities.

The statistical problem that accompanies the spatial structuring of ecological data can be illustrated using the following common case of spatially autocorrelated data. The observed values of the variable of interest - for instance, species composition - are most often influenced, at any given locality, by the species assemblage structure at surrounding localities, because of contagious biotic processes such as growth, reproduction, mortality, migration, and so on. In such a case, since the value at any one locality can be at least partly predicted by the values at neighbouring points, these values are not stochastically independent from one another. This may come as a surprise to ecologists who have been trained in the belief that nature follows the assumptions of classical statistics, one of them being the independence of the observations. However, field ecologists know from experience that living beings in nature are distributed neither uniformly nor at random; the same applies to the physical variables that we use to describe environments. Spatial heterogeneity is functional in ecosystems, and not the result of some random, noise-generating process, so that it becomes important to study it for its own sake. The first message of this paper is then that we have to revise our theories and models, to make them include realistic assumptions about spatial and temporal structuring of communities.

Autocorrelation in a variable brings with it a statistical problem; it impairs our ability to perform standard statistical tests of hypotheses. Fortunately, statistical concepts and techniques are now becoming available to handle such data. These will be briefly mentioned.

Finally, ways will be described of introducing spatial structures into ecological modelling. Two families of techniques have been experienced with: the raw data approach, based on multiple regression, and the matrix approach, based on the Mantel test.

REFERENCE

Legendre, P. Spatial autocorrelation: Trouble or new paradigm? Ecology (Submitted as part of a Special Feature).

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

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

LOCALLY WEIGHTED ANALYSIS OF SPATIALLY AGGREGATE BIRTH DATA: UNCERTAINTY ESTIMATION AND DISPLAY

D.R. Brillinger[1]

ABSTRACT

The concern is the analysis and display of data that is aggregate over geographic regions (such as census divisions) for phenomena that are felt to vary smoothly in space. In Brillinger (1990b) some spatial locally weighted analyses were provided for births taking place to women aged 25 to 29 in the years 1986 and 1987 for the province of Saskatchewan at the census division level. Various results were displayed via contour plots. The present work develops these ideas further and is, in particular, concerned with the computation and display of appropriate uncertainty levels for contour plots. A weekday effect is noted, but it does not vary appreciably with space. The approach provides an alternative to the empirical Bayes methods that have been proposed for similar problems.

KEY WORDS: Aggregate data; Binomial-logitnormal distribution; Contouring; Extra-variation; Interpolation; Locally weighted analysis; Logitnormal distribution; Maps; Simulation; Spatial data; Uncertainty presentation; Unmeasured covariates.

1. INTRODUCTION

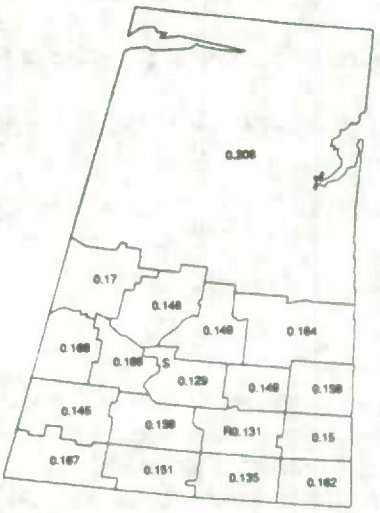
The concern is with analyzing smoothly varying spatial phenomena and with the provision of some indication of the uncertainty of the results of the analysis.

The data studied is spatially aggregate, representing totals over census divisions. The province of concern is Saskatchewan with 18 census divisions. Census division aggregate daily totals of births to women aged 25-29 for the two year period 1986-1987 are available for analysis. Also available are the census population totals for Census Day, 3 June 1986. It is desired to study and display the spatial structure of this data.

Figure 1 shows the Saskatchewan census divisions and the observed annual birth rate for each division. The letters R and S indicate the locations of the cities of Regina and Saskatoon respectively. The rates are estimated directly from the counts for the two years. More details concerning this data set may be found in Brillinger (1990a,b).

Figure 1

Saskatchewan: 1986-87 births, ages 25 to 29



annual birth rate

[1] D.R. Brillinger, Statistics Department, University of California, Berkeley, CA., U.S.A.

2. WEIGHTS AND ESTIMATION

The goal is spatial analysis and display, by means of contour plots. It is usual, in preparing contours from function values at scattered points, to first interpolate the values to a regular grid, see Pelto *et al.* (1968). Given (x_i, y_i, z_i), $i = 1, \ldots, n$ with z_i representing the value of a variate measured at location (x_i, y_i) a variety of methods have been proposed for this interpolation. A popular scheme is due to Shepard (1968). He computes z at (x,y) via:

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

where $w_i(x,y) = [(x-x_i)^2 + (y-y_i)^2]^{-\mu}$ with $\mu > 0$. The points are weighted inversely with their distance to (x,y). Other schemes are described in Franke (1982) and Sabin (1985).

In the present case, sampling fluctuations are present and one cannot simply interpolate. Further the data are counts, and proportions are computed from these, so the fluctuations are not elementary. Locally weighted likelihood analysis is a pertinent estimation technique for nonelementary distributions varying in space, see for example: Gilchrist (1967), Brillinger (1977), Tibshirani and Hastie (1987), Cleveland and Devlin (1988), Staniswalis (1989), Brillinger (1990a,b). Suppose a variate Z has probability function $p(z \mid \theta)$ depending on an unknown parameter θ. Let $\psi(z \mid \theta)$ denote the score function, $\partial \log p / \partial \theta$, and chose $\hat{\theta}$, the estimate of θ at location (x,y), to satisfy

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

for some weight function $w_i(x,y)$. As in Shepard's method, $w_i(x,y)$ depends on the distance of the point (x_i, y_i) to the location (x,y).

As a simple example of a locally weighted estimate, consider the case of B_i binomial with parameters π, N_i. One computes directly the estimate

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

This estimate is a natural extension of (2.1).

In the present case, where the data is aggregate over regions R_i, the choice for the weight is

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

with $W(.)$ the biweight,

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

and equal 0 otherwise where $u = b\sqrt{x^2+y^2}$ for some $b > 0$. In (2.4) $|R_i|$ is the area of division i. One can view $w_i(x,y)$ here as representing the influence of census division i on a person at location (x,y), the influence resulting from items like travel, nutrition, climate, ethnicity, education, television, laws. These weights are evaluated via a Fourier transform, taking advantage of the convolutional form. A naive weight would be $w_i(x,y) = 1/|R_i|$ for (x,y) in R_i and $= 0$ otherwise. This corresponds to $W(.)$ a delta function.

Figure 2 shows the effect of varying the parameter b for Census Division 18, the northern half of the province. The values of b for the three cases illustrated in Figure 2 correspond to no smoothing, a small amount of smoothing, and a moderate amount. This last value is employed in the computations of the paper. Figure 3 gives a plot of (2.4) for all of 18 of the divisions.

Figure 2

Census division 18 - effect of smoothing

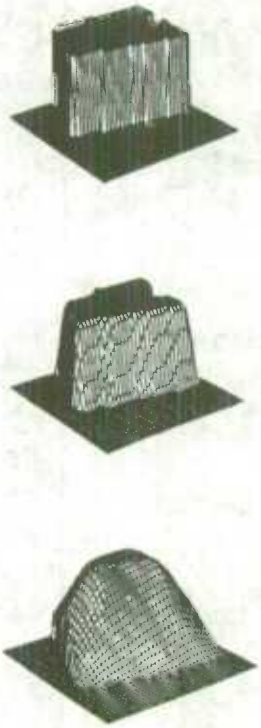
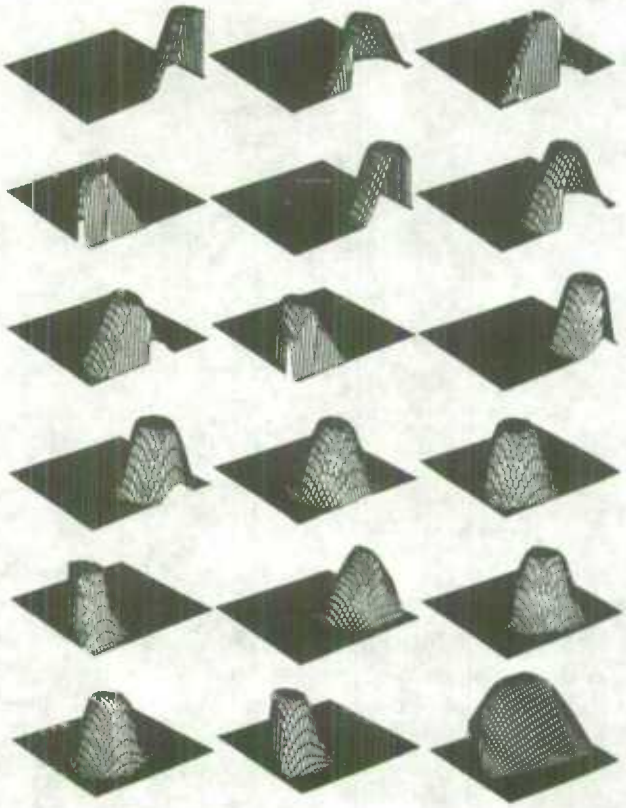


Figure 3



Alternate weights that might be employed are given in Tobler (1979) and Dyn and Wahba (1982). An advantage of the present weights is that they are local in character.

3. ESTIMATION: BINOMIAL CASE

Let B_{ijk} denote the number of births to women ages 25 to 29 in census division i and year j with $k = 1, 2$ depending on whether data is for a weekday or weekend. Let N_i denote the census population of census division i for the age group. The count B_{ijk} may be thought of as the number of births from this population. Its distribution may be approximated by a binomial. (This seems a better approximation than the Poisson used in Brillinger (1990a,b), since the chance that a woman has a baby in a year appears to be able to be as high as .2). Set $x_1 = 2$ and $x_2 = -5$. This will make the estimates orthogonal. It will be assumed specifically that B_{ijk} is binomial with parameters π_k and N_i where

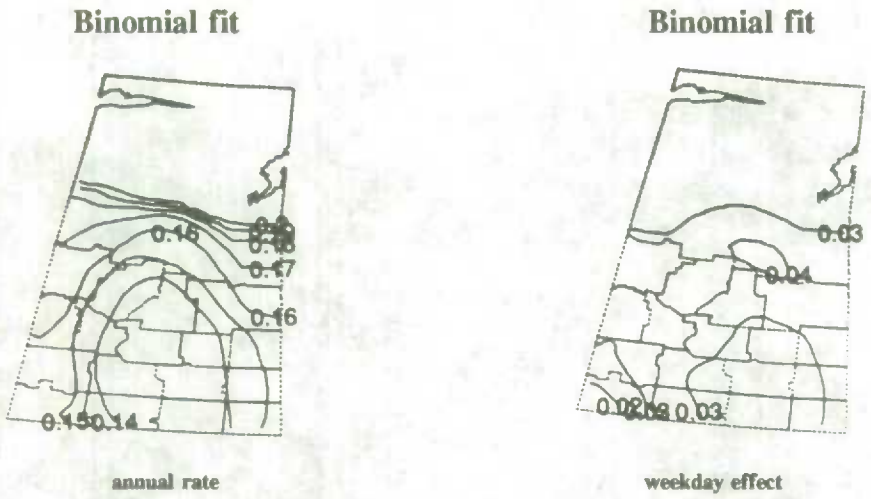
$$logit\ \pi_k = \log\,(\pi_k/(1-\pi_k)) = \alpha + \beta x_k. \tag{3.1}$$

The left-hand panel of Figure 4 provides contour plots of the annual birth rate estimated by the expression

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

where $\hat{\eta}_k = \hat{\alpha} + \hat{\beta}x_k$, for $k = 1, 2$ and it is to be remembered that $\hat{\alpha} = \hat{\alpha}(x,y)$ and $\hat{\beta} = \hat{\beta}(x,y)$. The first term in (3.2) corresponds to weekdays, the second to weekends. One notes contours rising up and out from the census divisions including Regina and Saskatoon. The right-hand panel graphs the weekday effect estimate $\hat{\beta}(x,y)$. All contours are seen to be positive, corresponding to an increased number of births on weekdays. It must be remarked that these quantities are all subject to sampling fluctuations. The provision of an indication of sampling fluctuations will be discussed in Section 5.

Figure 4



4. ESTIMATION: BINOMIAL-LOGITNORMAL CASE

It is argued in Brillinger (1990a,b) that a variety of pertinent explanatory variables, *e.g.* diet, lifestyle, weather, environment, holidays, age structure, urbanicity, will have gone unmeasured. This will lead to extravariation in the number of births beyond that of a binomial. One way to proceed is to introduce a random effect, σz, and replace (3.1) by

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

where z is a standard normal variate. The model is now binomiallogitnormal. The z's for the different census divisions are assumed independent. This model could be fit making use of numerical integration, see Bock and Lieberman (1970), Pierce and Sands (1975), Sanathanan and Blumenthal (1978), Brillinger (1990a,b) for example. In the present case the N_i are large and one can approximate the model by a logitnormal, *i.e.* by assuming the logits of the B_{ijk}/N_i are normal. Specifically this model may be written

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

with the ϵ_{ijk} independent normals of mean 0 and variance σ^2. The assumption here may be checked, in part, by fitting the logitnormal and examining probability plots of the residuals. This was done for the data consisting of the annual totals for each day of the week and census division. No strong departure was noted although there was an intriguing outlier.

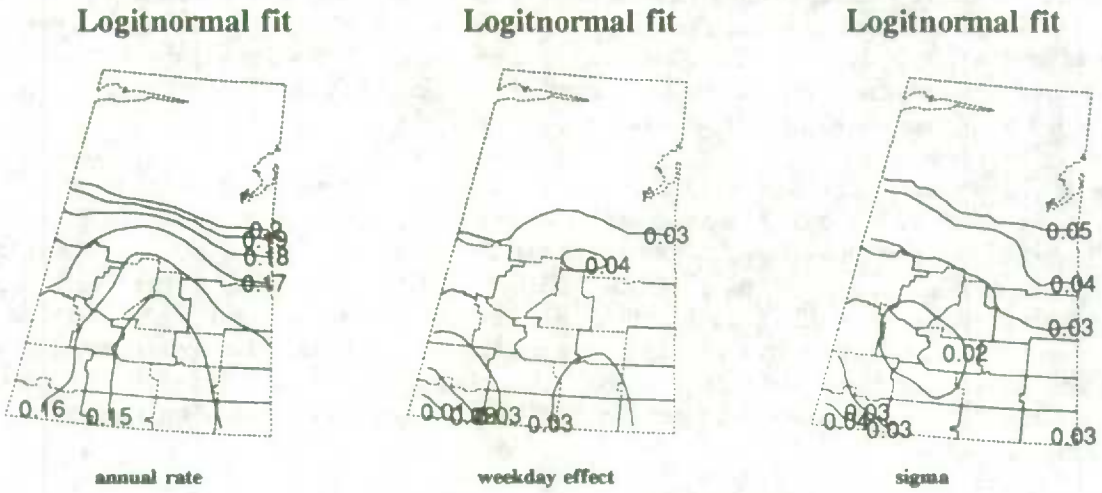
The annual birth rate is now estimated via

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

where $\eta_k = \hat{\alpha} + \hat{\beta}x_k + \hat{\sigma}z$ for $k = 1, 2$ and with $\phi(.)$ the normal density. The first term corresponds to weekdays, the second to weekends, as in (3.2). Crouch and Spiegelman (1990) discuss the numerical evaluation of integrals such as those appearing in (4.3). In the present case Gaussian integration with 21 nodes is employed.

The top left-hand panel of figure 5 provides the estimated rate function (4.3). As compared with the binomial fit of Figure 4, the birthrate surface estimate appears flatter. The top right-hand panel is $\hat{\beta}(x,y)$. This surface appears less flat. The final panel gives the estimate $\hat{\sigma}(x,y)$. It appears less in the region around Saskatoon, but of course is subject to sampling fluctuations.

Figure 5



5. UNCERTAINTY COMPUTATION AND DISPLAY

Simple maps are flaunt with difficulty of presentation and interpretation, see for example Monmonier (1991). The provision of associated indications of uncertainty seems even more difficult. This section presents a few procedures for the case of contours.

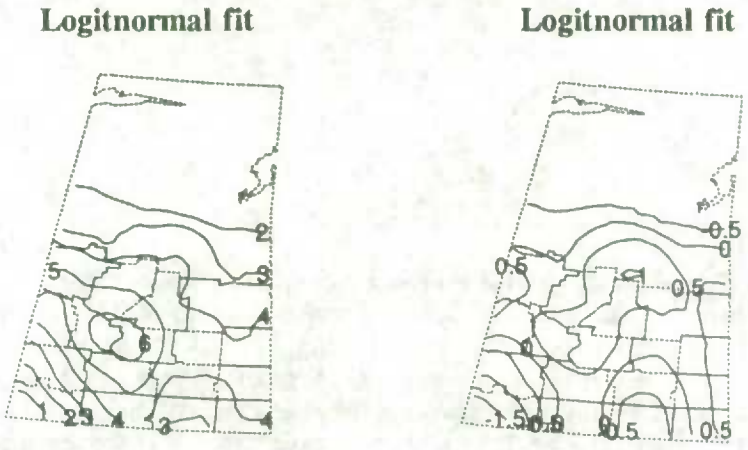
The actual computation of the uncertainties at each (x,y) seems not to be the problem. The estimates produced by the logitnormal fitting are weighted least squares so, writing in traditional notation, the variances of $\hat{\alpha}$ and $\hat{\beta}$ may be estimated by

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

and the standard deviation of $\hat{\sigma}$ by

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

Figure 6

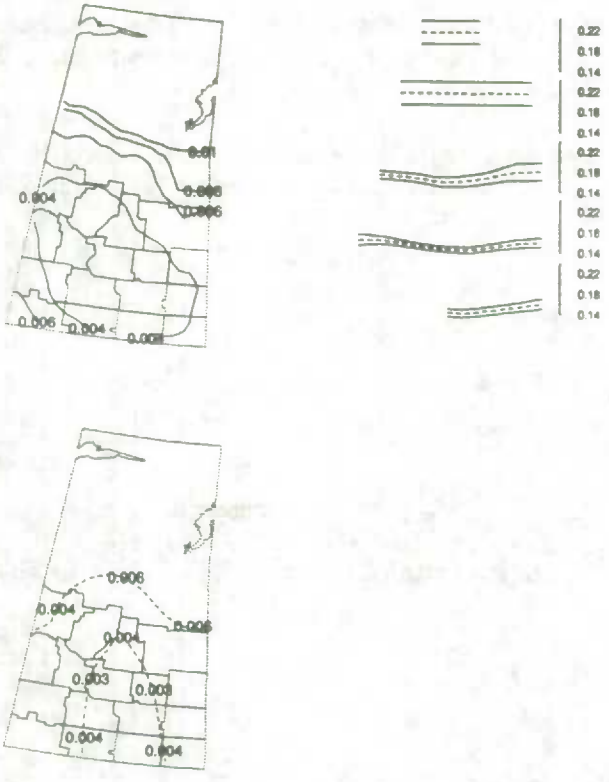


The uncertainty estimates may be employed to examine hypotheses. Figure 6 is directed at the issue of whether there is a weekday effect and whether or not it varies with space. The left-hand panel of Figure 6 provides the estimate, $\hat{\beta}(x,y)$, divided by its estimated standard error. The values range from 1 to 6, providing evidence for the presence of an effect. To study whether the effect varies spatially, the t -statistic is recomputed but now with its numerator having the estimated value for the whole province subtracted. Now the t -values range from -1.5 to 1 and there is no real evidence that the effect varies spatially.

The next two figures are directed at displaying the uncertainty of the annual rate estimate given by (4.3) and graphed in Figure 5. The standard error of the estimate (4.3) is estimated by the delta-method. The first panel of Figure 7 is a contour plot of this estimate. Contour levels range from .004 to .010, and suggest a valley centered at Regina and Saskatoon. The top right-hand panel of the figure is based on five east-west slices through the province. Provided are ±2 standard error limits about the estimated rate along the slice. This is a traditional means of indicating uncertainty for functions of a single variable. The final panel graphs the .15 and .17 birthrate contours as dashed lines and superposes the estimated standard errors at selected positions along these contours. It appears simpler to take in the uncertainty this way, but one cannot read off the standard error estimates for most locations.

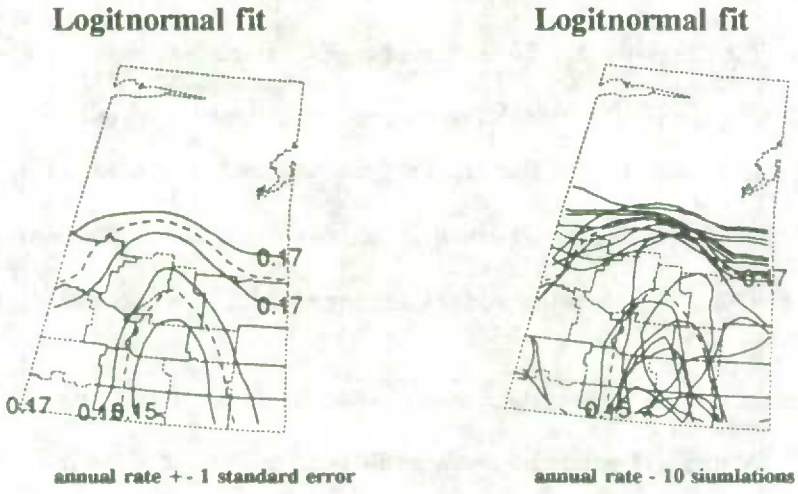
Figure 7

Logitnormal fit



The left-hand panel of Figure 8 indicates the shift in contour lines produced by adding and subtracting one standard error to the estimated rate function. Smooth bands appear about the estimated rate contours. The intention of the bands is clear, but thoughtful interpretation is needed. The right-hand panel provides 10 simulations of the process. The individual estimates $\hat{\alpha}(x,y)$, $\hat{\beta}(x,y)$, $\hat{\sigma}(x,y)$ are first aggregated over census divisions. Then independent realizations of the process are generated according to (4.2). The model is refit for each simulation, in bootstrap style, and the .15 and .17 contours determined. Diaconnis and Efron (1983) propose bootstrapping of contours. The .17 contour realizations fall in an apparent band, but the .15 move about a fair amount with some quite wild curves. The dashed lines are the original .15 and .17 contours.

Figure 8



6. DISCUSSION AND SUMMARY

The combination of the weight function, $w(.)$, and the random effect σz, allows the estimate at location (x,y) to borrow strength from the values of all census divisions, see for example Mallows and Tukey (1982). Hence this approach provides an alternative to the empirical Bayes estimates developed in Clayton and Kaldor (1987), Tsutakawa (1988), Cressie and Read (1989), Manton *et al.* (1989) for this type of data.

The computation of uncertainty has allowed examination of hypotheses of no weekday effect and weekday effect constant across the province. An advantage of the graphical approach is that were there spatial variation, then the plots might have suggested its character.

A substantial amount of work remains for the future. This includes: choice of the bandwidth, b, in (2.5), use of other weight functions, informal and formal analysis of goodness of fit, other methods to display uncertainty, including measured explanatory variables and finally appropriate asymptotics to employ in studying the technique.

ACKNOWLEDGEMENTS

The research was partially supported by National Science Foundation Grant DMS-8900613.

REFERENCES

Bock, R.D., and 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). Mapping aggregate birth data. *Proceedings of the 1989 Symposium on Analysis of Data in Time* (Eds. A.C. Singh and P. Whitridge), Statistics Canada, Ottawa, Canada, 77-83.

Brillinger, D.R. (1990b). Spatial-temporal modelling of spatially aggregate birth data. *Survey Methodology*, 16, 255-269.

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

Cleveland, W.S., and 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., and Read, T.R.C. (1989). Spatial analysis of regional counts. *Biometrika*, 31, 699-719.

Crouch, E.A.C., and 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., and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 248, 116-130.

Dyn, N., and 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., and Tukey, J.W. (1982). An overview of techniques of data analysis emphasizing its exploratory aspects. *Some Recent Advances in Statistics* (Eds. J. Tiago de Oliveira *et al.*). Academic, London, 111-172.

Manton, K.G., Woodbury, M.A., Stallard, E., Riggan, W.B., Creason, J.P., and 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., and Boyd, H.A. (1968). Automatic contouring of irregularly spaced data. *Geophysics*, 33, 424-430.

Pierce, D.A., and 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 (Ed. R.A. Earnshaw). *NATO ASI Series*, F17. New York: Springer-Verlag.

Sanathanan, L., and 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., and 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.

FIGURE LEGENDS

Figure 1. Annual birth rate for women aged 25-29, years 1986-7, by census division. R and S indicate the locations of Regina and Saskatoon respectively.

Figure 2. Provides (2.4) for census division 18 in the cases of no, of a small amount and of a moderate amount of smoothing.

Figure 3. (2.4) as employed for all 18 census divisions.

Figure 4. The binomial analysis. The rate estimate is from (3.2). The righthand panel gives $\hat{\beta}(x,y)$.

Figure 5. The logitnormal analysis. The rate estimate is from (4.3). The other panels give $\hat{\beta}(x,y)$ and $\hat{\sigma}(x,y)$.

Figure 6. t-statistics directed at the hypothesis of no weekday effect and of spatially constant weekday effect.

Figure 7. The first panel is a contour plot of the estimated standard error of the estimated birth rate. The second panel provides plus and minus two standard error limits about S east-west slices through the estimated birth rate surface. The final panel plots the estimated standard errors at selected points on the .15 and .17 contours.

Figure 8. The first panel gives the .15 and .17 contours as dashed lines, then the same contours for the surface increased and decreased by one standard error. The second panel shows the results of simulating realizations of the process (4.2) and determining and displaying estimates in the manner of the paper. Ten simulations are shown.

SESSION 4

Spatial Developments in Data Processing

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

AUTOMATED CODING OF MOBILITY PLACE NAME DATA FOR THE 1991 CENSUS

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

ABSTRACT

Spatial data on "place of residence five years ago" are collected in Canada's Census of Population. The Census question on mobility asks migrants to write in the names of the municipality, Census Division (CD) or county, and province/territory in which they lived five years previously. As part of data processing, these write-ins are converted to the 7-digit numeric Standard Geographical Classification (SGC) code consisting of a 2-digit province code, a 2-digit CD code, and a 3-digit Census Subdivision (CSD) code. Until the most recent Census (1991), mobility place name write-ins were converted manually to SGC codes. Now, for the first time in the Canadian Census, automated coding has been implemented to convert write-ins to numeric codes for a number of variables, including mobility. This paper describes the strategy and database structure used in the automated coding of place name write-ins and the resolution of a variety of response problems, such as: use of common rather than official, place names; incomplete information (partial responses); and duplicate place names. Implications of automated coding of spatial data are discussed, especially the impact of improved coding accuracy on CSD-level migration data.

KEY WORDS: Autocoding; Census; Migration; Mobility; Place names.

1. INTRODUCTION

1.1 Background

Since 1961, Canada's Census has included a question on place of residence 5 years ago. It asks respondents whether or not they had moved, that is, if they had lived at a different address 5 years ago, and if so, whether or not they had lived in a different city, town, village. In cases where they lived in a different city, town etc. respondents are asked to write in the name of the city, county and province. The version of the '5-year ago' mobility question from the 1991 Census is shown with an example of a response in Figure 1.

During census operations these place name write-ins are assigned the Standard Geographical Classification (SGC) which corresponds to the Census Subdivision (CSD), that is city, town, village, etc., Census Division (CD), (such as county) and province. The SGC is a 7-digit code consisting of a 2-digit province code, a 2-digit CD code, and a 3-digit CSD code.

From the coding of these place name write-ins we can derive for a particular CSD the number of people who had lived there 5 years ago. In other words, for each CSD we can obtain the number of out-migrants, as well as in-migrants (those who had lived outside the particular CSD 5 years ago). These CSD-level migration data are used by planners and researchers alike. The accuracy of coding these write-ins to corresponding geographical codes is a crucial factor in the quality of migration data for CSDs, particularly out-migration.

[1] M.J. Norris & S. Coyne, Demography Division, Statistics Canada, 6-A6, Jean Talon Bldg., Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

FIGURE 1. 1991 Census Mobility Question on Place of Residence 5 Years Ago

21. Did this person live at this present address **5 years ago**, that is on June 4, 1986?

25.

01 ○ Yes, lived at the same address as now
Go to Question 23

02 ⊗ No, lived at a different address

22. Where did this person live **5 years ago**, that is, on June 4, 1986?

Some large cities are made up of smaller cities or towns called municipalities. Where applicable, distinguish between the municipality and the large city, such as Anjou and Montréal, Scarborough and Toronto, Burnaby and Vancouver, Saanich and Victoria.

Mark one circle only.

03 ○ Lived in the same city, town, village, township, municipality or Indian reserve

OR

04 ⊗ Lived in a different city, town village, township, municipality or Indian reserve in Canada
Print below.

City, town, village, township, municipality or Indian reserve

05 WINDSOR

County (if known)

ESSEX

Province/territory

ONT

OR

06 ○ Lived outside Canada
Print name of country.

07

FIGURE 2. Example of Content from Place Name Code Book Used in 1986 Census

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	

Until the most recent Census (1991), mobility place name write-ins were converted manually to SGC codes as part of regional operations. This coding operation was based on the Place Name Code Book (PNCB) a reference manual containing some 40,000 place names in Canada (for 1986), including official and unofficial place names, such as neighbourhood and unincorporated place names. These place names correspond to some 6,000 standard geographical codes. An example of the type of information contained in the PNCB is given in Figure 2. Now, for the first time in the Canadian Census automated coding has been implemented to convert write-ins to numeric codes for a number of variables, such as major field of study and place of birth, as well as mobility.

1.2 Organization of the Paper

This paper provides a review of the developments in the automated coding of mobility spatial data, commencing with some background on the need for autocoding of mobility place name write-ins and an examination of the objectives of place name autocoding. This discussion is followed by a description of the approach and strategy to autocode place name write-ins using the Automated Coding by Text Recognition (ACTR) software developed by Statistics Canada. In addition to the strategy and resolution of response problems during autocoding, developments related to the production of coded mobility data, including quality control and edit and imputation are also considered. Results of testing and production to date are analyzed, followed by an assessment of the autocoding of this spatial data thus far. Implications of the autocoding of mobility spatial data are discussed, especially in relation to improved quality of CSD-level migration data, and new information on place name usage.

2. NEED FOR AUTOCODING OF MOBILITY DATA

The mobility variable was considered as a candidate for automated coding for three main reasons: data quality, volume and cost. The need for autocoding of mobility place names was recognized following an evaluation study of mobility and migration data from the 1986 Census at the small area level. This study (Norland 1988) showed that migration data for small areas, specifically CSDs, and to some extent CDs, suffered from data quality problems. These data quality problems were traced to two sources of error: coding error and respondent error. In addition to quality considerations, simply from the viewpoint of volume, automated coding was justified, given close to 1 million write-ins of place names and the potential for reducing production costs.

2.1 Quality of Small-Area Migration Data

The combination of coding and respondent errors resulted in a number of data quality problems at the CSD level such that: migration rates for "small CSDs" (CSDs with a population below 250) are unreliable; a significant number of larger CSDs have excessive out-migration rates; special problems involve data for "duplicate name places" (*e.g.* Barrie, for which there exists the township of Barrie in Frontenac County and the city of Barrie in Simcoe County); and, special problems involve selected CSDs within CMAs, such as Saanich and Victoria. The latter situation is more a case of respondent error in that the respondent tends to confuse the name of the larger metropolitan area (*e.g.* Victoria) with the suburban municipality (*e.g.* Saanich). However in the other types of problems coding error has been a significant source of error, particularly in the resolution of SGC codes for duplicate place names. Respondent and coder errors affect the correct assignment of the SGC codes for "CSD place of residence 5 years ago" and therefore the measure of out-migrants from the CSDs and the corresponding out-migration rates. As the study noted, it is for this reason that out-migration rates for CSDs tend to be worse than in-migration rates, which are not affected by these errors in the same way since they are based on SGC codes of residence at census time.

Some examples of the types of problems that were identified in the study of the 1986 migration data are given in Table 1. Out-migration rates for the selected CSDs are shown to indicate the extent of the suspected problem. The CSD of Tungsten in the Northwest Territories illustrates the problem of small population size, with some 200 people and a derived 1981-86 out-migration rate of 149 migrants per 100 residents. Suspiciously high out-migration rates are observed even for larger CSDs, such as Dawson Creek in British Columbia, with a population of over 9,000 and an out-migration rate of 50 per cent. The special problem of duplicate place names is illustrated by the case of Barrie City and Barrie Township both in Ontario. The smaller CSD of Barrie Township has an out-migration rate of 133%, compared to 19% for Barrie City, probably as a result of coders incorrectly assigning the SGC code of Barrie Township instead of Barrie City, to the place name write-in of only "Barrie".

The special problem of respondents confusing specific CSDs with the larger CMA is illustrated by the case of Saanich and Victoria. It can also become a coding problem in the case of both place names being provided.

As a result of these various data quality problems, the study made a number of recommendations regarding use of CSD-level migration data such that users refer to areas with large base populations and be aware of "special situations" such as "duplicate place names". As well, original plans to publish data on in-, out- and net-migration for CSDs were altered because of the significant number of CSDs with excessive out-migration rates. Only data on mobility status were finally published for CSD profiles. In addition, the study recommended that improvements be made to coding procedures and manuals for the 1991 census (recommendations were made on the basis of manual coding at the time). Further information on data quality problems is provided in the 'User's Guide to 1986 Census Data on Mobility' (Norris 1990). As well, changes made to the '5-year ago' mobility question in order to reduce some of the respondent-based problems for 1991 are provided in the National Census Test report on mobility (Norris 1989).

Automated coding was recognized as having the potential for improving the quality of migration data from the 1991 Census by providing a more accurate and consistent means of coding place name write-ins. Manual coding has been problematic for data quality due to coder error, inconsistent and subjective coding, with complex instructions and manuals to follow. Unlike manual coding, automated coding offers consistency and the ability to systematically isolate "special problem" types of responses.

Table 1: Examples Of CSDS With Suspected Data Quality Problems, 1986

Type of problem	CSD	1986 Population (Aged 5+)	Out-migration rate[1] (per 100 population) (1981-86)
Small population size	Tungsten, NWT	205	149
Larger CSDs with excessive out-migration rates	Dawson Creek, BC	9,470	50
Duplicate place names	Barrie City, Ontario	44,440	19
	Barrie Township, Ontario	690	133
Selected CSDs wihin CMAs	Saanich, BC	77,045	2
	Victoria, BC	60,540	52

[1] Out-migrants are derived from the 1986 Census question on place of residence 5 years ago, which yields the number of out-migrants aged 5 and over, for the 1981-86 period.

Source: 1986 Census, unpublished mobility data.

3. OBJECTIVES OF AUTOMATED CODING OF MOBILITY DATA

The main objective of the automated coding of mobility place name data is to improve quality of coding over that of manual coding. As well, other major goals for autocoding in general include the reduction of processing time and costs.

To be specific, the objective of mobility autocoding is to accurately assign a 7-digit geographic code to a corresponding place name write-in. In other words, the system of automated coding is required to automatically code, with minimal error, as many of the one million entries as possible. Because of the nature of respondent problems associated with mobility, it was realized during development of automated coding that the system could not assign codes for all write-ins accurately. In order to be both efficient and accurate, the system was developed with two objectives: to achieve automated coding for at least 70% of write-ins with an acceptable level of accuracy; and to allocate special coding problems, such as duplicate place names, to manual resolution. Thus, the manual resolution phase of autocoding would deal with those write-ins for which the system could not accurately assign codes.

3.1 Response-Based Coding Problems

At first, the assignment of a SGC code to a place name write-in might seem straightforward. After all, the respondent is asked to provide the name of the city, town, village, reserve, etc., the county and the province, which neatly corresponds to the CSD, CD and province of the SGC code. However, there are a variety of respondent and place name problems that can make coding of place name data fairly complicated, especially when using an automated system. Table 2 provides some types and examples of problems that can make the linkage of place name write-ins to SGC codes problematic. For example, an incomplete response, can sometimes, but not always lead to problems of duplicate place names — the incomplete response of 'Kingston' can refer to six possible CSDs across Canada with the place name of 'Kingston'. In the instance of a simple, but unanticipated spelling or key-entry mistake, the automated system cannot assign a code, such as in the case of 'Totonto'.

Table 2: Problems In Assignment Of SGC Codes To Mobility Place Name Data

Type of response	Examples
• Incomplete response	Kingston
• Common usage, such as neighbourhood and unincorporated place names	Glebe (neighbourhood), Lake Park
• Spelling errors	Totonto
• Abbreviations	Mtl.
• Incorrect combinations of city, county and/or province	Kingston, Manitoba
• Incorrect county names	Provincial electoral districts reported instead of counties, particularly for Quebec
• More than one place name	Saanich/Victoria Wawa/Hawk Junction
• Word order	Poplar Point/Point Poplar
• Street Address	235 Montreal Rd Ottawa
• Non-geographic response	On the farm
• Out-of-date names and boundaries	Galt

4. AUTOMATED CODING SYSTEM

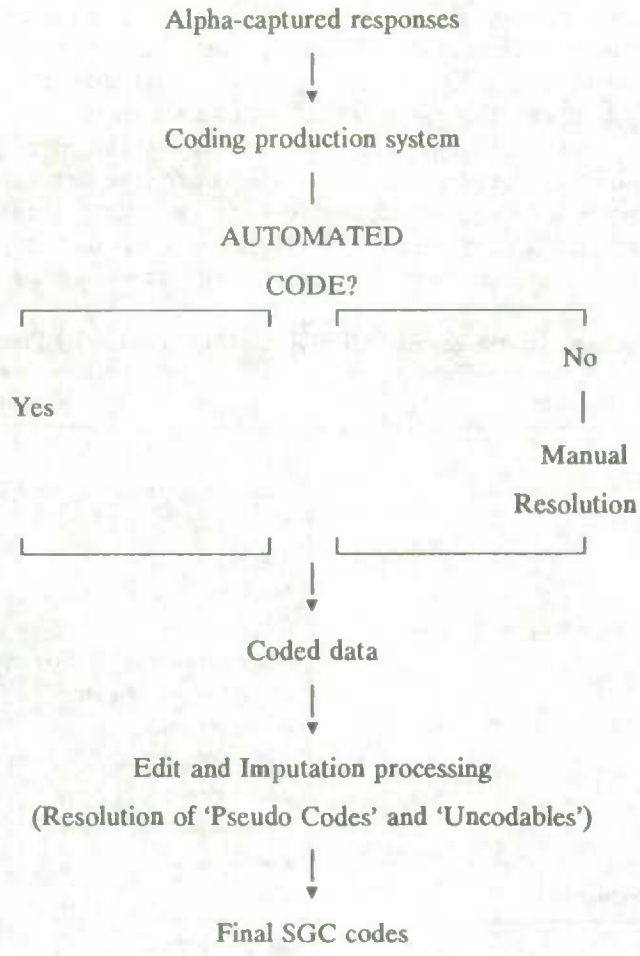
4.1 Overview of Automated Coding

Figure 3 provides an overview of the automated coding process for the 1991 Census. It starts with the alpha-captured names, that is, keyed in place name write-ins in the case of mobility. These responses are entered into the autocoding system which interprets keyed-in responses in order to assign a numeric code. There are two possible outcomes: either an automated code is assigned to the write-in, or, failing that, the write-in is directed to manual resolution. In manual resolution, specially trained coders try to assign codes to those write-ins that the automated system could not code. Some of the codes assigned during manual resolution are 7-digit 'pseudo' codes, as well as actual SGC codes. Coders also determine whether or not a write-in is codable. Thus, output from the coding production system consists of coded data, the majority of which are SGC codes; as well as 'pseudo codes and 'uncodables', which are later resolved in edit and imputation (E&I), from which final SGC codes are obtained.

4.2 Automated Coding by Text Recognition (ACTR)

The automated coding production system is based on 'ACTR', which stands for Automated Coding by Text Recognition, a computer software developed by the Research and General Systems Section of Statistics Canada. Research results on the use of ACTR as an automated coding system for Census are presented in the report by Fyffe *et al.* (1988). The ACTR system interprets keyed write-in responses in order to assign a corresponding numeric code. For example, in the application of ACTR to mobility, the place name write-in of Ottawa, Ontario is assigned the SGC code of 3506014 by ACTR.

Figure 3: Overview Of Automated Coding



In order to match the write-in response to a numeric code, the ACTR system contains a **reference** file consisting of phrases (*e.g.*, place names) and their corresponding numeric codes (*e.g.* SGC codes). As well, a **parsing** strategy is used to standardize as much as possible numerous variations in write-in responses, such as abbreviations, common misspellings, missing or extraneous words. The automated coding process consists of a series of successive steps that involve matching the respondent's write-in to a corresponding phrase against a phrase on the reference file.

Each subject matter area uses ACTR to develop its own reference file and parsing strategy specific to the census variables, such as major field of study. In the case of mobility, the coding of place name data represents the geographic application of a generalized coding system.

4.3 Geographic Reference File

Adapting a generalized coding software such as ACTR to a geographic application entailed working within some limitations and constraints. For example, the ACTR version that is in production for the 1991 Census does not distinguish between word order, so that, two separate place names, such as Spring Hill and Hill Spring, are interpreted to be the same phrase. In addition to the constraints of the ACTR system, the design of the mobility question itself imposed some restrictions on the usage of ACTR features such as the "filter" option. Had province been asked first, followed by county and city, the database or reference file could have been constructed such that matching of place name phrases could have been done on a province-by-province basis. However, since city is asked first and province last, respondents are less likely to provide province, and hence, it would be difficult to first match according to province and then by more geographic detail. Thus, these limitations had to be taken

into consideration when developing the geographic reference file, parsing and matching strategies for mobility place name data.

The geographic reference file used in the automated coding of mobility was based on the PNCB. Data for the reference file were derived from the machine-readable format of the PNCB, containing: place names for census subdivisions (CSDs), counties (CDs) and provinces; abbreviations for CSD type (eg. C for City, R for Reserve); and, corresponding SGC codes. In addition, the PNCB contains common, but non-CSD place names, such as neighbourhood or unincorporated place names and the SGC codes that they correspond to.

The reference file was developed to maximize automated code assignment. The strategy entailed building a hierarchy of place name responses according to completeness. Records containing 'partial' combinations of place name information were included in the reference file. Based on write-ins from the 1986 Census respondents do not give complete answers. The usual or most frequent response given, consists of the CSD place name and the province (*e.g.* Hull, Quebec). Sometimes only the city name is given without the province (*e.g.* Toronto). In the case of 'duplicates' within the same county, CSD type is required to differentiate the two places (*e.g.* Kingston township *vs.* Kingston city), and occasionally respondents will indicate the CSD type where it is relevant. Thus, while respondents are asked for the CSD name, CD name and province, often only partial information is supplied.

In the case of partial responses, automated code assignment depends on the type of response. If for example a partial response corresponds to a unique place name, then the system can assign a code. If however, a partial response corresponds to more than one place, or if there is insufficient information such as province only, then the system is prevented from assigning a code and the write-in is referred to manual resolution. During manual resolution, pseudo codes may be assigned where codes for partial write-ins cannot be resolved. In addition to the SGC codes, the reference file contains these pseudo codes and identifies preferred actual SGC codes which coders assign during manual resolution.

In order to accommodate the variety of responses, and to achieve as many unique matches as possible, a geographic hierarchy of response phrases was devised within the reference file. This structure of responses and their conditions for automated code assignment are outlined Figure 4. The structure of this reference file has generated some 170,000 records corresponding to about 40,000 place names and 6,000 CSDs.

Figure 4: Hierarchy of Responses in Geographic Reference File

	Response Phrase				Automated Code Assignment
	CSD/ PLACE NAME	CD	CSD TYPE	PROV/ TERR	Response Phrase Will Result in Automated Code Assignment:
1.	X	X	X	X	Always, response phrase unique within Canada
2.	X	X		X	Only if place name is unique within CD
3.	X			X	Only if place name is unique within province
4.	X	X			Only if place name is unique within CD and response phrase unique within Canada
5.	X				Only if place name is unique within Canada
6.	X		X	X	Only if place name and CSD type unique within province
7.	X	X	X		Only if response phrase of place name, CD and CSD type unique within Canada
8.	X		X		Only if response phrase of place name and CSD type unique within Canada

4.4 Examples of Geographic Reference File

The example of Kingston in Figure 5 serves to illustrate both the hierarchial structure of the geographic reference file phrases and the problem of partial responses. This chart shows the four components of geographic information a response could provide: the CSD name, the CSD type, the CD and province. We have eight possible variations of reference file phrases for 'Kingston City'. In the first case we have a complete response of 'Kingston City, Frontenac County, Ontario' which corresponds to one unique place, for which an SGC code can be automatically assigned. The next three partial response phrases containing the phrase 'Kingston City' also correspond to the SGC codes for Kingston City. However, the following three partial response phrases, of simply 'Kingston' in combination with either Frontenac County and/or Ontario suggest 2 possibilities — either Kingston City or Kingston Township, both of which are in Frontenac County. They would not result in an automated code assignment but would be referred to manual resolution. Finally, in the worst-case scenario, we have the partial response of simply 'Kingston' which could correspond to six possible place names in Canada, and would thus be referred to manual resolution.

4.5 Parsing Strategy

In the case of mobility, parsing was minimal as compared to other coded variables such as "major field of study". Parsing is minimal for mobility because for place names, small differences are important, *e.g.*, Saint John, St. John's. Examples of parsing for place names include: double letters, *e.g.*, BB becomes B; standardized province names, *e.g.*, any expected version of Ontario becomes ON, while 2-word provinces become one word, *e.g.*, all versions of British Columbia become BC. Parsing is also used for common misspellings and abbreviations of place names.

Figure 5: Example of Geographic Reference File Based on 'Kingston'

Response Phrases				Number of places in Canada phrase refers to
CSD	CSD type	CD	Province	
1. Kingston	City	Frontenac	Ontario	1
2. Kingston	City	Frontenac		1
3. Kingston	City		Ontario	1
4. Kingston	City			1
5. Kingston		Frontenac		2
6. Kingston		Frontenac	Ontario	2
7. Kingston			Ontario	2
8. Kingston				6

4.6 Computer-Assisted Manual Resolution

The manual resolution phase of automated coding is computer-assisted. For most write-ins that are sent to manual resolution, ACTR suggests code/phrase combinations. In some cases, codes cannot be suggested by the system for write-ins that are non-geographic, (*e.g.* on the farm) or spelling mistakes. In the example in Figure 6, ACTR suggests three possible codes for the write-in of 'Kingston, Ontario' for Person 1 in the household: the SGC code for Kingston City (3510009); the SGC code for Kingston Township (3510011); and the pseudo code for Kingston, Ontario (9935172). If the coders can obtain additional household information, such as a more complete response from other members, then the coder chooses the appropriate code. The system provides additional respondent household information on-line that the coder can easily access. As well, additional geographic reference material is provided to coders. If, during manual resolution the coder cannot determine the correct response then either a pseudo code or a preferred actual SGC code is assigned in the case of duplicate name problems. In other cases, an 'unable to code' assignment is made. In the example in Figure 6,

the other person in the household, husband/wife of Person 1, has not provided any additional information only the same write-in of 'Kingston, Ontario'. Thus, in this case, the pseudo code for 'Kingston, Ontario' is assigned.

4.7 Resolution of Pseudo Codes

Pseudo codes that are assigned by coders during the manual resolution phase of autocoding are resolved during edit and imputation (E&I) of census data. They are distributed into N-way splits based on the number of actual SGC codes that they correspond to, generally in the range from 2 to 8. Again, we can use the example of Kingston, Ontario. During manual resolution, the coder can find no additional household information to resolve the response 'Kingston, Ontario' which could correspond to either Kingston City or Kingston Township. Thus a pseudo code, 9935172 is assigned in manual resolution.

In E&I, responses which have been assigned the pseudo code of 9935172 for Kingston Ontario are distributed proportionally according to population size between the two SGC codes for Kingston City and Kingston Township (3510009 and 3510011 respectively). In addition to pseudo code resolution, other problems are resolved using family or donor imputation, for cases such as: responses that couldn't be coded; missing responses where a write-in should have been provided; and partial responses of only 'county' and/or 'province'. The assignment of SGC codes to 'CSD place of residence 5 years ago' is finalized in edit and imputation.

Figure 6: Example of Computer-Assisted Manual Resolution for 'Kingston, Ontario'

```
MANCDMOB     1991 CENSUS OF POPULATION / AUTOMATED CODING                04/06/91
MMANUALI  MANUAL CODING - MOBILITY 5 YEARS AGO INSIDE CANADA            12:00:00.0
Write-in to be coded                                  Type    Code

KINGSTON ON                                           M       __

Phrases returned by ACTR                                      Codes    (S)elect
KINGSTON ON (DUP)                                 __          3510009  - -
KINGSTON ON (DUP)                                 __          3510011  -
KINGSTON ON (DUP - ON)                            __          9935172  - -

ID: 35009020 100 2 22
================================================================================
Data for same question from each household member                Persnr:  1
Check-boxes
_____ _____ _____ _____ _____ _____ _____ _____
Write-ins
KINGSTON ONTARIO

Enter-PF1---PF2---PF3------PF4----PF5--------PF6---------PF7-----PF8------PF9--------PF10----PF11-----PF12---
   HELP   UP   DOWN  <<<<  >>>>  MORE  +UP  +DOWN  REFER  VALID  COMIT  QUIT
```

```
NDISMOB          1991 CENSUS OF POPULATION / AUTOMATED CODING            04/06/91
MMOBLTY1                    Mobility                                     12:00:00.0

                 1 YEAR AGO
                 Inside Canada     : SAMPRTE
                                   : _________
                 Outside Canada    : _________
                                   : _________
                 5 YEARS AGO
                 Inside Canada     : DIFCITY
                                   : KINGSTON ON
                                   : _________
                 Outside Canada    : _________
                                   : _____
                                   : _________
        Relationship to person 1   : HUSBAND/WIFE OF PERSON 1
                      Birth date   : 23-06-1962

ID: 35009020 100 2 22

Enter --   PF1 --- PF2 --- PF3 --- PF4 --- PF5 --- PF6 --- PF7 --- PF8 --- PF9 --- PF10 --- PF11 --- PF12 ---
           HELP                                                                                  QUIT
```

5. RESULTS AND BENEFITS OF AUTOMATED CODING OF MOBILITY

5.1 Coding Results

Given that automated coding for the 1991 Census is currently in the middle of production, only preliminary results on the automated coding of mobility are available. To provide some background, results from the research and testing phase of mobility autocoding, as well as from 'production-to-date' output, are presented in Table 3.

Table 3: Results of Automated Coding of Mobility Place Names

Selected Indicators	Research & Testing (Based on 1986 Census Sample of Write-ins)	Production-to-Date (November '91)
• Percent of write-ins resulting in automated code assignment:		
Range	60 - 80%	70 - 80%
Average	70%	75%
• Error rates in coding		
a) Based on sample of 1986 write-ins coded by system:		
- Automated	1%	Not applicable
- Manual (1986 Census coding results)	4%	
b) With measure of error incorporated for duplicate place name resolution:		
- Combined automated and manual resolution (expected)	Not applicable	1 - 2%
- Manual (estimate based on 1986 manual coding results)	8%	Not applicable

• Research and Testing

Automated coding for mobility was developed using a sample of mobility write-ins from the 1986 Census. During research and testing, automated code assignment was achieved for close to 80% of place name write-ins. Error rates of coding, based on these write-ins for which the automated system assigned codes, were 1% for automated coding compared to 4%, based on the manual assignment for these same records by coders in 1986. These error rates do not include code assignment for duplicate place names, involving pseudo or preferred codes. (Pseudo codes were not used in the 1986 coding of mobility.) Details of early research and testing of automated coding for mobility are available in the report by Norris and Kirk (1989).

• Production-to-date

Preliminary results based on production-to-date, which are provided in Table 3, indicate that on average, 75% of place-name write-ins result in automated coding assignments. For the various production runs so far, automated code assignments range from 70 to 80%. While it is difficult to precisely measure coding error rates at this early stage, it is anticipated that the combined automated manual resolution for 1991 will result in error rates of 1 to 2%. This is a significant improvement over an estimated error rate of roughly 8% for 1986 manual coding, which incorporates a measure of error for duplicate place name resolution.

5.2 Additional Benefits of Automated Coding

Along with improved coding accuracy, automated coding has provided additional benefits related to production, evaluation and further development.

• More efficient manual resolution

First of all, the manual resolution phase of automated coding is much more efficient than manual coding operations of 1986. Because manual resolution is computer-assisted, it eliminates excessive paper burden of the

previous census — there is no time lost flipping through files and code books. Responses in work units are grouped alphabetically. Coder time per write-in is much less in 1991 than in 1986.

• Better control and monitoring

Automated coding has ensured that there is better control and monitoring of coding operations for 1991. In the previous census, coding of mobility data occurred in regional offices across Canada. Now, coders can easily consult with subject matter specialists because operations are centralized. Also, results can be easily monitored with computerized coding reports on production.

• Lower production costs

Early estimates based on production-to-date indicate that production costs for 1991 coding of mobility will be significantly lower than those in 1986. Rough estimates for 1991 suggest costs of about $260,000 and four person-years *vs.* $466,000 and almost 17 person-years in 1986. A major portion of the cost for 1991 is alpha capture for the keying-in of place name write-ins.

• Computerized quality control (QC)

The computerized quality control (QC) allows one to easily track and monitor both manual and automated assignment of codes to unparsed write-ins. A log of unparsed write-ins and corresponding codes is maintained with the system QC and as well an *ad hoc* query system enables us to examine manual code assignments. Most importantly, with the QC approach it is possible to identify and rectify systematic error (via post-production fixes). Quality assurance strategies that have been used in automated coding for the 1991 Census are outlined in the paper by Ciok (1991).

• Better understanding of small area data quality

Output from automated coding of actual place name write-ins and their corresponding code assignments, will prove useful in the analysis of small area data quality. For the first time, it will be possible to know the extent of various response problems, such as: reporting of neighbourhoods or common, rather than official, place names, misspellings, incomplete responses, incorrect county names, more than one place name and street addresses. Evaluation of mobility place-name data from autocoding will provide a better understanding and assessment of the quality of small-area migration data for 1991.

• Input for next census

The 1991 Census experience with autocoding will provide input for the development of mobility in the next census in a number of areas. For example, with respect to production for the next census, the percentage of place-name write-ins resulting in automated code assignment could be increased, with further streamlining in manual resolution and with more resolution of SGC codes being done in E&I. Evaluation and analysis of place-name data from the 1991 Census will yield new knowledge and insight about respondent usage of place names and about respondent comprehension of mobility questions. This type of new information, which is now possible because of automated coding, will prove invaluable in the development of new mobility questions for the next census.

6. CONCLUSION

In conclusion, the original objectives for automated coding of mobility place name data are being met and at the same time, automated coding is providing additional benefits for production, evaluation and future development. Not only will coding accuracy and hence quality of small area migration data be improved, but a better understanding of the quality of place name data will be gained. For example, it will now be possible to know on what type of information code assignments have been made by examining actual write-ins and their codes; something which could not be done systematically in 1986, short of going to the actual questionnaires. As well, the automated system also yields new information on respondent usage of place name — for example, the extent

to which respondents report neighbourhood names instead of municipalities. Finally, the evaluation of these data from 1991 will provide input for an enhanced geographic reference file and the development of mobility questions for the next census.

ACKNOWLEDGEMENTS

The authors would like to express their thanks for the contributions of the following people from Statistics Canada: George Mori and Art Gardner, for their invaluable suggestions and guidance during the research phase of mobility autocoding; Rick Ciok, for incorporating special requirements of mobility into the QC system; Anna Rigakis for her input in the development of manual resolution strategies for mobility; and, Judy Kirk for her reports and observations on production-related issues. Thanks are also due to Audrey Miles for text processing.

REFERENCES

Ciok, R. (1991). The use of automated coding in the 1991 Canadian census of population, paper presented at 1991 American Statistical Association, Atlanta, Georgia.

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

Norland, J.A. (1989). Evaluation of mobility data form the 1986 Census, Statistics Canada, Demography Division (Internal Report, February 1989).

Norris, M.J. (1990). User's guide to 1986 census data on mobility, Statistics Canada, November 1990.

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

Norris, M.J., and Kirk, J. (1989). Research and testing of an automatic coding system for the mobility status variable using the ACTR system: Analysis report. Statistics Canada, 1991 Census of Canada, Ottawa (Internal Report, April 1989).

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

AN EXPERT ASSISTANT IN STATISTICAL ANALYSIS AND KNOWLEDGE ACQUISITION

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

SUMMARY

This paper describes the STATEX system, an expert assistant in statistical analysis and knowledge acquisition. The architecture of the system is based on the blackboard approach in order to allow communication between the user, the expert systems modules and the statistical modules. An integrated environment has been built, a miniature world with icons and graphs, in order to combine consultation, execution, and interpretation in one system. The user will be assisted by the system in all phases of statistical analysis. A data analysis model based on a theory of human perception has been developed. This implies that the necessary condition for distinguishing an object, to "see" it, is contrast, and that relationships among objects are determined by their association. This model served as the guide for the construction of a flexible and graphic statistical analysis system. The CSPAL program is used for statistical calculations. STATEX leads the user through the stages of quantitative analysis and creates an environment that changes dynamically with the needs of the user and the requirements of established statistical practice. The system then attempts an interpretation of the results in the idiom of the user. A statistical analysis method is proposed, and parts of the system can be used for arithmetic data related knowledge acquisition.

KEY WORDS: Expert assistant; Statistical system; Expert system; Modelling expert and user systems; Graphic interface.

1. INTRODUCTION

In their struggle for existence, organisms must be structurally adapted to their environments (Maturana and Varela 1984). In order to do this, they learn to make distinctions. The environment is very complex and changeable. Among their tools, humans have acquired one that serves to express their knowledge of their environment. This tool consists of the collection and analysis of numerical data, which uses mathematical methods and techniques. Powerful computer systems have been created to do the work of statistical calculation, but data analysis also uses more or less formal methods, including analytical strategies and rules for the interpretation of results, in order to reveal significant objects, facts, or events, depending upon the researcher's field. We have used artificial intelligence methods and techniques, and cognitive science, in order to model the knowledge of these methods and develop a knowledge acquisition system to assist in data analysis.

Artificial intelligence allows us to develop systems that represent knowledge in a given field, in this case data analysis. The cognitive sciences supply us with theoretical and practical bases, so that the developed tool will take into account the characteristics of the user as a function of the job that must be carried out. An artificial environment, a miniature world, has been created on computers in order to make knowledge acquisition easier.

---

[1] J. Muzard and E. Falardeau, Canadian Workplace Automation Research Centre, 1575 Chomedey Boulevard, Laval, Quebec, Canada H7V 2X2.

[2] M.G. Strobel, Université de Montréal, P.O. Box 6128, Station A, Montreal, Quebec, Canada H3C 3J7.

Current very powerful statistical programs are not very useful to users who are not familiar with the strategy, methods, and rules of the art of data analysis (Gale 1986). Large-scale accumulation of results is also difficult for users who do not have the background to be able to interpret the results of the analysis. The purpose of the STATEX project is: (a) to provide decision makers with a means to carry out analyses, in order to detect, compare, and predict trends; (b) to provide researchers and cognitive scientists with a knowledge acquisition tool; (c) to facilitate the analysis of data contained in databases; (d) to model knowledge in the data analysis field; (e) to create an expert system that will aid users in their work; (f) to increase knowledge in the field of workplace automation through the contributions of artificial intelligence and the cognitive sciences in a complex field; (g) to contribute to a better understanding of human thinking.

2. MODELS, METAPHORS AND HEURISTIC ANALOGIES

We make an act of distinction, in other words we separate an object from its context, in order to understand our environment (Maturana and Varela 1984). An object, entity, or unit contains implicit distinctions. The distinctions that we make are adaptive, and do not imply the real existence of the object. Using these categories, we produce schemas that establish relationships between categories, units of knowledge, or chunks. It seems that, in order to understand, we must be capable of making the necessary distinctions, have an initial understanding of the field, and be capable of progressing from a false concept to a new one (Carey 1990). Knowledge is "restructured" during the process of acquisition. There is a need to describe the restructuring that takes place during the process of acquiring complex knowledge. Thus, we can distinguish objects, situations, and events in the environment. Biological organisms need to orient themselves in relation to objects, identify situations, and predict events in order to increase their structural links to their environment as a survival measure. They possess mechanisms that allow them to perceive objects, make distinctions, create and order categories, establish relationships between events, and predict changes. Similarly, statistics involves describing objects, situations, and events, establishing the differences between them, and making associations or correlations between them. There is a progression in time and in complexity. Thus, this analogy involves initially a description of the objects or units, in order to then establish their differences and relationships.

3. ANALYSIS OF TASK COOPERATION BETWEEN TWO EXPERTS

3.1 Task Analysis

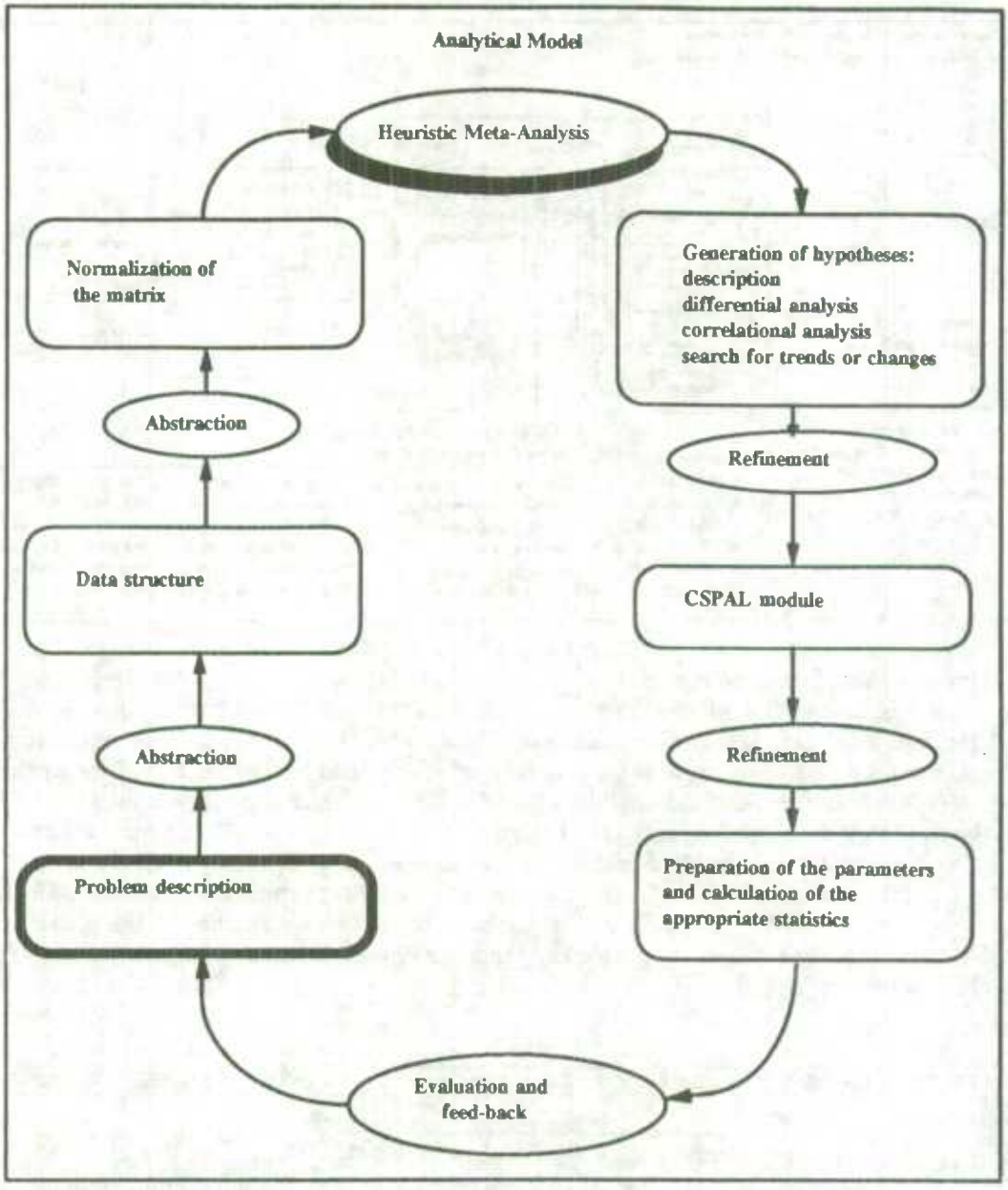
Consultation between a user and a statistical expert is characterized by communication and cooperation between two experts, in order to increase knowledge in the user's field. The client consults the expert because he is looking for assistance. He communicates the purposes of his research, and provides the data file. The statistician asks questions dealing with the data, the attributes of the problem, and the nature of the research. Subsequently, the statistician makes certain recommendations and participates in the observation and cleaning-up of the data, the structuring of the data matrix and the choice of strategy and methods, as well as the process of analyzing and interpreting the results in terms the user understands. The process of problem resolution is essentially a conversation during which reciprocal learning takes place. The client increases his understanding of data analysis, and the statistician increases his knowledge of the client's field. Each of the participants cooperates to reach the goal, which consists of acquiring new knowledge in the client's field.

3.2 The STATEX Model

We have established an analytical model (Fig. 1) on the basis of these observations. The first stage involves a complete description of the problem. A process of abstraction establishes the data structures, such as variables, groups, factors, rank and type of variables, *etc.* In fact, STATEX constructs an internal representation of the data by using the information on the data or meta-data (Hand 1991; Oldford 1990). Another process of abstraction and classification provides the structure of the matrix. Strategy hypotheses, such as correlational analysis, for example, are generated using the heuristic analogies described previously. The method to be used is established by a process of refinement, and the parameters for calculating the appropriate statistics are prepared. Then the results are presented and interpreted in the terms provided by the user while describing the problem. A new loop may be created on the basis of these results. This model describes how STATEX works.

The blackboard architecture used makes the process easier by taking into account the data, the system knowledge, the results, and the user's choices (Muzard, Falardeau, and Strobel 1991).

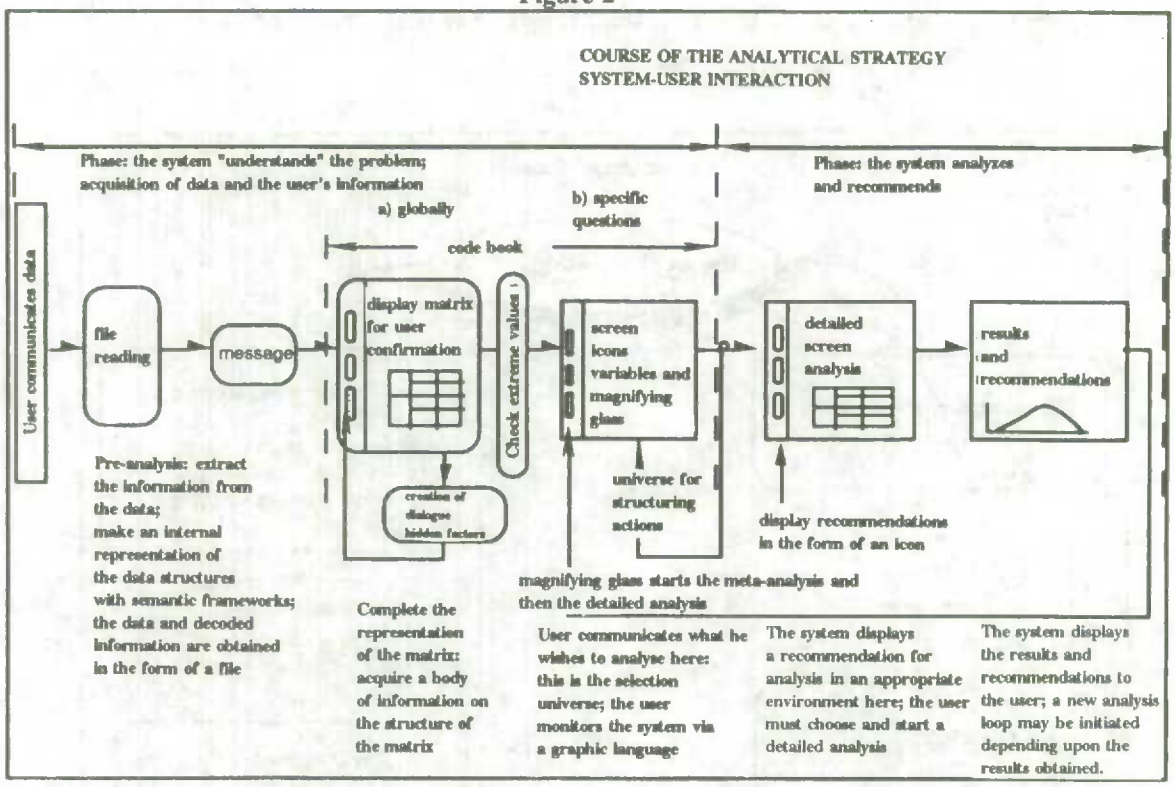
Figure 1



4. SYSTEM-USER INTERACTION

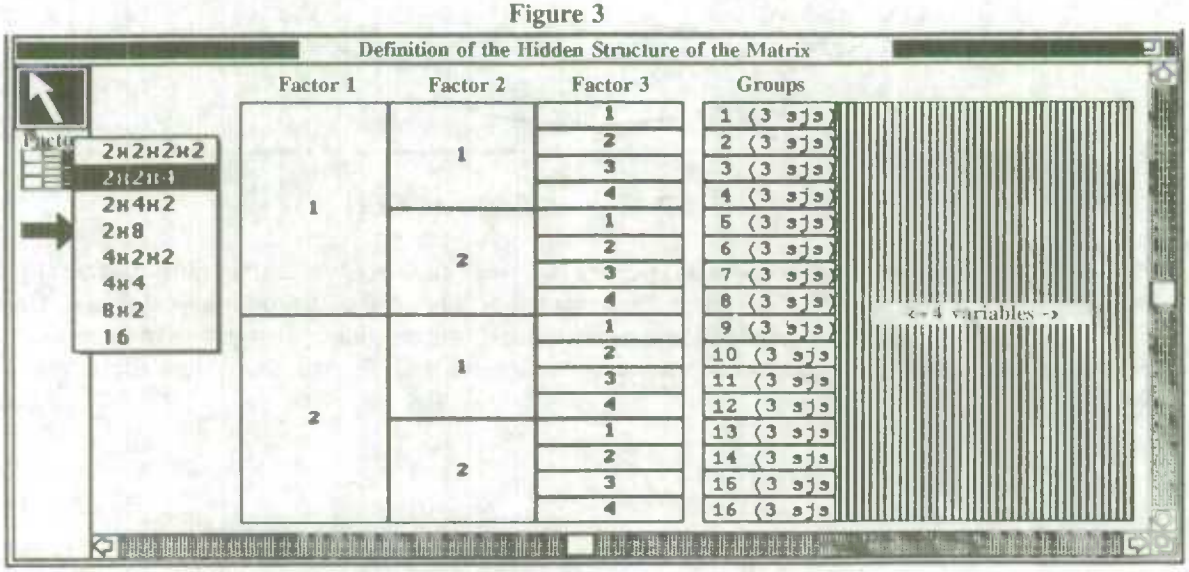
The progress of the analytical session is always under the user's control. We distinguish two general phases in order to describe the interaction between the system and a person who analyzes the data: first, the system acquires the data and information on the data in order to create an internal representation of the problem. This internal representation is the system's understanding of the problem. Second, the system analyzes and makes recommendations (Fig. 2).

Figure 2



At the start, the user is asked to provide information about himself, his name, and his knowledge. The system constructs a representation of the user and updates it as a function of the use of the system. The system adapts its interface to suit this information, offering more possibilities as the user becomes more adept at using the system. Next, the user is asked to communicate his data. The data file should be in the following format: one line per subject, variables and columns separated by tabs or spaces, and groups separated by alphabetical characters. STATEX is also capable of reading files created by SPSS if the control file is available. STATEX assumes that the data come from random samples. Before reading the data, STATEX opens a window with a picture of a matrix, and the user is asked to describe his research plan using, for example, a 2 x 2 x 4 pull-down menu. STATEX redesigns the matrix as a function of what is needed to provide the user with feedback and obtain his approval. In this manner, STATEX will be informed of the hidden structure of the matrix and the user will be able to see that the system has an adequate representation of his data at the same time. The user may also visualize his matrix (Fig. 3).

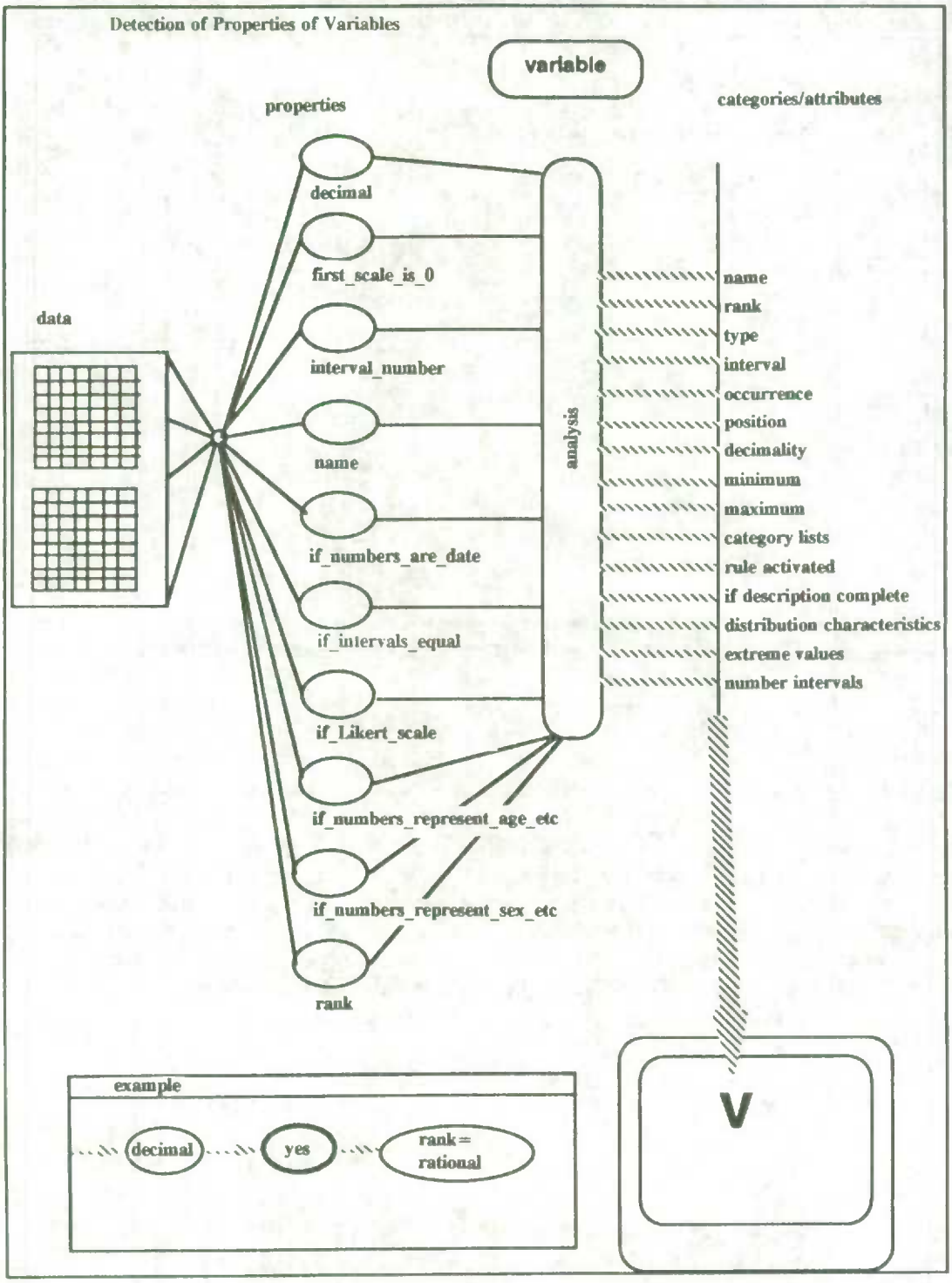
Figure 3



STATEX provides a graphic representation as often as possible, in order to stimulate the right hemisphere of the brain, which is responsible for pattern recognition and the overall view (Good 1983). We feel that this type of system can establish a closer relationship with user and work in harmony with mental representations in order to facilitate communication.

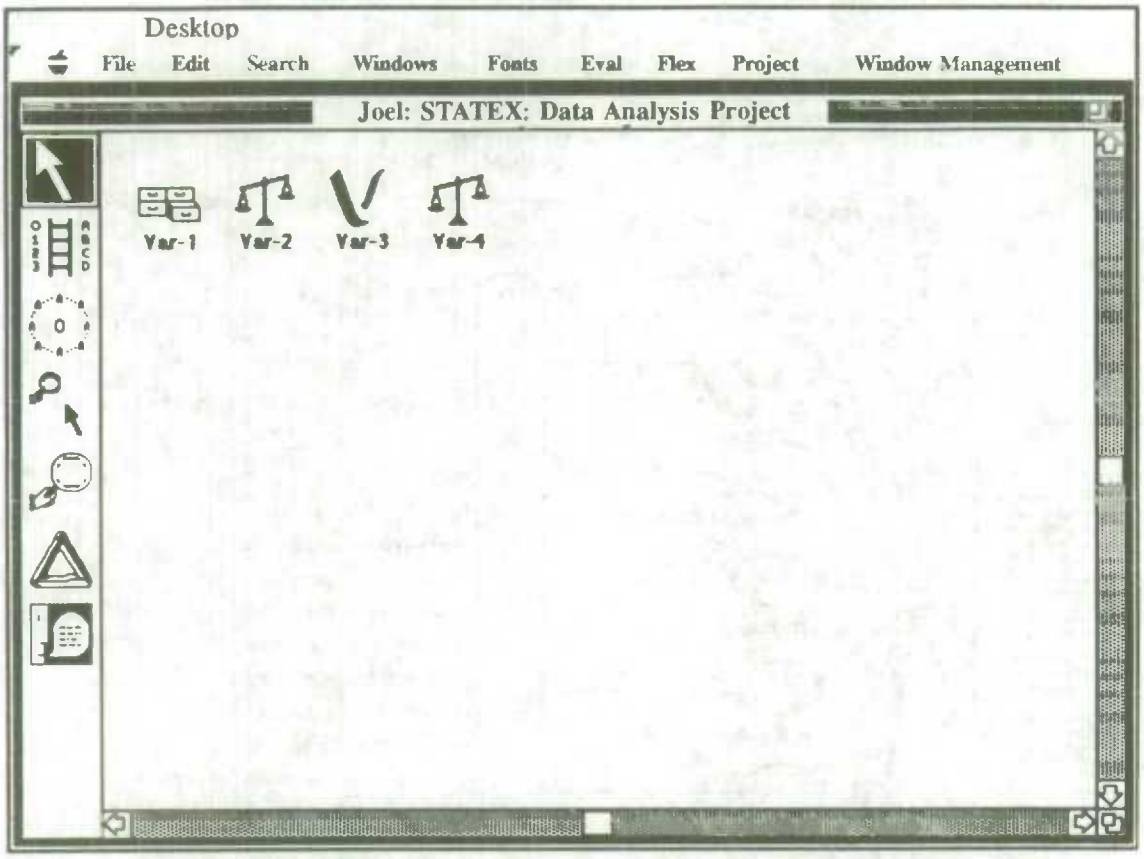
During pre-analysis, STATEX examines each variable in order to construct an internal representation of the data. This internal representation will be displayed using icons on the desktop. It will also be accessible to the basic information modules. STATEX will also use this information when displaying results. STATEX will attempt to distinguish the rank and type of each variable according to their properties (Fig. 4).

Figure 4



Screen display will be a function of this pre-analysis. For example, if the variable has a decimal number, a rule will be activated: if decimal = yes, then rank = rational. And if the rank is rational, the type will be classed as quantitative, or "suitable for averaging." Qualitative variables are shown as drawer icons, in an analogy to classification, and quantitative variables are represented by a scale, in an analogy to their being "suitable for averaging." STATEX uses the information on the data or meta-data to guide the analytical strategy. Once pre-analysis has been finished, STATEX will ask the user to complete the code book and display the desktop (Fig. 5).

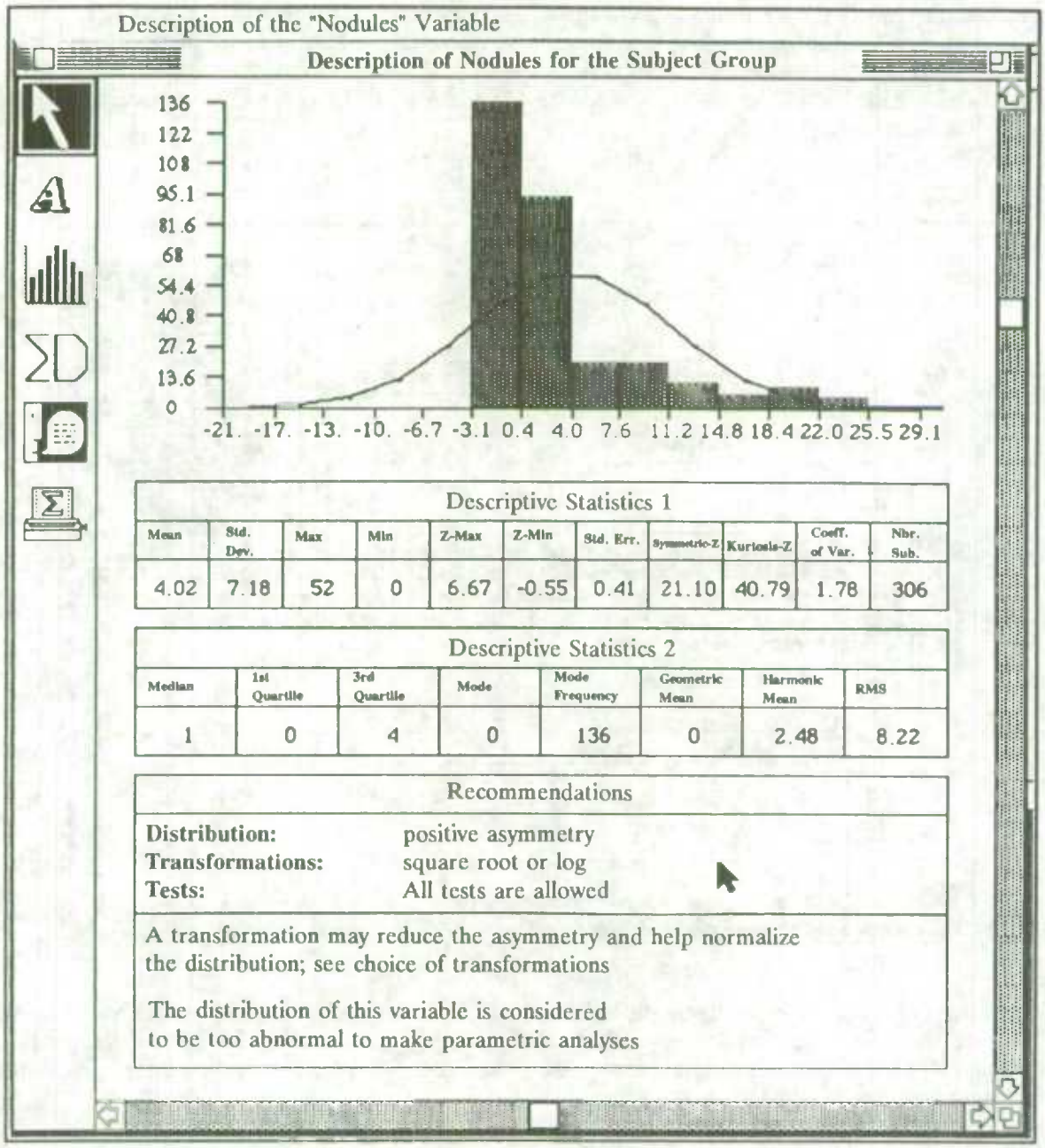
Figure 5



4.1 The Code Book

We have grouped all the activities that ask the user to provide information on variables, categories, factors, repeated measures, missing data, *etc.* under the heading "code book". For example, the user is asked to name the variables that are significant to him. During this stage, the user is asked to examine the data and see if there are any dubious data. STATEX shows him the data and the position of each dubious piece of information found to be to within more or less three standard deviations. The researcher may describe the variable (Fig. 6). He may also change the rank proposed by STATEX. If the users needs assistance finding a rank, STATEX will provide it.

Figure 6



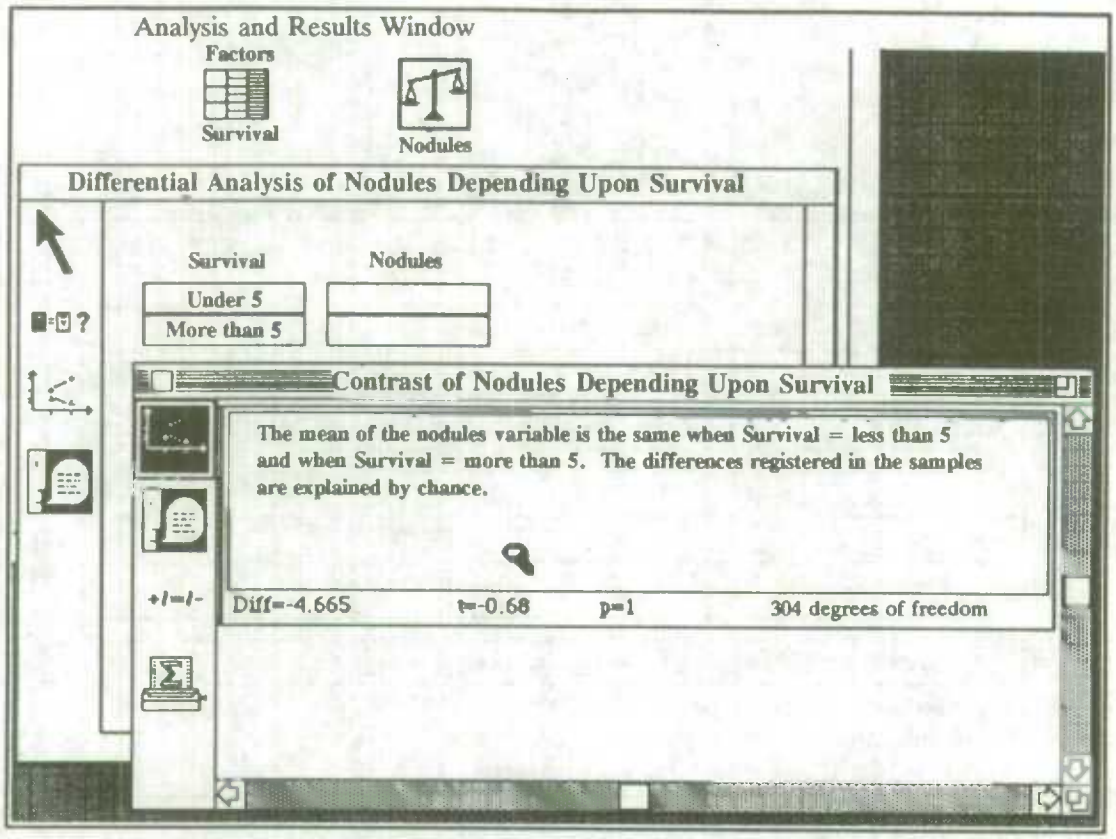
4.2 Meta-Analysis

Once the preparation of the data is complete, the user may save his project if desired. STATEX allows the user complete management of his work sessions through project management. A pull-down menu simplifies this operation. He may then select the variables and groups he wishes to analyze. To do this, he clicks on the corresponding icons with the "Arrow on magnifying glass" tool. Next, he starts the meta-analysis by clicking on the magnifying glass. During meta-analysis, STATEX classifies the selected data according to the matrix pattern, "many-groups-one variable suitable for averaging," for example. Two types of recommendations are made on the basis of this classification, the stage in progress, and the user's level: structural actions, or analytical strategies depending upon the heuristic analogies described earlier. The user may request explanations to guide him through the proposed alternatives. STATEX executes the choice. Structuring actions are, for example, "degrade a quantitative variable" for a correlational analysis with another classification variable.

4.3 Analytical Strategy

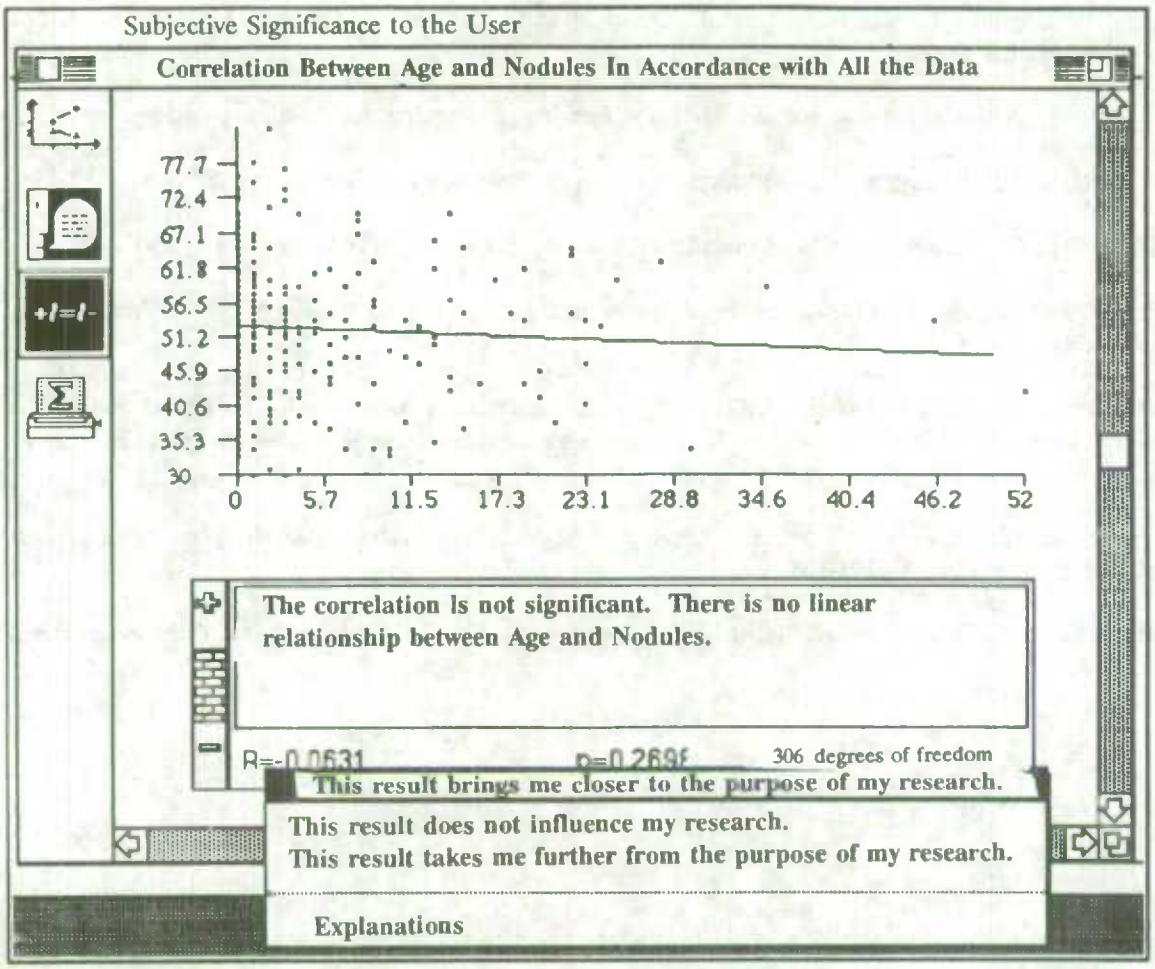
If an analytical strategy is proposed during meta-analysis, and the user makes that choice, STATEX starts a process that displays an analysis window based on the characteristics of the selected data, the statistical information in STATEX, and the available possibilities for analysis. This analysis window (Fig. 7) shows the user the form of the matrix that will result from his choice of variables and factors, and icons representing the appropriate statistical tools.

Figure 7



If the user selects one, the corresponding statistical calculation is started and displayed graphically and in text form. Then the user is asked to consider the results and indicate whether or not the result is significant for him. In this way, STATEX can organize the results and take them into account for the final report, to modify the user's status, and for later analysis (Fig. 8).

Figure 8



5. CONCLUSIONS

A prototype of a tool that is capable of effectively contributing to the resolution of statistical problems has been created. The prototype was developed in Quintus Prolog and Flex, Mac Prolog and MPW Fortran for CSPAL on a Macintosh. CSPAL (Strobel 1978) was modified and modularized in order to be integrated into STATEX. The statistical routines implemented in STATEX are basic statistics, such as Student's *t* test, correlations, analysis of variance, *etc.* This very interactive tool simplifies the choice of analytical strategies, the choice and calculation of the appropriate statistic, the presentation of results in graphic form, and the interpretation of results in the user's language. This is a tool capable of meeting the needs of a user who wishes to analyze data through an ergonomic interface, and capable of adequate support. The system facilitates the management of analytical projects. The experience accumulated in these analyses may be used to improve the system. It is possible to visualize a user's analytical strategy and follow his progress, in addition to collecting statistics on the use of the system. The system makes the production of reports easier, by allowing windows to be printed and displays to be transferred to a word processing program. It is very useful for interactive exploration of data and for making it easier to find the information contained in data. Working with this tool allows the user to improve his understanding of the field of data analysis.

REFERENCES

Carey, S. (1990). Cognitive Development. (Eds. D. Osherson and 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. In Society for Artificial Intelligence and Statistics.

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

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

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.

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

A MULTIVARIATE APPROACH TO RESPONDENT LOCATION

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

ABSTRACT

This paper reports on the development of a multivariate approach to the problems of locating respondents and data linkage. The paper describes how addresses, postal codes, telephone numbers, place names and legal land descriptions can be used for locating respondents thus linking data to the appropriate geostatistical area. A strategy for combining these spatial elements to improve locational specificity and robustness is detailed. Decision rules to enhance fault tolerance and for tie breaking are outlined. Finally, the potential utility of the multivariate approach to various survey operations and topics for further research are put forth for consideration.

KEY WORDS: Geocoding; Geographic referencing; Data linkage; Automated coding.

1. INTRODUCTION

In census and survey operations, defining the whereabouts or location of a respondent is necessary for correct linkage of the data, data quality evaluation, reconciling the results between different censuses and/or surveys, and a variety of other tasks. The same requirement is true for conversion of administrative data to geostatistical reporting units for statistical report preparation.

Traditionally, the tasks of geographic editing and data linkage have been time consuming, manual processes. Beginning in the late 1960s, tools such as Statistics Canada's Postal Code Conversion File and Area Master File and the U.S. Bureau of the Census' DIME, and now TIGER, files have enabled automation of parts of these tasks. Univariate approaches have been most often used, with the postal code being a common choice as the locational key (Wilkin 1988a; Wilkin 1988b; Nadwodney 1989). More recently, multivariate approaches have attracted interest (Norris and Kirk 1989; Schneider 1987; Yergen 1987), as they offer potential for better spatial resolution and more robustness (*i.e.* tolerance to missing or erroneous data).

This paper reports on research at Statistics Canada on the development of a multivariate approach to the problem of respondent location and data linkage. The paper begins with a discussion of how different geographic elements can be used to locate respondents and link their data to an appropriate geostatistical unit, and then how they can be combined to improve the respondent location process. Alternative approaches are reviewed. Decision rules to resolve uncertainties arising from improper responses are detailed. Figures are provided to illustrate the spatial resolution of univariate and multivariate processes. Finally, the potential utility of the multivariate approach to various survey operations and items for future research are put forth for consideration.

---

[1] L. Li and G. Deecker, Geography Division, Statistics Canada, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A OT6.
P. Daoust, Business Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A OT6.

2. GEOGRAPHIC ELEMENTS FOR RESPONDENT LOCATION

2.1 Addresses

The respondent's address is one of the most commonly available geographic element in surveys and administrative files. Responses typically reflect the postal address of the respondent. The characteristics of postal addresses are detailed by Deguire (1988). Generally, in urban areas, the postal address refers to a dwelling, although super mailboxes, postal boxes, general delivery and other forms of nonlocationally specific addresses do exist. In rural areas, postal addresses are less informative. Rural route addresses, such as 'RR3, Almonte, Ontario', provides lower spatial resolution. Use of postal boxes and general delivery addresses, which only link the respondent to the post office, is also more prevalent.

In urban areas, addresses can be used in combination with street network files, like Statistics Canada's Area Master File or the US Bureau of Census' DIME and TIGER files, to locate respondents (Yergen,1987,221-230). The process is carried out by matching the address of the respondent to the address range and street name which is attached to each street segment in the street network file. Today, many commercial Geographic Information System (GIS) software packages, like MAPINFO (Mapping Information Systems Corporation 1989, 4-63) and ARC/INFO (ESRI 1989), include specific programs for this task.

In rural areas, use of postal addresses for locating respondents is much more difficult. This is due to the poor spatial resolution/specificity of rural addresses, and the lack of street network files for rural areas in Canada.

Another method of using postal addresses for respondent location is to convert the address to a postal code, and then use tools such as Statistics Canada's Postal Code Conversion File (Geography Division 1989) to link the respondent to an Enumeration Area or several Enumeration Areas. Conversion of postal addresses to postal codes is usually done in a two step process. First, the raw addresses are analyzed to derive an address search key (Deguire 1988; Yergen 1987, 221-231). The second step takes the address search key and retrieves a postal code from the Postal Code Master File, a file of all postal codes from Canada Post. The whole process is similar to going to the post office and looking in the postal code directories to find a postal code for a given address.

2.2 Postal Codes

The postal code, a six character, alphanumeric code is found on a large majority of survey responses and administrative files. It has been widely used as a locational key for survey processing (Nadwodney 1989), data analysis (Wilkin 1988a) and a host of other applications (Maloney 1988; and Nadwodney,1989)

As noted above, a postal code can be used with Statistics Canada's Postal Code Conversion File (PCCF) to link respondents to an Enumeration Area or Enumeration Areas. The PCCF serves as a look-up table to facilitate the conversion.

The spatial resolution of postal codes vary significantly between urban to rural areas. In urban areas, a six digit postal code can be as precise as a block face, a large apartment building or even one floor of an office building (Canada Post 1983, 7). In rural areas, a postal code represents a service area, which can encompass parts of several geostatistical areas. Table 1 shows the comparative resolution of urban and rural postal codes by province.

TABLE 1: AVERAGE NUMBER OF EAs or PART EAs PER POSTAL CODE BY PROVINCE

	Can.	B.C.	Alta.	Sask.	Man.	Ont.	Que.	N.B.	N.S.	P.E.I.	Nfld.
Urban	1.04	1.04	1.03	1.03	1.04	1.04	1.03	1.04	1.03	1.02	1.05
Rural	4.31	5.13	6.16	4.92	3.90	4.31	2.99	4.29	3.92	5.18	2.14

2.3 Telephone Numbers

A telephone number is composed of three components, a three digit area code followed by a three digit exchange and then the four digit local. The area code delineates large regions. For example, all of Manitoba is served by the area code, 204. A telephone exchange delimits a relatively small area, usually served by a single telephone switching station, and often equal to a municipality in size, but not necessarily coincident with the municipal boundaries. The four digit local is unique to each telephone customer, except for those served by party lines.

For locating respondents, the area code and exchange are the most useful at this time, since they delimit specific service areas, and each exchange is unique within its respective area code. The local which may facilitate linkage of the respondent to a household holds future promise, but data unavailability and cost limits its present utility at the national level.

In contrast to postal codes, telephone exchanges provide relatively better spatial resolution in the rural area than the urban area. Table 2 shows the average number of EAs for an exchange by province.

TABLE 2: AVERAGE NUMBER OF EAs PER TELEPHONE EXCHANGE BY PROVINCE[2]

	Can.	B.C.	Alta.	Sask.	Man.	Ont.	Que.	N.B.	N.S.	P.E.I.	Nfld.
Urban	---------- Data Not Available ----------										
Rural	5.84	4.18	7.96	7.88	---	7.13	5.29	5.22	5.13*		2.30

* for Nova Scotia and P.E.I. together.

2.4 Place Names

A place name, denoting the place of residence of the respondent, is often available from survey or administrative files. Where such places cannot be located as part of an address search or postal code based search, use of the place name directly for respondent location can be tried. The keys to making place names useful for respondent location lie in the development of a place name reference file, to provide the XY coordinates or the appropriate geostatistical unit for each place, and an appropriate parsing and text matching strategy.

A place name reference file, with approximately 160,000 entries, is being compiled at Statistics Canada from previous census data, names associated with standard geostatistical areas, unincorporated places and other relevant sources (Norris and Kirk 1989, 12-13). An additional source which requires further investigation is the Toponomic File from the Geographic Services Division of Energy, Mines and Resources Canada which incorporates official place names from provincial gazetteers.

Place name matching has some similarities to the problems of automated text retrieval, automated coding and record linkage. The large body of literature on automated text retrieval (Ashford and Willet 1988; Saffady 1989; Stanfill and Kahle 1986; Bouchard 1979) offers insights on useful approaches and methodologies including compromises between recall and precision and methods for incorporating fault tolerance into searches. Means of search optimization are examined by Sellis (1988), Wu and Burkhard (1987), Ramamohanarao and Lloyd (1983) and Bouchard (1979).

Place name matching, however, differ substantially from text retrieval from free text documents in that the uniqueness of place names is context sensitive, (a name may only be unique within its locale; sometimes towns

---

[2] These averages should be treated as approximate estimates, as some Enumeration Areas included more than one exchange, thus were counted more than once. Further the data source for the telephone information was the 1986 Census of Agriculture which does not have many respondents from urban Enumeration Areas. Thus, it does not have full coverage of all telephone exchanges in the urban areas.

with the same name occur quite close together) and aliases are common. In Canada, the process is further complicated by the presence of English, French and aboriginal names.

The work of Norris and Kirk (1989), and the US Bureau of the Census (Schneider 1987) provides useful insights into automated place name coding. Norris and Kirk, using the Automated Coding and Text Recognition system (ACTR) (Development Division 1989) which incorporates an entropy based algorithm for match determination, have achieved success rates of approximately 80% with nation wide samples. Schneider (1987, B-3 to B-6), in some of their preparatory research for their 1990 census reported varying success rates, approximately 87 % for Los Angeles and 44 % for Mississippi.

The current research builds upon the general approaches used by Norris and Kirk (1989), Schneider (1987) and Yergen (1987). The process is as follows: Input names are first parsed, then matches are sought based on the significant words within the name string. Direct matches are then sought for each of the input variables. For variables which are not directly matched, closest matches are retrieved. An intersection set is then defined by cross comparison of the candidates which were retrieved for each input variable. If multiple candidates remain in the intersection file a set of rules can be applied to eliminate unsuitable candidates from the set by considering dissimilarities between the characteristics of the places being matched. A distance decay function is also used to determine the likelihood of each candidate being the place of the respondent, if appropriate.

The spatial resolution of locating respondents through place names varies greatly with the nature of the question which is posed to the respondent. For national applications, such as for migration studies, location of subjects to a municipality seems to be the desired resolution (Norris and Kirk 1989). For census data quality studies and data editing, more specific locations, such as an individual Enumeration Area, may be desired.

In 1986, Canada had a total of 6009 legally incorporated municipalities. Approximately 41% of them encompass only one Enumeration Area. 17% of the municipalities encompass two EAs, 9% includes three EAs, and 33% includes four or more EAs. From these figures, it is easy to understand why municipal names, which are often the same as postal office locations, serve as a useful key for geographic data linkage.

Unincorporated places which are generally sub-municipal entities are noted as part of the census field process. They provide an additional source of information for determining the location of respondents for data linkage.

From some of our tests, municipal and unincorporate place names have yielded deterministic mapping of respondents to an EA in up to 70% of the assignable cases in rural areas, given reasonably error free inputs. In urban areas, where there is a predominance of one to many relationships between a municipality and EAs, deterministics assignments at the EA level are very few.

The results of our research and the experiences of Norris and Kirk (1989) and Schneider (1987, B-7a - B12) indicate that the factors responsible for unsuccessful matches or mismatches include: abbreviations, spelling errors, alias names, entries in to incorrect fields, incomplete names, historical names which have since changed, and ambiguous or non-unique names.

2.5 Legal Land Descriptions

The nature of legal description of land units varies significantly from province to province, within some provinces, and also between urban and rural areas. The heterogeneity reflects the colourful history of Canadian settlement, with the interplay of French and British influences in the east, the more orderly settlement of the Prairies and the strong dominance of the rugged physical environment in British Columbia. This heterogeneity makes it very difficult to formulate a generalized question to obtain legal land descriptions, and to devise a standard structure for capturing respondents' answers for national surveys and censuses.

For example, township-range-section data from the Prairies and the B.C. Peace River District are numeric descriptors with occasional alphabetic modifiers. The descriptions for all other provinces are alphanumeric. Responses from Quebec, the Maritimes and British Columbia are extremely variable. Variations which can be expected include answers which do not follow the expected structure and sequencing of variables, inclusion of

street addresses, property identification numbers, town names, seigneury names, lot numbers from plan of subdivisions and others.

In spite of the heterogeneity of the data, the legal land description provides a powerful means of locating respondents in rural areas, since it is an extremely specific descriptor of the location of the respondent, and is commonly known in many rural locales. At the most precise level, an individual ownership parcel can be identified, however, both the compilation cost and the data volumes of the required reference file would be very high for all of Canada. At a more aggregate spatial level, the township concession lot or township-range-section represents a more practical scale for national applications. Township concession lots or its equivalents in eastern Canada are typically 100 to 200 acres. In the Prairies, a township section covers approximately 640 acres or a square mile.

To use the legal land descriptions for respondent location, dual strategies were tested. For the Prairies and the B.C. Peace District, where responses usually adhere to the expected data format, direct matching on the full section-township-range and meridian was often successful. When a direct match is not possible, a partial match without the section and/or the range was attempted.

For some of the other provinces, particularly Quebec and British Columbia, where the input is much more variable, the input position of the response can not considered to be a significant factor in match evaluation. For example, if Oxford was reported in the "County" variable, a match on Oxford County would not be given precedence over Oxford Township. The two competing candidates would be equally considered and other tie breaking rules, incorporating a distance decay function and characteristics of the places, would be applied to eliminate extraneous candidates from consideration.

3. MULTIVARIATE STRATEGY

The above discussion outlines five univariate approaches to the respondent location problem. Combining all or several of the approaches holds promise to make the process more robust, as well as improving the spatial specificity of the assignments by taking advantage of the strengths of one assignment path to cover the weakness in another.

Several researchers have tested multivariate strategies for respondent location or similar problems (Norris and Kirk 1989; Schneider 1987; Yergen 1987; Drew, Armstrong and Dibbs 1987). The current research extends the general approaches used by the above-noted authors to include use of telephone exchanges and legal land descriptions. Error tolerance is introduced into the individual candidate identification process, with the incorporation of rules for checking and refining missing or erroneous telephone area codes and the province designator in the address. Further, use of a separate variable as a probability indicator for tie breaking amongst final candidates is tested for locating farms and farmers for the Census of Agriculture.

The first step in the process is to refine the input data to ensure essential components are available to maximize its utility/suitability for subsequent processes. The refinements are carried out by taking in a respondent's postal address, postal code, telephone number and legal land description. The input data is refined by correcting missing province data, which is critical for address processing, by using the postal code and telephone area code to retrieve the province from a reference file. Telephone area codes are verified against a list of all known area codes from the telephone companies. Incorrect or missing area codes are corrected by using the post office name and province to retrieve the appropriate area code for the respondent from a reference file. Addresses are parsed, and a postal code retrieved for each address using the PCODE program (Research and General Systems Subdivision 1987). Separately reported legal land descriptions are parsed to remove noise and facilitate identification of the most significant components for subsequent matching.

The second step is to retrieve candidate locations/EAs where the respondent may be found via each of the input locational elements - postal code, address, telephone exchange and legal land description. This is accomplished via a series of straight look-ups on a postal code to EA file (the PCCF), telephone exchange to EA file, and county-township-concession-lot to EA file.

Having retrieved up to four sets of candidate locations/EAs, a cross checking process is invoked to identify the candidates which are common to the highest number of the input sets. The resultant intersection set defines the potential locations where the respondent should be found. This cross checking process is useful for weeding out erroneous input, since the process requires that several independent geographic elements point to the same potential locations before they are included in the intersection set.

An indication of the spatial resolution of such a multivariate process is shown in Table 3.

TABLE 3: AVERAGE NUMBER OF EAs ASSOCIATED WITH UNIQUE COMBINATIONS OF POSTAL CODE AND TELEPHONE EXCHANGE NUMBER BY PROVINCE[2]

	Can.	B.C.	Alta.	Sask.	Man.	Ont.	Que.	N.B.	N.S.	P.E.I.	Nfld.
Urban	---------- Data Not Available ----------										
Rural	2.65	2.80	2.82	2.81	2.07	2.61	1.99	2.49	2.28	2.93	1.50

Note: Estimate based on data from the 1986 Census of Agriculture, which under represents rural nonfarm areas.

In respondent location applications, a definitive location is often required. For migration studies, the required resolution may be a municipality. For data quality studies, EA accuracy may be desired, and prioritization of EAs in which the respondent may be found is useful for reducing the search space for in-depth manual searches or further automated matching on specific respondent characteristics, such as sex and birth date. Therefore, if multiple candidates are identified in the intersection set, a mechanism/process is required to determine the likelihood of the respondent in each of the candidate locations. The process which has been adopted brings in an associated variable as an indicator of the probability of the respondent being found in the given location. For example, for locating a given farmer, the number of farms in each of the candidate EAs may be used as probability indicator to rank the EA candidate for more stringent searches.

In the future, the performance of the process should be improved by the expansion of the rule set to consider other key characteristics of the respondent in order to better define specific candidate locations with the same characteristics. In the case of the Census of Agriculture, the expanded rule set may consider the crops which are being grown by the farmer and other characteristics such as farm size.

4. MACHINE LEARNING TO IMPROVE SYSTEM PERFORMANCE

Compilation of a reference file of legal land units (township-concession-lots) for Canada is a time consuming and costly process. As part of our research, a process is being developed to mimic the on-the-job knowledge acquisition capability of staff engaged in manual editing.

The process assimilates previously unknown land unit descriptions and their associate EAs from clean records; checks the unknown parts of the geographic descriptors against existing information and keeps a running tally of the number of times each of these relationships is reported. Once the tally exceeds a threshold value, which is being empirically defined from past census data, the given relationship is entered into the land unit to EA reference file to improve its coverage.

Test results indicate that this process works well for areas with relatively clean data, and where there is a relatively large number of respondents present within each of the areal units for which the spatial relationship is to be learned. In areas where the land descriptions are highly variable and responses differ considerably, such as in the province of Quebec, the required number of repetitious responses needed to elevate a particular spatial relationship to the "knowledge base" is seldom met. Further research to improve understanding of the characteristics of the responses, leading to the development of improved parsing and phrase substitution strategies would likely improve the performance of such as "learning" process.

5. CONCLUSIONS

Many areas of survey taking, collection and data analysis need to locate their subjects and link their data to a set of geostatistical areas. The current research has demonstrated that a multivariate approach can improve the spatial resolution of locational edits and improve the robustness of the process. The approach, which is being developed, draws from strategies and methods which are used in automated text retrieval, record linkage and the work of others in multivariate geocoding. Applications have been implemented for the Census of Agriculture geographic edit, and for quality studies within the Census of Population and Housing.

At this time, the implemented systems are still in production, thus definitive results on their performance are not yet available. Preliminary indications are that they are performing very close to expectations. Analysis of these results is planned for the near future.

From our experience so far, a number of opportunities for future research have been identified. Further research is required into the derivation of postal code boundaries for rural areas. Our current lack of detailed knowledge of place names and land description across the nation are also impediments to better system performance. Consultation with field personnel to examine other sources of geographic information and maps which may be available, and the integration of such information into the automated process would also be helpful. Focused research to better our understanding of the relationship between respondents and characteristics of various geographic locations and appropriate distance decay functions for various interactions are needed for expansion of the rules for elimination of extraneous candidates. On this latter topic, adoption of some of the concepts and approaches inherent in donor imputation may prove valuable. Investigation of artificial intelligence capabilities may also hold promise for development and enhancement of rules in real time leading to overall improvements in the processes.

ACKNOWLEDGEMENTS

The majority of this research was conducted under the Statistics Canada's Internal Sabbatical Program. Additional funding and data was also provided by the Census of Agriculture. Heartfelt thanks are due to Catherine Cromey, Marcelle Dion, Dave Dolson, Doug Drew, Rennie Molnar, Mary-Jane Norris and Peter Schut for their advice and guidance. The hard work of the research and system development team under Danny Wall and Joe Gomboc were also instrumental to the achievements. The authors gratefully acknowledge all of their support and contributions.

REFERENCES

Ashford, J., and 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. *Survey Methodology*, 14, 2, 317-325.

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. and 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., and 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., and 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. (July 22, 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., and 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., and MacDonald, R. (1988). *Potential uses of postal codes with vital statistics data*. Health Division, Statistics Canada, Ottawa.

Wu, C.T., and 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

Geographic Innovations in Data Collection

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

APPLICATIONS OF TIGER TO THE 1990 CENSUS: BENEFITS TO DATA ANALYSIS AND PROSPECTIVE APPLICATIONS TO SURVEY TAKING

R.W. Marx[1]

ABSTRACT

The TIGER System means different things to different people. During the past four years, the TIGER System has fulfilled the geographic support functions of the 1990 decennial census for which the Geography Division of the United States Bureau of the Census designed it. It also has provided most of the geographic and cartographic products needed to complete the tabulation of the collected data and make those data useful to the numerous constituencies that carry out the myriad tasks that define our lives. Simultaneously, work is underway to define a framework for the future of this bold new product -- and this future looks bright. Similar planning is going on in offices and institutions across the United States and around the world. This is true especially in the context of geographic information system (GIS) applications involving the machine-readable products of the TIGER System and the data products of the 1990 decennial census.

KEY WORDS: Automated cartography; Decennial census; Future; Geographic information systems; TIGER System; U.S. Census Bureau.

1. INTRODUCTION

The TIGER System means different things to different people. First and foremost, it is an innovation (Carbaugh and Marx 1990). The focus of most earlier articles about the TIGER System has been on its development and use within the United States Bureau of the Census (Marx *et al.* 1990); these articles have described a series of massive geographic and cartographic production operations performed in conjunction with the 1990 decennial census of the United States. This article summarizes the salient points of that history and discusses several aspects of where the future might lie for the TIGER System, both within the Census Bureau and as a tool for geographers, cartographers, and others developing automated geographic and cartographic systems. It also describes the potential for using this new system -- and the associated 1990 census statistical data -- to perform spatial analysis using geographic information system (GIS) technology.

2. WHAT IS THE TIGER SYSTEM?

The Geography Division, one component of the Census Bureau, developed the TIGER System in response to a major goal the agency set in 1981:

> "to automate the full range of geographic and cartographic support processes in time to serve the data collection, tabulation and dissemination needs of the 1990 decennial census -- the Bicentennial Census of the United States."

This decision gave the Geography Division six short years to build a computer data base containing every known street and road in the United States (a task accomplished only with significant assistance from the United States

---

[1] R.W. Marx, Geography Division, Bureau of the Census, U.S. Department of Commerce, Washington, DC 20233 U.S.A.

Geological Survey -- USGS), the name of each, and the range of address numbers located along each segment of every street in the 345 largest urban cores of the United States; six short years to include all the railroads in the United States along with all significant water features and their associated names; and six short years to enter and verify the boundaries, names, and numeric codes for all the geographic entities used by the Census Bureau to tabulate the results of both the 1980 and 1990 decennial censuses. Many of these same entities are used to tabulate the results of the economic and agriculture censuses as well (see Chart 1).

This development task is complete; the overall goal, met -- and on time! Is the TIGER data base perfect? Of course not! But most of the information it contains is correct and the information is much more up-to-date than the information on the traditional maps and in the GBF/DIME-Files that the Census Bureau used for the 1980 census. More importantly, now all this information is in the computer! As a result, the Census Bureau can enter changes easily as people identify needed updates.

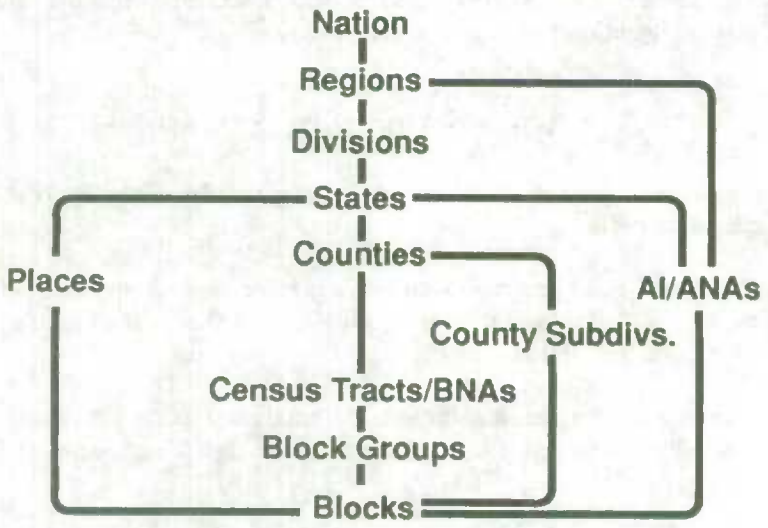
3. HOW HAS THE CENSUS BUREAU USED THE TIGER SYSTEM?

The success of a census or sample survey rests not only on collecting data but also on linking those data to geographic areas. At this juncture, the Census Bureau has relied on TIGER System products and services to help compile, validate, and geographically assign (geocode) the 1990 decennial census housing unit and group quarters address list; to conduct and evaluate the data collection operations of that census; to tabulate the collected data and make those data useful to the many constituencies that by law, or choice, rely on the results of the decennial census to carry out the myriad tasks that influence our daily lives (Tomasi 1990); and to provide a wide variety of "public" products in both paper (see Chart 2) and machine readable (see Chart 3) formats that match, in exact geographic detail, the geographic entities used to report the statistical data of the 1990 census (LaMacchia, Tomasi, and Piepenburg 1987). These TIGER Extract products support the GIS applications of people and organizations that will make use of the 1990 census data products as the Census Bureau releases them over the next two years.

4. WHAT LEVEL OF GEOGRAPHIC DETAIL DOES TIGER PROVIDE?

The Census Bureau uses a basic geographic hierarchy for most of its statistical programs (see Figure 1 and also Chart 1). The entities in this hierarchy generally share a "nesting" relationship; that is, each major entity generally contains multiple smaller entities.

Figure 1: The Census Bureau's Basic Geographic Hierarchy



4.1 The "Higher Level" Entities

The hierarchy starts with the United States -- the entirety of the 50 states plus Washington, DC. Dividing the United States, there are 4 regions (that are groups of states) that are divided into 9 divisions (that also are groups of states). Then there are the 50 states and Washington, DC. The Census Bureau built the TIGER data base to include, in addition, Puerto Rico and the other outlying areas in which the Census Bureau helps conduct the censuses of population and housing -- American Samoa, Guam, the Northern Mariana Islands, Palau, and the Virgin Islands of the United States.

Dividing the states and those other statistically-equivalent entities, there are more than 3,200 entities that people generally call counties. Most of these entities cover fairly large geographic areas, which means that they provide fairly coarse "resolution" for the characteristics for the people, housing units, farms, and businesses of the United States -- useful for analysis mostly on a national or global scale. These counties are extremely important for data analysis purposes because the Census Bureau also makes statistical data available for them from the economic and agriculture censuses it conducts every 5 years; censuses that include topics as diverse as manufactures, retail trade, service industries, wholesale trade, construction industries, mineral industries, transportation, enterprise statistics, minority and women-owned businesses, and all sorts of agricultural information. Many other governmental agencies and private organizations also collect and report data for counties, resulting in a "rich" data series for them.

At the next level in the Census Bureau's geographic hierarchy there are more than 60,000 units of local government and statistically equivalent entities subdividing those 3,200-plus counties -- townships, cities, villages, census designated places, and so forth. The relationship of places to the hierarchy is not quite as neat and tidy in all instances, as places do not provide "wall-to-wall" coverage the way the previously discussed entities do. In terms of data cells, these "governmental" tabulation units provide more than a twenty-fold increase in entities over counties. This still offers fairly coarse resolution in a GIS context because many subcounty governments also cover very large areas. The good news is that demographic data are available for all these entities and economic data are available for the most populous of them.

4.2 The "Lower Level" Entities

To zoom in closer, the hierarchy includes more than 62,000 census tracts -- and their cousins, the block numbering areas -- for the 1990 census. There has been a big increase in the number of these entities since the 1980 census because these entities now cover the entire United States and its territories. These census tracts and BNAs "cut-up" the large governmental units into geographic "chunks" that average about 4,000 people. This starts to make the resolution of the 1990 census tabulation entities much more useful in most GIS applications. As an example, at the county subdivision/place level in the Census Bureau's geographic hierarchy, there is one entity called "the City of Los Angeles;" at the census tract level, Los Angeles has 737 entities or which the Census Bureau tabulates detailed statistical data. These census tracts are relatively small geographic areas within the much larger City of Los Angeles.

The census tracts and BNAs are themselves subdivided into approximately 229,000 block groups that further segment the governmental units for purposes of statistical data presentation. At both the census tract/BNA and the block group levels (as at all previously discussed "higher" levels), a GIS user has access to the full range of decennial census statistical data -- those data collected from every person and about every housing unit and those data collected from only a sample of the population and their housing units.

Finally, for the 1990 census the Census Bureau identified and numbered more than 7 million census blocks -- people polygons -- covering every portion of the United States. The Census Bureau has tabulated the data it collected from every person and about every housing unit for each of these blocks. As with the census tracts and BNAs, there has been a huge increase in the number of these entities since the 1980 census when block-level data were available primarily for the urban cores of metropolitan areas. These "millions and millions" of blocks provide a very fine-grained resolution to the demographic data sets available from the Census Bureau. They are defined by the roads, rivers, railroads, and governmental unit boundaries that are the underlying "geometry" of the TIGER data base; the same features that already comprise at least one layer in most people's GIS.

4.3 Special Purpose Entities

The Census Bureau's geographic structure includes several other categories of entities for which it tabulates many of the statistical data items it collects; these entities provide varying levels of geographic resolution and coverage (see Chart 1).

- Metropolitan areas generally are composed of a county containing a large city or major urban center, plus adjacent counties that are linked economically to the central one. In the New England states, metropolitan areas are composed of clusters of cities and towns instead of clusters of counties. Collectively, all the metropolitan areas comprise the "metropolitan population of the United States;" everyone outside these metropolitan areas is "non-metropolitan."

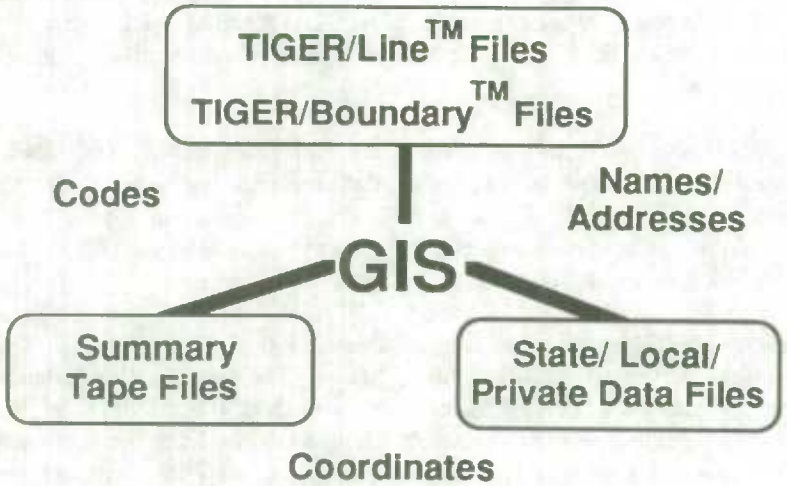
- Urbanized areas are quite different. Whereas metropolitan areas are defined as collections of governmental units -- that can vary widely in population density and areal extent -- urbanized areas are defined strictly on the basis of population density. Thus, every metropolitan area contains at least one urbanized area at its core -- some contain two or more. These urbanized areas generally cover much smaller geographic areas than do metropolitan areas -- but at a much higher average population density. A number of urbanized areas also exist in counties that do not qualify for definition as metropolitan areas. Collectively, all people in these urbanized areas, together with all people in other incorporated and census designated places of 2,500 population or more, comprise the "urban population of the United States;" everyone else is part of the "rural" population.

- The 1990 census also includes data tabulations for American Indian and Alaska Native areas, Congressional Districts, and soon will include data for ZIP Codes plus a wide variety of other entities.

5. WHAT MAKES TIGER USEFUL FOR GIS APPLICATIONS?

Numeric codes for each geographic entity, that match the entity codes used in the Census Bureau's statistical data products, are the most important contribution of the TIGER data base to the "world of GIS." Using these codes, a GIS can link the statistical data directly with the network of features bounding the millions of 1990 census blocks, and thus, to the many other data sets that increasingly are available; state and local data sets as well as the data sets that flow from the decennial, economic, and agriculture censuses (see Figure 2).

Figure 2: GIS and the TIGER System



In earlier days, when data analysts used geographic entity names and codes to make maps showing the distribution of various data items, they often performed their analysis using paper maps and colored pencils! In a GIS environment -- using the TIGER Extract products -- the computer does the "coloring" using the numeric codes in the TIGER product as a means to link with, display, and analyze the statistical data from the decennial, economic, and agriculture censuses. These codes let the data analyst examine the characteristics of the people who occupy the land -- their houses, their farms, their businesses, and their industrial activities. They also let the GIS display these data in the context of the governments responsible for managing an area and examine the inhabitants in conjunction with other geographically distributed data sets -- soil categories, hazardous waste sites, water quality, land use/land cover, sales data, and so forth.

6. WHAT ABOUT THE FUTURE OF THE TIGER SYSTEM?

Simultaneously with the Census Bureau's use of the TIGER System to support its 1990 census data tabulation and dissemination needs, work is underway to define a framework for the future of this bold new product; and this future looks bright. One aspect of this framework definition exercise has been to document the fact that the TIGER System and the 1990 census Address Control File are "national resources." Although the TIGER data base provides a computer-readable street map of the entire United States that the Census Bureau is using to support the geographic needs of all its statistical programs, TIGER provides only half of what each census and sample survey needs: Each also needs an address list; the decennial census needs the biggest list no matter what methodology is selected for the 2000 decennial census and that list also serves as the sampling frame for the Census Bureau's demographic sample surveys.

Unlike fine wine, neither the TIGER data base nor the 1990 census address list improves with age, however. Both become more out-of-date every minute as the bulldozers keep scraping new streets out of farm fields and woodlands, and as the construction industry builds its millions of new housing units each year. In addition, these two massive files still exist separately. As things stand, Census Bureau staff must apply every needed update twice; once to the TIGER data base and once to the address list.

A second major part of the framework definition exercise has been the preparation of a "geographic vision statement" to guide the planning process; it reads as follows:

> We will manage, in an automated format, a continually updated and increasingly accurate map base for the United States and its territories that includes the correct mailing address and geographic location of each housing unit, group quarters, farm, business, and industrial establishment. This computer file consistently will exceed the expectations of our customers, both internal and external, for current maps and correctly geocoded statistical data. It also will provide complete address lists to serve as the frame for approved statistical activities.
>
> To do this, we will seek to build partnerships with others who share our concern for having current and correct geographic and address information. Doing so will enable us to extend and improve on our earlier geographic innovations so that future products and processes will better serve the Census Bureau's respondents, customers, and partners.

6.1 A Comprehensive Address List

It appears that the nature of the Census Bureau's work will remain essentially the same in the foreseeable future. For this reason, the Census Bureau likely will continue to use address lists -- whether purchased from commercial vendors, compiled by staff in various field operations, or maintained as internal files updated through computer matches with the address lists maintained by others -- as a principle framework for its censuses and sample surveys.

To the extent that the Census Bureau obtains new or updated address lists for each of its statistical programs, the address matching function of the TIGER System will continue to play a major role. This function provides one of the most significant sources of information for updating the TIGER data base, especially in areas with structure-number/street name-style address systems. The resolution of the addresses that the TIGER System

cannot match and assign to a correct location highlights new features, feature names, and address ranges that need to be added to the TIGER data base as well as new addresses that need to be added to the Census Bureau's master address list. These additions ultimately will enhance the set of information available to do even more automated address matching and produce more effective cartographic displays.

6.2 Grid Reference Maps

A byproduct of having coordinate values that identify the location of every street intersection and end point in the TIGER data base is the capability the TIGER System offers to replace a labor intensive and error-prone methodology for preparing an often requested product; a street name index referenced to a map grid. The Census Bureau used automated processes to prepare a more generalized version of this product when it created its county-based street name-to-enumerator assignment area reference listings as part of the 1990 census map production process. While using the TIGER System to prepare the maps for the printed 1990 census reports, staff inflicted a uniform grid, identified with letters and/or numbers, over the entire United States. Having done so, the software can document every grid-cell through which a street (or other mapped feature) passes, sort the resulting street name list alphanumerically, and generate the desired listing. This product will be especially useful to the sample survey operations of the Census Bureau in which the field staff must visit only a sample of housing units. The TIGER System will provide not only the map and index products that inform the field representative of the location of the sample unit, but also will provide the potential, using GIS technology, to calculate the shortest route to get him or her to the unit.

6.3 Geographic Relationship Files

Similarly, while essentially a non-cartographic function, the preparation of various geographic reference files as extracts of the TIGER data base will remain a primary function of the TIGER System. These files document the hierarchy of and relationships between and among the geographic entities used for the tabulation of Census Bureau data. They provide the geographic entity "stubs" for the printed reports and summary tape files that flow from each census and sample survey--including those that will come after the 1990 census.

6.4 Structure of the TIGER Data Base

The success of the structure used for the TIGER data base reigns as one of the most significant conceptual advances in the entire TIGER System. It provides direct linkages among the geographic entities for which the Census Bureau tabulates data. It also provides direct linkages between the object representation of those entities and the object representation of the underlying feature network that forms the cartographic base for many maps and geographic information system displays. Significantly, it does this without the need for multiple overlays of boundary polygons typical of many automated cartographic systems.

The storage and retrieval approach used by the TIGER data base avoids all the traditional problems associated with multiple representations of a single line in many different feature (object) categories; for example, a state boundary that also is a county boundary, that also is a township boundary, that also is a city boundary, that also is a road. No matter what scale a cartographer selects for a map (graphic image), and no matter which categories of information the cartographer chooses to display, the relationships remain constant. In this way, there is never a chance that the simultaneous display of more than one category of object (for example, a boundary that follows a road) will result in slightly different alignments or different generalizations in relation to the basic underlying features.

Nevertheless, the structure of the files comprising the TIGER data base clearly warrant review based on the experience gained from the demands of 1990 census processing. Although the address matching function and the geographic reference file function of the TIGER System continue to be critical elements in planning the future applications software for the TIGER System, the records of computer hours used for each category of applications demonstrate that the most significant activity -- from a computational standpoint -- was the preparation of the cartographic products and the organization of the information in the TIGER data base to support those cartographic tasks. As a result, major attention will be given to reevaluating the original decision to avoid some redundancy in storing cartographically oriented data at the expense of computer processing costs.

6.5 Cooperation for Updates

The Census Bureau traditionally has used labor intensive map-to-map comparison techniques as the primary means to update the cartographic base features on its maps (and now in its TIGER data base). The increasing availability of automated mapping systems and GIS technology, coupled with the increasing sophistication of map makers and map users at all levels of government and in the private sector, makes the pursuit of cooperative programs and automated transfer approaches very appealing.

One of the great successes of the TIGER System to date has been the cooperative relationship that developed between the Census Bureau and the USGS. The likelihood is great that the USGS will continue to be a major cooperator and there appear to be other agencies interested as well. For example, an experiment is underway to explore cooperation with the United States Postal Service (USPS) as a means to update both the road and address information in a merged TIGER data base and nationwide address list (see Chart 4). Another experiment in North Carolina involves cooperation among multiple agencies and organizations to develop and maintain more detailed digital cartographic files through the cooperation of a state agency, local governments, the American Association of State Highway and Transportation Officials, the USGS, and the Census Bureau. A similar experimental study is underway in Texas involving the agencies establishing emergency service "911" telephone systems. The door to innovation and cooperation remains open at the Census Bureau.

6.6 Digital Products and GIS

If the TIGER System is to be judged truly useful outside the Census Bureau, similar planning will need to be going on in offices and institutions across the United States and around the world in the countries wishing to do business in the United States. The analytical power available to study the interrelationships among diverse data sets, using the GIS technology now available in the private sector, make better understanding a very real possibility.

REFERENCES

Carbaugh, L.W., and 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., and 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. (Guest Editor), *et al.* (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.

Chart 1. Geographic Entities of the 1990 and Other Recent Censuses

Level of Resolution	Type of Geographic Entity	Number of Geographic Entities with Data*[1]			
		Decennial Censuses		1987 Census	
		1980 [2]	1990 [3]	Economic [4]	Agriculture [5]
Very coarse (global/ national studies)	Nation (the United States) [1]	1	1	1	1
	Regions (of the United States) [1]	4	4	4	---
	Divisions (of the United States) [1]	9	9	9	---
	States and statistically equivalent areas	57 [2]	57 [3]	55 [4]	53 [5]
Coarse (national/ state studies)	Counties and statistically equivalent areas	3,231	3,248	3,228 [6]	3,179 [7]
	County subdivisions and places	**59,451**	**60,228**	7,287 [10-13]	---
County subdivisions	Minor civil divisions - MCDs	30,450	30,386	---	---
	Sub-MCDs	265	145	---	---
	Census county divisions - CCDs	5,512	5,581	---	---
	Unorganized territories - UTs	274	282	---	---
	Other statistically equivalent areas	41 [8]	40 [9]	---	---
▶Places	Incorporated places	19,176 [10]	19,365 [10]	6,776 [11]	---
	Consolidated cities	---	6	---	---
	Census designated places - CDPs	3,733 [10]	4,423 [10]	44 [10]	---
Other related areas	Special economic urban areas - SEUAs	---	---	433 [12]	---
	Balances of metropolitan areas	---	---	34 [13]	---
	American Indian/Alaska Native areas (AI/ANAs)	**499**	**579**	---	---
American Indian areas	American Indian reservations (no trust lands)	241	259	---	---
	American Indian entities with trust lands	37	52	---	---
	Tribal jurisdiction statistical areas - TJSAs	---	17	---	---
	Tribal designated statistical areas - TDSAs	---	19	---	---
Alaska Native areas	Alaska Native villages - ANVs	209	(See ANVSA)	---	---
	Alaska Native village statistical areas - ANVSAs	(See ANV)	217	---	---
	Alaska Native Regional Corporations - ANRCs	12	12	---	---
	Metropolitan areas (MAs)/urbanized areas (UAs)	---	---	---	---
"Metropolitan" [17]	Standard metropolitan statistical areas - SMSAs	323	---	---	---
	Standard consolidated statistical areas - SCSAs	17	---	---	---
	Metropolitan statistical areas - MSAs	---	267	265	---
	Consolidated metropolitan statistical areas - CMSAs	---	21	21	---
"Urban" [18] all places ▶of 2,500 or more plus all UAs	Primary metropolitan statistical areas - PMSAs	---	73	73	---
	Urbanized areas - UAs	373	405	---	---
	Special-purpose entities	---	---	---	---
	Congressional districts	435	435	---	---
	School districts	16,075	16,000 [E]	---	---
Medium (county/ local studies)	**Census tracts/block numbering areas (BNAs)**	**47,114**	**62,276**	---	---
	Tabulated parts	NA	145,035	---	---
	Block groups (BGs)	**258,398** [14]	**229,192**	---	---
	Tabulated parts	300,192	356,742	--	---
	Special purpose entities	---	---	---	---
	Neighborhoods	28,381	---	---	---
	Traffic analysis zones - TAZs	160,000 [E]	200,000 [E]	---	---
	Voting districts - VTDs	36,361 [15]	148,874 [15]	---	---
	ZIP Codes	37,000 [E]	40,000 [E]	31,000 [E]	31,000 [E]
Fine (local/neighborhood studies)	**Blocks**	**2,545,416** [16]	**7,017,427**	---	---

* Numbers represent conditions as of January 1 in the census year listed.

[1-18] See the facing page for a detailed explanation of each numbered item.

--- Not applicable [E] = Estimated number NA = Not available

Chart 1 (Continued). Geographic Entities of the 1990 and Other Recent Censuses

[1] Officially, "the United States" consists of the 50 states and the District of Columbia. For each census, the Census Bureau makes extensive data tabulations available for the 51 entities that comprise the United States and several statistically equivalent entities (see Notes 2-5 for the details of each recent census); the latter entities often are referred to collectively as "Puerto Rico and the Outlying Areas." All information in this table refers to the number of geographic entities within these sets of "states." The TIGER data base also includes -- on an equal basis:

- The Federated States of Micronesia and the Marshall Islands (which were part of the former Trust Territory of the Pacific Islands). The Census Bureau included these entities in case it was asked to assist in conducting a census in one or both of them.

- The Midway Islands. The Census Bureau included the Midway Islands to completely cover the area encompassed by the boundaries of the State of Hawaii.

The numbers in this table include only the geographic entities in the set of "states" discussed in Note 3. If the table were to include the three additional entities cited above, the counts of geographic entities at the time of the 1990 decennial census would be: 60 "states," 3,286 "counties," 60,420 county subdivisions and places (including 30,536 MCDs, 186 sub-MCDs, and 4,423 CDPs), 62,392 census tracts/block numbering areas (having 145,196 tabulated parts), 229,466 block groups (having 357,038 tabulated parts), and 7,020,924 blocks.

[2] In addition to the 50 states and the District of Columbia (the United States), the 1980 decennial census included American Samoa, Guam, the Northern Mariana Islands, Puerto Rico, the balance of the Trust Territory of the Pacific Islands, and the Virgin Islands of the United States.

[3] In addition to the 50 states and the District of Columbia (the United States), the 1990 decennial census includes American Samoa, Guam, the Northern Mariana Islands, Palau, Puerto Rico, and the Virgin Islands of the United States.

[4] In addition to the 50 states and the District of Columbia (the United States), the 1987 economic censuses included Guam, the Northern Mariana Islands, Puerto Rico, and the Virgin Islands of the United States.

[5] In addition to the 50 states (the United States for purposes of this data series, as there is no agriculture census taken in the District of Columbia), the 1987 Census of Agriculture included Guam, Puerto Rico, and the Virgin Islands of the United States. The Census Bureau conducted comparable agriculture censuses for American Samoa and the Northern Mariana Islands in conjunction with the 1990 decennial census of those entities.

[6] In addition to the county-level entities comprising the included states (see Note 4), the 1987 economic censuses tabulated data for the seven offshore entities listed below, treating them as the statistical equivalents of counties:

Alaska	Louisiana	Atlantic	Northern Gulf of Mexico
California	Texas	Pacific	

[7] The 1987 Census of Agriculture tabulated data for the county-level entities comprising the included states (see Note 5) with the following exceptions:

- It aggregated the data for Alaska's then 23 boroughs and census areas into 5 geographic entities that it treated as the statistical equivalents of counties.

- It did not provide separate data for independent cities and most counties that were coextensive with an incorporated place.

- It did not include a few other counties and statistically equivalent entities.

[8] The 41 entities include the 37 "census subareas" in Alaska and the 4 "quadrants" in the District of Columbia.

[9] The 40 entities include the 40 "census subareas" in Alaska. At the request of the government of the District of Columbia, the Census Bureau did not tabulate 1990 census data for the "quadrants"; they were replaced by a single MCD called "Washington."

[10] In agreement with the State of Hawaii, the Census Bureau does not recognize the city of Honolulu, which is coextensive with Honolulu County, as an incorporated place for statistical presentations purposes. Instead, the State delineates, and the Census Bureau tabulates data for, CDPs that define the separate communities within Honolulu County.

[11] The 1987 economic censuses included only those incorporated places having a population of 2,500 or more, except for three smaller primarily commercial/industrial places.

Chart 1 (Continued). Geographic Entities of the 1990 and Other Recent Censuses

[12] The 1987 economic censuses included as "places" those minor civil divisions in the six New England States, New Jersey, and Pennsylvania that contained 10,000 or more people.

[13] In the six New England States, the 1987 economic censuses aggregated the data for those portions of counties that were not included in some metropolitan area as the statistical equivalents of places.

[14] This number includes 102,235 entities that the Census Bureau called enumeration districts -- EDs -- for the 1980 decennial census. All portions of the United States are covered by block groups for the 1990 census.

[15] Includes only those eligible entities participating under the provisions of Public Law 94-171.

[16] The Census Bureau tabulated data by census block only for limited areas in the 1980 decennial census -- urbanized areas and their vicinity, other incorporated places with a population of 10,000 or more, and other entities that chose to contract with the Census Bureau for such data. (The latter category included 5 entire states: Georgia, Mississippi, New York, Rhode Island, and Virginia.)

[17] The "metropolitan" population of the United States is the sum of all people living in metropolitan areas (MSAs and CMSAs); the "non-metropolitan" population is all other people in the United States.

[18] The "urban" population of the United States is the sum of all people living in urbanized areas plus all people living in places (incorporated and CDP) of 2,500 population or more outside urbanized areas; the "rural" population is all other people in the United States.

Chart 2. Paper Map Products Available from the TIGER System

County Block Maps (Electrostatic plots) County-based maps showing American Indian/Alaska Native areas, county subdivisions (MCDs or CCDs), places (incorporated and census designated), census tracts or block numbering areas (BNAs), and 1990 census blocks.

P.L. 94-171 County Block Map (1990) (Available) -- show voting districts delineated by states.

County Block Map (1990) (Available) -- do **not** show voting districts.

Entity Based Block Maps (1990) (Late 1991) -- Four series of maps -- American Indian Area Block Map (1990); Alaska Native Area Block Map (1990); County Subdivision Block Map (1990); and Place Block Map (1990) -- designed to focus map sheet coverage on selected governmental units and statistically equivalent entities other than counties. Map content is the same as the County Block Map (1990).

County Block Map (1990) Version 2 (Fall 1992) -- Map sheets prepared when the computer processing to produce the County Block Map (1990) Meta Files results in a map sheet layout that differs from the original County Block Map (1990) map sheet layout; this version of the maps will match the Meta Files on CD-ROM.

Outline Maps (Electrostatic plots). The outline maps show American Indian/Alaska Native areas (AIANAs), states, counties, county subdivisions (MCDs/CCDs), places, the map series subject area, and selected base features at a small scale.

County Subdivision Outline Maps (Now - January 1992) -- State based.

Voting District Outline Maps (Available) -- County based; cover only those counties with VTDs in states that participated in the Census Bureau's voting district delineation program.

Census Tract/Block Numbering Area Outline Maps (Available) -- County based.

Urbanized Area Boundary Maps (Now - December 1991) -- Urbanized area based.

Outline Maps (Printed reports) The outline maps show American Indian/Alaska Native areas (AIANAs), states, counties, county subdivisions (MCDs/CCDs), places, the map series subject area, and selected base features at a small scale.

American Indian/Alaska Native Area Outline Maps (Spring - Summer 1992) -- American Indian or Alaska Native area based.

State/Metropolitan Area Outline Maps (Spring - Summer 1992) -- State based.

State/County Outline Maps (Available) -- State based.

Congressional District Outline Maps -- 103rd Congress (Fall 1992) -- State based.

County Subdivision Outline Maps (Available) -- State based.

Urbanized Area Outline Maps (Spring - Summer 1992) -- Urbanized area based.

Census Tract/Block Numbering Area Outline Maps (Fall 1992) -- County based.

Chart 3. Machine Readable Products Available from the TIGER System

County Block Map (1990) Meta Files (On CD-ROM only - Fall 1992.) Text files (ASCII) that contain both simple and complex graphic commands, such as, "pen up," "pen down," and "fill polygon," used to produce the County Block Map (1990) map sheets. The information also defines the spatial extent of each map image. Provides a resource to view the information available on the electrostatically plotted maps using a computer terminal screen.

1990 Census TIGER/Line™ files (Available -- on CD-ROM by state or groups of states and on magnetic tape by county or groups of counties within state as specified by the purchaser.) The CD-ROM version also contains the GRF-N for the state(s) that is (are) on the CD.

TIGER/SDTS™ files (On magnetic tape only - Mid-1992; Prototype for selected test counties on CD-ROM early Winter 1991-1992.) This file presents the point, line, and area information from the TIGER data base in a format that complies with the proposed FIPS Spatial Data Transfer Standard (SDTS).

TIGER/GRF-N™ (Geographic Reference File - Names) files (Available -- on magnetic tape; also provided with the 1990 Census TIGER/Line™ files on CD-ROM.) State files of geographic names and codes for 1990 census geographic entities presented in entity-type sort.

TIGER/GICS™ (On magnetic tape - Spring 1992; published form - Mid-1992.) State files of geographic names and codes for 1990 census geographic entities at the place level and higher, presented in hierarchical order. The tape version also will contain area measurement information and internal point coordinates for all included geographic entities.

TIGER/Boundary™ files (On magnetic tape - beginning in Mid-1992.) Six files that contain a digital representation of the boundaries for the entities specified in the title of each product. They also include the shorelines of major water features. These files will be available in two formats: the first with a full set of coordinates; the second with a "thinned" set of coordinates more suitable for use on micro-computers.

State/County (National file)

AIANA/County Subdivision/Place (State files)

Census Tract/BNA/BG (State files)

CD (National file)

AIANA (National file)

UA (National file)

TIGER/Census Tract Comparability™ file: (Magnetic tape - Late 1991.) Available by groups of counties, within state, as specified by the purchaser.

TIGER/Census Tract Street Index™ (Magnetic tape and printed listings - Late 1991; updated version on CD-ROM - Late 1992.) County-based files with the census tract/BNA number for named streets having address ranges.

TIGER/Map Sheet - Geography™ file (National file on magnetic tape - Winter 1991-1992.) This file identifies the County Block Map (1990) sheets that are required for coverage of AIANAs, county subdivisions, places, and census tracts/BNAs.

TIGER/Map Sheet Corner Point Coordinate™ file (National file on magnetic tape - Available.) This file provides the corner point coordinates (latitude and longitude) for each map sheet in the County Block Map (1990) series.

Chart 4. Proposed Census Bureau/U.S. Postal Service TIGER Cooperative Project

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

THE CREATION OF A RESIDENTIAL ADDRESS REGISTER AT STATISTICS CANADA

L. Swain, J.D. Drew, B. Lafrance, K. Lance[1]

ABSTRACT

The Address Register is a frame of residential addresses for medium and large urban centres covered by Geography Division's Area Master File (AMF) at Statistics Canada. For British Columbia, the Address Register was extended to include smaller urban population centres as well as some rural areas. The paper provides an historical overview of the project, its objective as a means of reducing undercoverage in the 1991 Census of Canada, its sources and product, the methodology required for its initial production, the proposed post-censal evaluation and prospects for the future.

KEY WORDS: Address Register; Census undercoverage; Geographical Information Systems (GIS).

1. OBJECTIVE

The concept of an Address Register at Statistics Canada dates back to the 1960s. Fellegi and Krotki (1967) first considered building one for the 1971 Census using administrative source files as the base. Their approach was mostly manual and yielded a very complete set of addresses with minimal undercoverage and overcoverage. In the mid-1970s (Booth 1976), the idea resurfaced in planning for the 1981 Census. This time the approach started with data capture of addresses from the previous Census and was augmented with information from Canada Post. In both cases, the generated address lists were being considered as a frame for a mail-out Census. However, costs of creation were high and would have needed offsetting reductions in other Census operations to be effective. In addition, the risks associated with changing the traditional enumeration method were considered too great. As a result, the construction of an Address Register was suspended in each case.

A renewed interest in the concept of an Address Register emerged from the International 1991 Census Planning Conference (Royce 1986, 1987) in October 1985. This interest derived from the potential for automation of Fellegi and Krotki's approach due to technological developments, such as the availability of machine readable administrative files with addresses and postal codes and the development of in-house software to parse addresses into standard components, to assign postal codes and to link postal codes to Census geography. It followed as well from the development of a statistical theory for record linkage (Fellegi and Sunter 1969) and computer systems based on this theory (Hill and Pring-Mill 1985).

As a result, a project was initiated in 1986 with the first research (Gamache-O'Leary *et al.* 1987) investigating the use of an Address Register for a mail-out Census rather than the traditional drop-off approach. It concluded that the new Census data collection approach would be less expensive only if the quality of the Address Register required minimal field updating prior to the Census. Two small pilot registers created in early 1987 put Address Register coverage at 90-95%, which was unacceptable without field updating (Drew *et al.* 1987), ruling out the use of an Address Register for a mail-out Census.

However, the two pilot registers revealed the potential for an Address Register to aid in coverage improvement when used in conjunction with the traditional drop-off methodology. This fitted well with the emergence of

---

[1] L. Swain and B. Lafrance, Social Survey Methods Division; J.D. Drew, Household Surveys Division; K. Lance, Geography Division, Statistics Canada, Ottawa, Ontario, Canada K1A 0T6.

coverage improvement as one of the top priorities for the 1991 Census. The results of the Reverse Record Check for the 1986 Census had indicated a dramatic rise in the undercoverage rate compared to previous Censuses (from 2.01% in 1981 to 3.21% in 1986 for the national total population; from 2.08% in 1981 to 3.28% in 1986 for the national urban population) (Statistics Canada 1990). It was therefore decided that *the research project should concentrate on the development of the Address Register as a coverage improvement method for the 1991 Census.*

The next section describes the two major tests conducted to develop and refine the procedures used to create the Address Register for the 1991 Census. As well, the second section outlines the joint agreement with the Province of British Columbia to extend the Address Register. The third section presents the administrative and geographic sources used in the production process and the structure and content of the Address Register booklets, the end product used by Census Representatives in the field. The fourth section describes the methodology used to exploit the sources in order to produce the Address Register booklets. In the fifth section, the proposed post-censal evaluation is discussed while the last section presents future prospects for the Address Register. A separate future report will detail an evaluation of the methodology.

2. BACKGROUND

2.1 The November 1987 Test of Coverage Improvement Methods

A substantial test of the use of the Address Register (AR) as a coverage improvement tool was conducted in November 1987 in five large Regional Office cities. It was designed to estimate both undercoverage and overcoverage of dwelling units for the traditional Census method of listing and for two experimental methods using an AR: Post-list and Pre-list. The Post-list approach had the enumerator compile the dwelling list in the usual Census manner (creating a Visitation Record) then reconcile it with a dwelling list for the Enumeration Area (EA) derived from the AR. Field follow-ups were done where necessary on any address discrepancies between lists. In the Pre-list method, the enumerator was given the AR in advance and updated it during a canvass of the EA to create the final dwelling list.

The results (van Baaren 1988) concluded that the Post-list method was the more effective in improving coverage. This approach as a simple add-on to the standard Census enumeration process was fail-safe. If for some reason we failed to produce the AR (either in whole or in part) on time for the 1991 Census, the AR reconciliation step could simply be dropped without affecting the traditional enumeration process. The test data also provided estimates of the degree of coverage improvement and costs (Royce and Drew 1988). It was estimated that 34,000 occupied dwellings and 68,000 persons would be added by the AR to the medium and large urban centres for which it would be constructed (these urban centres representing those areas for which an Area Master File exists, *i.e.*, covering about 65% of the Canadian population). This would represent an improvement in coverage of 0.26 percentage points (the national undercoverage rate in 1986 being estimated as 3.21 percent). Relative to the two previous attempts at AR construction, the costs were demonstrated to be low to the Census due to the highly automated approach and the proven benefit. As well, the risk was minimized since the traditional data collection method would still be used. Based on this cost, benefit and risk assessment, a go-ahead was given for creation of an AR for the 1991 Census.

From the November 1987 test, two concerns presented themselves. First, the ordering of the addresses in the AR booklets produced for each Enumeration Area (EA) didn't correspond to the order in the Visitation Records which made reconciliation a tedious and time-consuming task. Second, the overall overcoverage at 17% still seemed too high and more effort was required to eliminate erroneously placed or duplicate records. Both these problems were addressed by improving the methods for matching the AR to Census geography. Instead of linking addresses merely to EAs as had been done for the November test, procedures were developed to match the AR to the Area Master File (AMF) (Statistics Canada 1988) blockfaces. An algorithm was developed to sort addresses by block and within block in the same order they would be encountered in walking around the block.

2.2 The September 1989 Test to Refine Procedures

Another substantial test was conducted in September 1989 involving four cities of various sizes: Moncton, Laval, Brampton and Calgary. Each was chosen because of unique difficulties that could arise based on the November 1987 test. The results (Dick 1990) showed a significant decrease in coverage from 84% in the 1987 test to 73%, a discouraging outcome. On the other hand, this test revealed a considerable reduction in overcoverage down from 17% to 8%. Importantly, despite the reduced coverage of the AR, its performance as a coverage improvement tool for the Census was still viable. On analysis, the new geocoding operation was found to be problematic, both in terms of its high costs since it involved a great deal of clerical intervention and in terms of its quality. The geocoding steps were therefore revamped for production, a key aspect of which was the adoption of CANLINK record linkage software (Statistics Canada 1989b) to improve quality and reduce costs of the AR/AMF linkage.

2.3 Joint Agreement with the Province of British Columbia

The Ministry of Finance and Corporate Relations in British Columbia was concerned about the high rate of undercoverage in their province in the 1986 Census (4.49% in 1986, up from 3.16% in 1981, for the provincial total population) (Statistics Canada 1990). Statistics Canada entered into a joint agreement with the Planning and Statistics Division (the provincial statistical agency) of the Ministry on a project to help reduce undercoverage in British Columbia in the 1991 Census. The contract covered two important areas: first, working with Geography Division to expand Area Master File (AMF) coverage in British Columbia to include smaller urban areas, thereby increasing the population covered from 62% to 88% and second, building the AR for these urban areas (Stewart 1991).

The project was a co-operative effort with responsibilities and work in both AMF extension and AR creation shared between the agencies. For example, in AR creation, Statistics Canada software for standardization, merging and unduplication of administrative records was provided to the Planning and Statistics Division, which was responsible for acquiring administrative files and for production of the AR up to (but not including) the geocoding step. It carried out these steps for both the existing and extended AMF areas.

3. SOURCES AND PRODUCT

Production started in April 1990 and ended with the final Address Register (AR) booklet stapled in mid-May 1991. We had compiled 22,756 booklets containing 6.6 million addresses for the Census data collection process.

3.1 Administrative Sources

In the September 1989 test, it was concluded that wherever possible the following four administrative sources ought to be used as sources of addresses in creating the AR: telephone company billing files, municipal assessment rolls, hydro company billing files and the T1 Personal Income Tax file. However, the use of all four sources was possible only in Nova Scotia, New Brunswick, and eight major urban centres in Ontario (Ottawa, Toronto, Brampton, Etobicoke, London, Mississauga, Hamilton and Windsor). Because of the multiplicity of files, the cost of files and refusals, only three sources were used for Newfoundland, Québec, Manitoba, Alberta (telephone, hydro and tax files) and for Regina and the rest of Ontario (telephone, assessment and tax files). For Saskatoon, only telephone and tax files were available. The primary source files used by the British Columbia government were those of telephone and hydro although motor vehicles, cable and Elections files were also used (Stewart 1991).

3.2 Geography Sources

In building the AR we made extensive use of Geography Division products.

i. The Area Master File (AMF) (Statistics Canada 1988) is a digitized feature (streets, railroads, rivers, *etc.*) network for medium and large urban areas, generally with populations of 50,000 or more. Our

interest was in the street features which contained street name and civic number ranges which could be used to locate individual addresses onto a blockface, our primary linkage.

ii. The Computer Assisted Mapping System (CAM) orders blockfaces into blocks and blocks into a Census Enumeration Area (EA). CAM was used for the sequencing of addresses in the AR booklets. The EA maps produced by CAM were used by the Census Representatives for the 1991 Census. For the AR, the maps for all AMF areas were used in the second clerical operation.

iii. The 1990 Postal Code Conversion File (PCCF) (Statistics Canada 1991) is a national file of all postal codes, each of which is linked to a 1986 Census EA or a series of 1986 EAs. This input was used for secondary linkage of addresses at the EA level.

iv. The 1986/1991 EA Correspondence File relates the 1986 EA geography to the 1991 geography. This file was used for the secondary linkage at the EA level and the second clerical operation.

3.3 Address Register Booklets

The end product consisted of a set of booklets of residential addresses, one for each Enumeration Area, covering urban areas of Canada for which an Area Master File existed. Figure 1 contains a sample page from an AR booklet (reduced in size).

Figure 1: Sample Page from an AR Booklet (reduced in size)

ADDRESS REGISTER Protected PROVINCE 35 EA - SD 261 Page 21 of-de 22
REGISTRE DES ADRESSES Protégé FED - CÉF 038 VN - NV 0

Block No. No d'ilot	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	MAIN ST								1044566	5551111
4	19	MAIN ST								1044564	5561234
4	15	MAIN ST								1044562	5552321
4	11	MAIN ST								1044559	
4	9	MAIN ST								1044583	7475739
4	7	MAIN ST								1044581	5552222
5	30	CENTRE RD								1019615	5561029
5	34	CENTRE RD								1019617	
5	34	CENTRE RD	BT							1019618	5564261
5	60	CENTRE RD								1019627	
5	64	CENTRE RD								1019629	7478765
5	68	CENTRE RD								1019634	5556942
5	72	CENTRE RD								1019636	
5	76	CENTRE RD								1019640	
5	80	CENTRE RD								1019642	7476789
5	84	CENTRE RD								1019644	5568765
5	88	CENTRE RD								1019646	5559999
5	92	CENTRE RD								1019579	7473456
5	96	CENTRE RD								1019581	7450987
5	100	CENTRE RD								1019648	
5	108	CENTRE RD								1019579	5557171
5	112	CENTRE RD								1019581	5558888
5	116	CENTRE RD								1019583	7462009
5	120	CENTRE RD								1019586	7450235
5	124	CENTRE RD								1019588	5569630

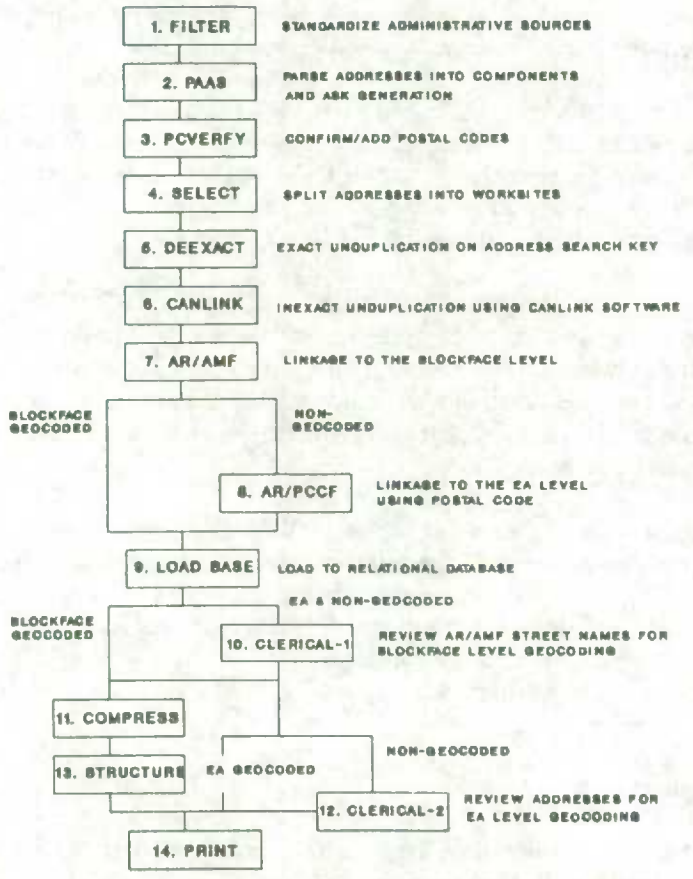
Each booklet was divided into two sections: a structured portion and an unstructured portion. The structured portion contained all the addresses tied to a blockface with all the blockfaces being sequenced into blocks within the EA. The sequencing mirrored that found on the map that the Census Representative (CR) used for listing the EA in his/her Visitation Record (VR). The unstructured portion contained the addresses that could be tied only to the EA rather than a blockface. These were sorted by odd/even civic numbers within street name. The volume of addresses was split 90%-10% between structured and unstructured.

Besides the address data, each page in an AR booklet contained a series of columns to be used in the reconciliation operation between the AR and VR. In the reconciliation, the Census Representative manually compared the Visitation Record with the AR to identify matches and non-matches. If the address was on the VR only, it was added to the AR (undercoverage in the AR). If the address was on the AR only, field resolution was usually required by the CR, with the result that the address was designated either as a new address to be enumerated for the Census by the CR (undercoverage in the Census) or as an invalid address classified by type of error (overcoverage in the AR). Addresses were denoted as invalid if they were duplicates, if they were outside the EA, or for any other reason. All valid addresses had the Census Household Number coded in the booklet by the CR. A telephone number for the address, if available, was pre-printed in the last column of the booklet for use by the CR during the Census follow-up operation.

4. METHODOLOGY

In this section, we describe the creation of the Address Register (AR) with Figure 2 providing an overview of the steps involved.

FIGURE 2: PROCESS FLOW



4.1 Overview of the Methodology

The free-format addresses contained on the source files were first standardized into their component parts (steps 1 and 2) in preparation for the use of subsequent software. Then, postal codes were confirmed or corrected (step 3) so that those areas for which the AR was to be created could be selected from among all the addresses and locations contained on the source files (step 4). Because the same addresses could be contained on more than one file or more than once on the same file, unduplication of addresses based on both exact and probabilistic matching took place (steps 5 and 6).

Next, automated linkages were made of the addresses to blockface using the Area Master File (step 7) or to Enumeration Area (EA) using the Postal Code Conversion File (step 8). After loading the addresses into a database management system (step 9), manual linkages were made of addresses to blockface (steps 10 and 11) or to EA (step 12). Addresses within each EA were then sequenced by and within blocks (step 13) before being printed and collated in booklets (step 14) for use in the Census.

4.2 Address Standardization (Steps 1, 2 and 3)

The Postal Address Analysis System (PAAS - step 2 of Figure 2) (Statistics Canada 1989c) performed two tasks: it broke up the free-format addresses from the source files into their component parts (street name, civic number, street designator, street direction, apartment number, municipality, province, postal code) and composed the address search key (ASK). ASK is an ordered concatenation of all the components of an address and is used during unduplication.

Although PAAS was an excellent product, analysis from the 1989 prototype had revealed certain shortcomings that we felt could be resolved by filtering the administrative file contents prior to using the generalized software. This FILTER step (step 1) concentrated on the following tasks: eliminating special characters with which PAAS refused to deal, repackaging address components in a manner compatible with PAAS, translating street designator short forms to acceptable ones, introducing commas between the street and municipality components of the free-format address to improve PAAS's comprehension, eliminating leading zeroes from civic numbers and numeric street names, and adding municipality and province names.

The FILTER and PAAS steps worked together in an iterative fashion. First we evaluated what anomalies needed filtering for each administrative source. If the PAAS error rate after filtering was greater than 5%, we reviewed error records looking for recurring problems to be eliminated by further filtering until an error rate of less than 5% was achieved. As any address record that failed address standardization was eliminated from the system, it was vital to have a PAAS success rate as high as possible.

The PCVERFY step (step 3) used the Automated Postal Coding System (PCODE) (Statistics Canada 1989a) package for confirmation and generation of postal codes. It was not quite as effective as the PAAS software at address analysis and could only confirm or add postal codes for 84% of the output from PAAS. It confirmed 78% of the postal codes and changed another 6%. Only .003% of the source administrative records had arrived with no existing postal code. It was crucial to have correct postal codes because these would be used for worksite selection in the subsequent step.

Two problems related to processing time arose in the PCVERFY step during production. If an address was missing a municipality/province component, the software continued to attempt to find a postal code instead of suspending further processing. As a consequence, enormous amounts of processing time could be spent trying to find postal codes. To prevent this happening, an additional task was included in the FILTER step to add municipality and province names. The second problem was such that when a street name was numeric, the processing time per address increased fourfold. Correction of this problem will require modifications to the PCODE software.

4.3 Worksite Selection (Step 4)

This step partitioned the country by postal code into manageable worksites, the sizes of which were based on the efficiency of CANLINK software for linkage of multiple large files. This led us to adopt geographic worksites

with dwelling counts in the 100,000 to 150,000 range based on the 1986 Census. Worksites were formed from an individual AMF (for a medium sized city), collections of physically adjacent AMFs (for small towns/townships), or parts of an AMF (for a large city). Where an AMF was split, it was done based on the first three characters of the postal code (Forward Sortation Area or FSA) and generally followed physical features (*e.g.*, a river) within the urban area. Geography Division's Postal Code Conversion File (PCCF) which links postal codes and detailed Census geography was used to do this partitioning in the SELECT step (step 4). Once partitioning was completed, there were 105 distinct worksites and the original 43.4 million addresses had been reduced to 20.5 million addresses, with the dropped addresses having postal codes outside the AMF areas (*i.e.*, smaller cities and rural areas).

4.4 Unduplication (Steps 5 and 6)

In order to delete addresses included more than once on the source files, an unduplication process was conducted in two stages: an exact match with DEEXACT (step 5) and a probabilistic match using CANLINK (step 6).

The DEEXACT step utilized the address search key (ASK) produced by the PAAS software and all records with an identical ASK were collapsed into a single record. With DEEXACT, the 20.5 million records from the SELECT step were reduced down to 10.1 million records. This volume reflects on the importance of performing the address standardization.

Step 6 utilized the CANLINK generalized record linkage software (Statistics Canada 1989b) that had been used in the prototype. It stratifies records into groups called "pockets" and only records within the same pocket are actually matched together. For this application, civic number was used as the pocket. The components of the address (street name, municipality name, postal code, *etc.*) were used for matching purposes and weights were assigned for agreement or disagreement of each component. The development of levels of partial agreement for street name, municipality name and LDU (the last three characters of the postal code) allowed for spelling variations and letter transpositions within the fields. The CANLINK step accounted for a further reduction to 6.7 million records. More details on the use of CANLINK in address unduplication are given in Drew *et al.* (1988), where its application in the November 1987 test is described.

4.5 AR/AMF Linkage (Step 7)

The major concern from the 1989 test was the strategy used to link addresses to their respective blockface. Because of the 11% drop in coverage from 84% to 73% compared to the 1987 test, a thorough investigation and possibly a new approach was needed. The other noticeable concern with the procedure was that automated matching accounted for only 80% of the records matched while the other 20% were picked up clerically. This would have represented a substantial manual workload during full production. In 1989, the automated process allowed only for an exact match on street name/street designator/street direction between an AR and Area Master File (AMF) record. Any inexact matching was performed clerically. In order to circumvent these two concerns, another CANLINK application was developed for the AR/AMF linkage (step 7) using probabilistic matching techniques.

The original 1989 test files for Brampton still existed, so this became the test site for developing this step. The revised approach yielded 10% more matches, which increased the coverage back up to 1987 levels. As well, the automated matching was now responsible for 97% of the matches with 3% being picked up clerically, a significant improvement on the earlier 80%-20% split. Based on these results, the CANLINK approach was adopted for Census production.

In the construction of the new matching strategy, the first area of study involved a comparison of the contents of fields that would be used for matching purposes. This revealed certain anomalies that could be corrected prior to use to improve the number of linkages. The processing modifications to existing fields covered the following areas: removal of blanks between compound street names; alignment of street directions and civic numbers; conversion of numeric street names to numbers (on the AMF); removal of special characters in street names (on the AMF); correction of spelling variations in municipalities (on the AR); and a recreation of certain PAAS translations for street names (on the AR). Several new fields were also generated: NYSIIS (New York State Identification and Intelligence System) and SOUNDEX versions of the street name, employing two

phonetic encoding packages used to eliminate the effects of common spelling errors (Statistics Canada 1989d); a duplicate street name flag (on the AMF) to identify situations where a street name was not unique; a unidirectional street flag (on the AMF) to identify streets that had only a single street direction coded; and an official street name flag (on the AR) to indicate that the street name matched an official AMF street name. The AMF records contained only street data so we appended the Census Subdivision name and a province code and then attempted to assign postal codes to blockface civic numbers. When the postal codes differed between the "from" and "to" civic numbers, we generated subblockfaces for each unique postal code.

For this application, three distinct pockets were created for each record, effectively triplicating the files. The primary pocket was the most stringent in nature and was designed to find all the good match possibilities quickly in the first pass of the files. It was composed of street name/Forward Sortation Area (FSA)/odd or even civic number flag. The second pocket was postal code/odd or even civic number flag which allowed for poorly parsed addresses to be matched on postal code. The third was the NYSIIS version of the street name/odd or even civic number flag which allowed records with spelling variations in street name and missing postal codes to be potential matches.

The function rules established for partial matches for street name, municipality name and LDU (last three characters of the postal code) were taken directly from our existing CANLINK application used for internal unduplication where they had already demonstrated their effectiveness.

An option with the CANLINK software is the ability to introduce custom code to suit the application at various key points of the linkage process. One of these, the "SELECT-PAIR" stage, permits the exclusion of unwanted linkages before records are matched, thereby reducing processing costs and increasing the quality of the links. These unwanted linkages are those with sufficiently high weights for matching but because of specific circumstances are not really valid matches and should be excluded from the comparison. In the production process, the following were among those excluded: when the civic number on the AR record fell outside the range of the AMF record; when an AMF street name was not unique and the street designators were different or non-existent on the AR record; when the AR and AMF street names were different and the AR had an official AMF street name.

There were three AMFs to which we had difficulty matching in the course of production: Red Deer, St. Thomas and Charny. The problem with all three was missing civic number data on the AMF. Knowing that these would require heavy clerical intervention, a field operation was mounted in December 1990 to update the CAM maps. CAM maps from Geography Division were sent to Regional Office staff who added the missing civic number ranges. These updated maps were subsequently forwarded to Geography Division for inclusion in the next round of updates to the AMF. For the creation of the AR, the civic number ranges for the three AMFs were used manually in the clerical operation.

Success in matching was quite similar across all provinces except for Québec. In Québec, the automatic matching to the blockface dipped by about 10-12% to 73% as it was not as effective at dealing with French addressing as it was with English addressing. Three situations were identified as causes for the drop in the automatic match rate: the use/non-use of articles within the street name (*e.g.*, Savane, de la Savane, la Savane), the use of complete personal names as street names with a high degree of spelling variability (*e.g.*, Jean-Francois Belanger, J.F. Belanger, Jean F. Belanger) and the lack of street designators. As a result, the clerical operations described below, especially the first one, were of increased importance for matching in Québec relative to the other provinces.

During the AR/AMF processing with the CANLINK software, the only problem that arose was in exceeding an internal pocket maximum on the number of records allowed. The solution was to identify the streets causing the problem from the pocket report (they were always major thoroughfares) and set up special pre-processing programs that would add the fifth digit of the postal code in calculating the pocket value for those streets to make it more discriminating. This had the effect of reducing the number of records within the pocket.

4.6 AR/PCCF Linkage (Step 8)

This step (step 8) attempted to obtain an automated link to the proper Enumeration Area (EA) for those addresses which could not be matched to the blockface using the AMF in step 7.

The principal inputs were the Postal Code Conversion File (PCCF), which gave the correspondence between postal codes and 1986 EAs, and the 1986 to 1991 EA Correspondence File. By matching the two together we could identify postal codes that were uniquely matched to a single 1991 EA, as well as postal codes matched to two or more possible 1991 EAs, requiring manual work to resolve later in step 12.

Again, Brampton became the test vehicle. The analysis of the postal code/EA matching revealed that 38% of the postal codes could be uniquely assigned to a 1991 EA. The linkage to these postal codes of the AR records unmatched to a blockface yielded a further 5% increase in total matches. Overall, the automated match rate increased to 89% (84% to the blockface and 5% to the EA), up from 64% in the September 1989 test, almost cutting in half the amount of manual intervention.

4.7 Loading the Base (Step 9)

To facilitate queries and in anticipation of future usage, ORACLE had been used in the 1989 test as the database management system and was used again for the 1991 production. The ORACLE load step (step 9) involved the transformation of the up-to-now sequential file into four separate component files, one for each of municipality, blockface, street and address.

4.8 Clerical Procedures (Steps 10, 11 and 12)

The clerical procedure for the 1989 test was a review of all unique combinations of street name/street designator/street direction from both AMF and AR records along with an AR record count for each street combination. The objective was to replace an unmatched AR street combination with the legimate AMF combination. By comparing similar street combinations and determining which ones should in fact have been identical, hitherto uncoded AR records could be matched manually to a particular blockface. This procedure had worked well in 1989 and had proved useful in two problem situations: those where there were large discrepancies in street name spelling and those where the AR street name field contained both the street name and a street designator short form that the PAAS software had not understood in parsing the address.

We expanded the capability of this clerical procedure (step 10) to compare AR street combinations with other similar AR street combinations to handle instances where a particular street might have a number of AR spelling variations with no AMF equivalent.

Following this expansion of the first clerical procedure (Clerical-1), we added a Compress step (step 11). For each unique value of street name/street designator/street direction within a worksite, all the corresponding address records were checked for uniqueness using the civic number/apartment number as the key. Where multiple records occurred, they were collapsed with all pertinent data blended into one single record.

Step 12 dealt with the residual addresses that could not be linked to a unique EA but could be matched to two or more possible EAs via step 8. A complete set of CAM-generated maps was produced for the AR project. The Clerical-2 step consisted of examining these maps for the candidate EAs to assign these residual addresses to the proper EA wherever possible.

Overall, the ratio of automated to manual matching was 91%-9%. The automated portion comprised 87% from the AR/AMF linkage to blockface, and 4% from the AR/PCCF linkage to EA. The manual portion was split 3% matched to the blockface from the Clerical-1 operation and 6% to the EA in Clerical-2.

Although ORACLE was an appropriate vehicle for the 1989 prototype, it proved to be costly and eventually a bottleneck once in full production with the AR as just one user on a Bureau-wide database. It allowed for only 8-10% of the worksites on-line at any one time, and had to export and import sites continuously to free up space and carry on processing. A second ORACLE database was therefore set up for exclusive use of the AR team.

In fairness to ORACLE, not all the processing being done was conducive to any database management system. We were building our product and as a consequence were looking at large portions of the tables to make sweeping field changes, to eliminate duplication and to select records for printing. ORACLE did offer tremendous flexibility to change software procedures quickly and generate new ones as production unfolded.

4.9 Use of the Computer Assisted Mapping System (Step 13)

The Computer Assisted Mapping System (CAM) was a new research initiative for the 1991 Census whose development ran concurrently with AR development. The system generated all the Enumeration Area maps within AMF coverage areas. This was a major departure from the manual map generation process of the past. CAM also provided a structure to EAs that located blockfaces within blocks and sequenced the blocks within the EA (step 13). An offshoot to CAM for AR purposes was set up to sequence the dwellings on the blockface. This was necessary to organize the address lists in a manner corresponding more closely to the way the Census Representatives do their listing.

For the 1989 test, CAM was not available so the maps produced for the test were generated using the computer on an EA by EA basis, to simulate the CAM product. The structure data that sequenced blockfaces in the EA and addresses within blockfaces were generated by some developmental software (Schut and Haythornthwaite 1990). The results of the 1989 test pinpointed some areas of concern. First, the EA maps would be more useful if they provided information on surrounding EAs. Second, the structure data had some problems: coding of blockfaces to the proper EA when an EA was contained within another EA; assigning sequence numbers to blockfaces in fishhook shaped streets (courts); missing blockface data; and duplication of blockface data. These problems were subsequently addressed in the development of CAM.

CAM was fully implemented by the time of AR production. In order to remain compatible with it, the same vintage of the AMF that CAM employed was used. However, a small portion of blockfaces had no structure data assigned to them. For any EA where this percentage was greater than 5%, either CAM was re-executed for that worksite if time permitted or an alternate system, Point-in-Polygon Assignments (PIPA), that locates blockfaces within their EA was executed. Although PIPA shifted addresses from the structured portion of the AR booklet (based on blockface coding) to the unstructured portion (EA coding), at least the affected addresses were not dropped during the print selection process, which was the case when sequencing data were missing.

4.10 Printing and Booklet Production (Step 14)

Previous tests gave no indication of the volume of printing and production of booklets (step 14) for the almost 23,000 Enumeration Areas containing at this point 6.6 million addresses. Major concerns included print speed and quality, durability of booklets and compilation costs.

The prototype system had used a cut-page printer which was almost ten times slower than a continuous-page printer. As a result, the print procedure was adapted for the faster machine after verification that the print quality was fully satisfactory. Another time consumer was the bursting process. We decided to drop it and leave the listing in its fan-like form.

To enhance the longevity of the booklets in the field operation, front and back covers were designed. The front cover had a window through which the Census staff could read the province code, the Federal Electoral District number, the EA number and the Census Commissioner District number in order to facilitate organization and distribution during the Census data collection operation. The other concern related to durability was the binding type. On investigation, the glued binding was found to be too brittle and couldn't withstand the constant opening and closing that was necessary, so we opted for staples.

Several options were available for compiling the booklets: contracting out the work, assembly by the print shop or gathering and attaching them ourselves. In the end, the last option was selected for considerations of cost and timeliness. In order to provide a booklet product quality acceptable to the Survey Operations Division field staff, we designed and had built wooden stapling beds to hold three staplers with document guides. We could now guarantee staple placement along the spine of the booklet as well as margin depth. The prototype performed admirably. From the timing estimates, a total of four stapling stations were constructed to meet

production requirements. Because of the print volume involved, print and spooling schedules were established with the Main Computer Centre to smooth our impact on other Bureau processing.

5. POST-CENSAL EVALUATION

The post-censal evaluation can be broadly categorized into four study areas: field operations, data capture of AR booklets, update of the AR and determination of the AR contribution to coverage improvements.

Evaluation of field operations will focus on the effectiveness of training, how complete the reconciliation work was, and causes of errors, with a view to improving the methodology for future Censuses.

The data capture operation will yield two separate outputs. First, addresses printed in the booklets will be deleted if invalid, and if valid their Census Household Number will be captured. Second, the new addresses added by the Census Representatives will be captured. It will then be possible to calculate the AR overcoverage and undercoverage rates and the AR contribution to Census coverage. Addresses placed in the wrong EA can be investigated and traced back to the source of error. Through the Census Household Number, the number of persons added and characteristics of dwellings and persons can be studied.

From a cost perspective, the unit cost per dwelling added by the AR will be calculated, in view of the cost of creating the AR and using it in the Census.

6. FUTURE DIRECTIONS

The Address Register (AR), although initially set up as one of the procedures for reducing Census undercoverage, is a developmental project with potential impact on other programs within Statistics Canada as well as other government agencies.

The more immediate objectives for the future development of the AR, as directed by senior management, are as follows: to incorporate the addresses identified during Census enumeration; to evaluate the effectiveness of the AR in improving coverage of the 1991 Census; to document and evaluate the production activities; and to develop a longer-term plan for the AR addressing its cost-effectiveness as a household frame, the optimal updating strategy and its potential for use by external agencies.

Within these guidelines, a project plan was prepared and is presented below under six main topic areas.

6.1 Relationships between the Census and the Address Register

Besides the potential for coverage improvement, other ways in which the AR could contribute to the Census will be explored. Some preliminary thoughts in this regard include possibilities for the AR to be used as a processing control file, for telephone numbers to be used for follow-up purposes, for creation of control numbers of dwellings in an Enumeration Area, for certification of dwelling counts for processing, or for migration analysis. Consideration will be given to whether the AR should be used before or after Census Day, and to how the AR might be used for those addresses where only a higher level of geography than the EA can be ascertained.

6.2 Relationships between Geography and the Address Register

As is evident in the description of the methodology, the creation of the AR relied heavily on many of the products from Geography Division (*e.g.*, the Area Master File, the Postal Code Conversion File). Their contributions and limitations in building the AR will be reviewed. For any new products developed by Geography Division, their possible use in the AR will be investigated with a view to incorporating the AR needs directly into the new product. As well, the AR will be integrated into the Geography Division's Geographical Information System (GIS).

The AR may be able to provide update indicators to the Area Master File (AMF) or for the delineation of Enumeration Areas. The AR could be used to establish priorities especially in high-growth areas or in areas where there are poor civic number ranges in the AMF. The updating of the Postal Code Conversion File might be served by postal code/Enumeration Area or postal code/blockface combinations from the AR. After each Census, all Census households are encoded with blockface centroids. Since the bulk of AR records have already been geocoded prior to the Census, a link of the AR with the Census Household Number will reduce the amount of manual geocoding work after the Census. This last project is already in progress.

6.3 Documentation, Evaluation and Improvement of Procedures

A user guide documenting procedures and a technical guide to document programs, sample problems and solutions and quality assurance are being prepared for the work done to date.

As with any new project, much is learned during the creative process and procedures are developed as required and as time and budget permit. After the fact, there are usually efficiencies to be gained by reviewing these procedures.

For the automated procedures, projects already underway include a more efficient use of ORACLE or choice of another system, the use of desk-top computers rather than the Statistics Canada mainframe computer, standardization of the filter, enhancements to PAAS, amalgamation of sites into provincial databases, the dropping of some fields earlier in the process, consideration of other postal coding software, improvement of address place name matching and an improvement of the Area Master File linkage with French addresses.

For the manual procedures, improved handling of adjacent Enumeration Areas across boundaries of Federal Electoral Districts and of the lack of civic numbers on CAM maps are to be pursued. The editing system to correct addresses will be reviewed for possible improvement as well.

Telephone numbers were added at a later stage within the AR production. A thorough evaluation of their coverage and accuracy will be undertaken especially in view of the potential uses of telephone numbers in the Census and other Statistics Canada surveys. For the latter, initial emphasis will be placed on testing within the context of the upcoming redesign of the Labour Force Survey.

Computer systems developed for the initial production have already been cleaned up to a large extent for better efficiency of mainframe expenditures, for programs and disk and tape storage, for file manipulation, for output, libraries and file access. Better system controls will be prepared.

This AR was produced only for urban areas. Future methodological development will examine the potential for extension to rural areas.

6.4 Updating Methodology

The AR was created from among four sets of administrative files: telephone files, municipal assessment files, hydro files and the T1 tax file from Revenue Canada. As well, the AR is currently being updated to be consistent with the 1991 Census so that the Census is also a source. The relative contributions of these source files, both in volume and quality, will be investigated so that a decision on acquisition of files for updating can be made.

An integral part of the updating strategy is the development of a methodology for updating. The definition of an update will be needed along with an update system. The cost effectiveness of ongoing updating, dependent on the various needs which result from projects identified throughout these future directions, will be considered as well. Is ongoing updating cost effective when compared to updating only in time for the Census? What requirements will there be from other possible uses? Answers to these questions will lead to an updating strategy.

6.5 Other Uses of the Address Register in Statistics Canada

Besides the Census and geographical relationships presented earlier, a number of other uses are suggested within Statistics Canada. The potential use of the AR in the Labour Force Survey (LFS) will be investigated as part of the LFS Redesign Project. The possibility of using the AR in urban areas either to improve sampling under the existing area frame or as a list frame to reduce the number of stages in the sample design are two major areas highlighted for research. With telephone numbers on the AR, more telephone interviewing would be possible.

The use of the AR as a survey frame for other Statistics Canada surveys will be examined. In addition, since the AR currently uses telephone files as a primary source of information, it has these files on hand for further exploitation. The Special Surveys Program, the General Social Survey and the existing Labour Force Survey are areas which use or require telephone files.

Another potential application within Statistics Canada is as a housing database if the AR were enriched with housing data from the 1991 Census and data obtained from municipal assessment files, for example. The existence of such a database might reduce the amount of information on housing that would have to be collected in future Censuses. Data needs and availability have to be explored.

6.6 Uses of the Address Register External to Statistics Canada

The AR is being considered as one of the possible joint ventures in discussions currently taking place with Canada Post Corporation. How Canada Post might contribute to future updating of the AR (as an additional source or potentially as the sole source) and how the AR might be used by Canada Post merit further investigation. As a start, a comparison between the AR and Canada Post's database will be made. In addition, the software for address standardization used in the AR is being considered as a joint project.

As previously mentioned, the AR had a joint venture with the Government of British Columbia for the extension of the AR to smaller urban centres in order to reduce Census undercoverage. Further ventures with British Columbia or other provinces might be possible.

If the AR is to be used outside Statistics Canada, issues of confidentiality of the source files must be addressed. Some source files were provided to Statistics Canada in confidence, either contractually (*e.g.*, some files from Alberta and the joint venture with the Government of British Columbia) or legally (the T1 file from Revenue Canada).

6.7 Conclusion

The breadth and diversity of the ideas contained above in future directions demonstrate the potential of the Address Register as a geographical product with methodological applications in many areas of Statistics Canada and elsewhere.

ACKNOWLEDGEMENTS

The authors would like to thank the many persons from the following areas for their dedication and perseverance in the creation of the Address Register: Phillip Reed and the AR Production Unit, Geography Division, the Labour Force Survey Sample Control Unit, Census Methodology, Survey Operations Division, the Main Computer Centre and Household Surveys Division. The authors would also like to thank Gordon Deecker, Peter Schut, Dick Carter, Phillip Reed and Carol Sol for their helpful suggestions for this paper.

REFERENCES

Booth, J.K. (1976). A summary report of all address register studies completed to date, Statistics Canada, Report E-414E.

Dick, P. (1990). Address register - September 1989 test, Statistics Canada, Draft internal report.

Drew, J.D., Armstrong, J.B., and 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., and Deguire, Y. (1988). Methodology for construction of address registers using several administrative sources, Statistics Canada, *Proceedings of the Symposium on the Statistical Uses of Administrative Data*, 181-190.

Fellegi, I.P., and 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., and Sunter, A.B. (1969). A theory for record linkage, *Journal of the American Statistical Association*, 64, 1183-1210.

Gamache-O'Leary, V., Nieman, L., and Dibbs, R. (1987). Cost implications of mail-out of Census questionnaires using an address register, Statistics Canada, Internal report.

Hill, T., and 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 of an address register in the Canadian Census, *Proceedings of the International 1991 Census Planning Conference*, Statistics Canada, 207-215.

Royce, D., and Drew, J.D. (1988). Address register research: Current status and future plans, Statistics Canada, Internal report.

Schut, P.H., and 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.

Statistics Canada (1988). Area Master File (AMF), User guide.

Statistics Canada (1989a). Automated Postal Coding System (PCODE), User and retrieval guide.

Statistics Canada (1989b). Generalized Iterative Record Linkage System, Concepts guide.

Statistics Canada (1989c). Postal Address Analysis System (PAAS), User guide.

Statistics Canada (1989d). Record linkage software, Reference guide.

Statistics Canada (1990). User's guide to the quality of 1986 Census data: Coverage, Publication 99-135E.

Statistics Canada (1991). Postal Code Conversion File (PCCF), the January 1991 version, User guide.

Stewart, A. (1991). Joint Census undercount project: B.C. address register creation, Ministry of Finance and Corporate Relations, Government of British Columbia, Report with restricted distribution.

van Baaren, A. (1988). Report on the November 1987 address register test, Statistics Canada, Internal report.

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

CURRENT AND FUTURE APPLICATIONS OF REMOTE SENSING IN SPATIAL DATA COLLECTION

R. Ryerson and M. Manore[1]

ABSTRACT

The coverage of censuses of population can be evaluated either by demographic methods or survey-based estimates of coverage errors, or by a combination of the two. This paper reviews the developments expected in the next five to ten years in remote sensing imagery, its interpretation and integration into routine spatial and point data collection.

The factors reviewed include the availability of a variety of imagery from space and higher resolution optical imagery from both space and from aircraft. As well, the paper reviews the nature of spatial information in the context of the historical development of remote sensing. In the image processing domain, the paper reviews the use of expert systems in interpretation and the often related integration of remote sensing derived information within the Geographic Information System (GIS) environment to better interpret changes. The paper closes with a look to the potential application of remote sensing to a variety of areas of potential interest to those involved in collecting spatial statistics.

KEY WORDS: Remote sensing; Spatial data; Geographic information systems.

1. INTRODUCTION

This paper provides a concise summary of the situation with regard to remote sensing today and projects that into the future in terms of likely applications and issues of interest to those involved in the area of spatial or geographic information. Remote sensing is defined here as the collection of natural resource information using imagery derived from airborne or spaceborne sensors. The benefits of using remote sensing are many, but limitations can also be significant. It is beyond the scope of this paper to detail all of these benefits and limitations, but a summary with special emphasis on spatial information is presented in Appendix A.

2. SATELLITE AND AIRBORNE REMOTE SENSING SYSTEMS

The United States launched the first earth observing satellite (Landsat 1) for natural resource monitoring in 1972. As recently as the mid 1980s the Landsat program provided the only satellite system specifically designed for natural resource monitoring. The system, based on the 80 metre spatial resolution Multispectral Scanner (MSS), was not particularly useful for broad application outside of certain agricultural, cartographic and oceanographic applications (Ryerson and Howarth 1983; Thompson *et al.* 1982; and Alfoldi and Munday 1978). The United States National Oceanographic and Atmospheric Administration's (NOAA) weather satellite series has also been used for broad area coverage for vegetation assessment with its one kilometre resolution Advanced Very High Resolution Radiometer (AVHRR).

With the advent of the higher resolution Landsat Thematic Mapper (TM), the range of application broadened considerably. (See special issue of *Geocarto International,* 1990 (1).) In 1986 France joined with its SPOT system.

---

[1] R. Ryerson and M. Manore, Canada Centre for Remote Sensing, Energy, Mines and Resources, 1547 Merivale Road, Ottawa, Ontario, Canada K1A 0Y7.

Today Europe, India, the Soviet Union and Japan all offer imagery on a commercial or quasi-commercial basis. The characteristics of the most readily available sensors are given in Table 1.

The characteristics of the SPOT, Landsat and NOAA AVHRR systems in terms of their application in forestry, agriculture and geology are well summarized in several information notes provided by the authors' organization (Anon. 1986, 1987, 1988).

At the same time, airborne sensing systems are becoming more sophisticated, covering larger areas with higher spatial and spectral resolutions. As well as optical and infrared sensors, there are active systems involving both lasers and radars (Till 1987 and Raney *et al.* 1991). For a variety of reasons outlined below, these systems provide a particular challenge to those involved in spatial data collection and handling.

Remote sensing systems available are becoming much more specialized and much more application driven compared to the sensors available in the 1970s and early 1980s. The new sensor's spatial and spectral characteristics are being closely matched to identifiable needs within the resource management or monitoring community. In conjunction with the so-called needs assessment, there are usually extensive tests and evaluations to ensure that there exists a market for any new sensor before it is actually launched or put into use.

3. THE SPATIAL NATURE OF REMOTELY SENSED IMAGERY

We assume here that spatial information is used to explain, measure or better understand the world in which we live. One application of spatial information is the development of models of the real world. The development of such models accounted for a great deal of effort in the late 1960s. Indeed, some of the geographers involved in the discussion at that time (including this paper's senior author), found themselves entering the new field of remote sensing as a means of obtaining better spatial information about features of interest in their particular geographic sub-disciplines.

That Geographers showed an early interest in remote sensing is simply a natural extension of traditional geographic study. The paradigm for this can be seen most clearly by following Haggett's review of model construction (Haggett 1965; 19-20) that saw several levels of abstraction from the real world of any particular system. The first level of abstraction he noted for a road system, for example, would be an aerial photograph of that system. The second would be a line map of the road net, while the third would be a measure of road density. In essence, these levels of abstraction can be seen as a continuum as represented below.

Real World → Image → Spatial Information → Model → Theory → Application

The fact that these abstractions were taken from air-photo-like images which are easily understood in the geographer's visually based paradigm of the world has been very important in remote sensing's development. Geographers have long applied principles of spatial organization in the extraction of information from the images using keys and similar tools (see for example the review in Ryerson 1989). It may be argued that if the images were not easily understood by the geographer, and did not fit into the geographer's world view or model construct, others from some other discipline would be developing the applications since these others would be better equipped to understand the images and hence be better equipped to extract information.

Indeed, it may be argued that the complexities associated with radar imagery have resulted in an increasing involvement of those from other areas such as electrical engineering and physics in developing or attempting to develop applications. This is entirely understandable since these newer forms of imagery are most easily understood in terms best understood by the physicist or electrical engineer. These more complex new image forms represent an abstraction of the real world one level removed from the visually based paradigm so well understood by the traditional geographer. An unfortunate result is that those based in electrical engineering and physics may well understand the imagery and even the physical nature of what is being imaged, but not the spatial or geographic context within which the features are found. Consequently, there have been some interesting problems develop in the area of information extraction.

4. INFORMATION EXTRACTION

For some years guidelines or keys were developed and used for the interpretation of imagery such as aerial photographs. These allowed a less experienced interpreter to follow a model developed by a more expert individual. This form of information extraction has not been popular over the past fifteen years or so.

From 1972 more attention was paid to multi-spectral information extraction. This approach was based on the assumption that like objects will have like spectral characteristics in several spectral regions, and the corollary that different objects will have different spectral characteristics. This oversimplification resulted in some startling errors - such as bare rock being confused with urban areas, for example. (In one case, future projections based on such errors resulted in meaningless results and a considerable waste of money for a large mapping program in the USA's portion of the Great lakes Basin. Following on the same path, it was decided that an additional dimension could be added by overlaying imagery from different dates to take advantage of seasonal changes that might permit separation of different features. While this has proven useful in some applications, the most important element for this discussion is that there now exist a suite of approaches to overlay different data sets on remotely sensed imagery. Much of this can now be done through Geographical Information Systems.

However, as discussed in more detail elsewhere (Ryerson 1989) the fundamental problem with the multi-spectral approach (even when it used multi-temporal data) is that it only uses one of several basic interpretation elements. It ignores texture, context, pattern and other spatial information. For that reason, more effort has recently been placed on the spatial information side.

There has also been a resurgence of interest in keys in terms of expert systems and artificial intelligence. Systems are now being planned which will take into consideration the special expertise of experienced interpreters, while allowing the analyst to make appropriate use of spectral and spatial algorithms available (Ryerson 1989).

It is in the domain of spatial expert that the geographer will find a new role and new uses for remotely sensed data. In addition, the new 'breed' of geographer with expertise in the computer processing of spatial data will be in a strong position to advance the technology of information extraction from remotely sensed sources.

5. REMOTE SENSING AS A SOURCE OF SPATIAL INFORMATION

In general, remote sensing is not very useful in providing *direct* socio-economic measures (beyond swimming pools!). Remote sensing is far better suited to natural resource surveys, land cover or land use, and change over time (in forests, land use, crop practices, water area, *etc.*). For example, the case has been made for the use of remote sensing for population studies in developing countries (Ryerson and Lo 1990). No such case could be made in Canada.

While it is clear that remote sensing has considerable potential to provide useful spatial information for natural resource surveys, few spatial information needs can be met by remote sensing alone. This is particularly so in a developed country like Canada where there exists a well developed data collection infrastructure with which remote sensing must compete. For example, crop areas may be easily determined for some specialized crops in Canada, but for large area crops other methods have proven far superior to remote sensing. Where these other methods do not exist however, remote sensing may well be competitive.

Remote sensing can often provide a new or different type of information that adds to the existing information set. A number of examples can be cited of satellite imagery being used for the collection of spatial data. The CCRS library contains over 80,000 papers, over half of which are concerned with some form of spatial information extraction. For example, a specialized measure of crop condition has been developed using NOAA weather satellite data to assist in crop yield prediction over large areas at relatively low cost. This has been produced as a weekly report by Statistics Canada. Other applications in Canada include the updating of thematic information such as forest cover and urban development on topographic maps. Forest inventories across the country are updated on a regular basis with satellite imagery. Recognizing the overview capabilities of satellite imagery, they are also used to help plan field work and select sample plots. In Australia agricultural extension

workers have saved 15 per cent of their transportation costs through the use of satellite imagery (Personal Communication, K. McLoy, NSW Dept. of Agriculture, Sydney).

6. THE 1990s AND BEYOND - ISSUES AND OPPORTUNITIES

A major new problem is now beginning to emerge in remote sensing. New imaging systems which use areas of the electromagnetic spectrum with which we have little experience are more difficult to interpret. Quite simply, these new systems provide imagery which we cannot relate to the world as we commonly visualize it. We carry a certain degree of bias associated with our visual sensor systems (eyes) which operate in the blue-green-red (*i.e.*, visible) portion of the spectrum.

To understand and use aerial photography or early satellite imagery, all that was required was a general understanding of the environmental system being imaged - be it southern Ontario urbanization, a New Brunswick potato field or a British Columbia forest.

Today that is not enough. One must understand these systems and also be able to understand the complex interactions between the active sensor and the features being sensed along a whole range of variables. As well, there is increasing diversity in the types of sensors available.

To solve these problems a variety of approaches are being taken. Multidisciplinary project teams have been built that bring together a range of expertise to address how best to transform the data into information. There has been an increasing emphasis on proper preparation before new sensors are launched. After imagery is available, other innovations have come about. A major factor in the recent increase in efficiency in the use of remote sensing imagery has been the integration of remote sensing with other data sets, often through the use of a GIS.

A simple example that illustrates how a variety of factors have some together to produce a new product which contains spatial information is the image map. To some it began with the objective of producing aesthetically pleasing posters. To others, they wanted to combine the information of a map with the detail of an image. What has now resulted in a whole new set of map products which combine the aesthetic appeal of the poster maps and the cartographic integrity of a map. Now the most common image maps overlay some standard topographic information (such as the road network) on image information. With the assistance of a simple key-legend, the user may extract information from these as needed. This one product appears to meet a variety of needs and a mass market.

Accompanying the trend for refined, "value-added" products such as the image map is the establishment of groups or agencies that specialize in the pre-processing of raw remote sensing data and its conversion to useable information products. For example, for monitoring vegetation condition in western Canada the Manitoba Remote Sensing Centre operationally geocodes and processes raw NOAA-AVHRR imagery into a standardized, near cloud-free product on a weekly basis. These data are, in turn, purchased by a variety of agencies including the Agriculture Division of Statistics Canada where further processing results in a series of easily understood cartographic, statistical, and graphical products for the end users. At each stage, a level of technical expertise which is not generally available is required to produce the value-added product, thus leading to the requirement of specialized shops to handle remotely sensed data. Because of the higher technical understanding required to interpret radar imagery, this type specialization is expected to increase as more of these data become available.

7. CONCLUSION

Remote sensing has already had a significant impact on how we collect and think about spatial information. It has provided us with a view of our world that makes it easier to understand, while at the same time underlining the complex interrelationships that exist at the interface between the built and natural environment. While the technology is becoming more complex, it is also being packaged in a form that is much easier and much less costly to use than it was even ten years ago. It can be said with some confidence that the development of new applications for spatial data collection is limited more by the imagination of the potential user than by the technology.

Table 1 - Satellite Sensor Characteristics

Landsat Sensors

1. Multispectral Scanner

Swath Width 185 km Spatial Resolution 80m

Spectral Bands:
1. 0.50 - 0.60 micrometre (green)
2. 0.60 - 0.70 micrometre (red)
3. 0.70 - 0.80 micrometre (near-infrared)
4. 0.80 - 1.10 micrometre (near-infrared)

Radiometric Resolution 64 grey levels

2. Thematic Mapper

Swath Width 185 km Spatial Resolution 30m (except band 6)

Spectral Bands:
1. 0.45 - 0.52 micrometre (blue)
2. 0.52 - 0.60 micrometre (green)
3. 0.63 - 0.69 micrometre (red)
4. 0.76 - 0.90 micrometre (near infrared)
5. 1.55 - 1.75 micrometre (shortwave infrared-SWIR)
6. 10.5 - 12.5 micrometre (thermal infrared - 120m)
7. 2.08 - 2.35 micrometre (shortwave infrared-SWIR)

Radiometric Resolution 256 grey levels

SPOT Sensors

1. PLA

Swath Width 60 or 117km Spatial Resolution 10m

Spectral Bands: 0.51 - 0.73 micrometre

Radiometric Resolution 64 grey levels

2. MLA

Swath Width 60 or 117km Spatial Resolution 20m

Spectral Bands:
0.50 - 0.59 micrometre (green)
0.61 - 0.68 micrometre (red)
0.79 - 0.89 micrometre (near infrared)

Radiometric Resolution 256 grey levels

NOAA AVHRR Sensor

Swath Width	2500 km	Spatial Resolution	1.1km

Spectral Bands:	0.58	- 0.68	micrometre (red)
	0.725	- 1.10	micrometre (near infrared)
	3.55	- 3.93	micrometre (shortwave infrared)
	10.5	- 11.3	micrometre (thermal infrared)
	11.5	- 12.5	micrometre (thermal infrared)

REFERENCES

Alfoldi, T.T., and 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., and Ahmed, S. (1991). Radarsat, *IEEE Proceedings*.

Ryerson, R.A., and 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., and 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.

Appendix A

BENEFITS AND LIMITATIONS OF REMOTE SENSING

Benefits:

Unlike an interview or field visit, once an image is acquired, it can be re-interpreted to derive new information. Such is not possible with field visits or interviews.

An image contains an overview of all that was in the field of view of the sensor.

Digital data, geometrically corrected can be combined with other spatial data within a GIS.

Can be used to help stratify for special surveys.

Increased spatial detail over map sources.

Limitations:

Spatial resolution and parts of the EM spectrum sensed are constrained by the sensor system being used.

Coverage is not timely for large areas at fine resolutions.

Obtaining data for sampling costs as much as purchase of total coverage.

High cost of large area coverage.

SESSION 6

Spatial Data Quality

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

DEALING WITH ERRORS IN SOCIO-ECONOMIC DATABASES: SELECTED FINDINGS OF A NATIONAL RESEARCH INITIATIVE

U. Deichmann, M.F. Goodchild and L. Anselin[1]

ABSTRACT

Due to the growing use of geographic information systems for a wide range of research and management applications, the problem of spatial database accuracy is receiving increasing attention. We review selected results conducted under the umbrella of a research initiative of the National Center for Geographic Information and Analysis. The issue is examined from the viewpoint of a user of published socio-economic data. It is argued that the increasing trend towards large, integrated databases and the automation of spatial data manipulation raise issues of geographic product quality that are often not adequately addressed in applied analysis. An example from an integrated multiregional modeling effort demonstrates the importance of sensitivity analysis in spatial analysis.

KEY WORDS: Geographic information systems; Spatial database accuracy; Socio-economic modeling.

1. INTRODUCTION

Geographic Information Systems (GIS) are being increasingly applied by geographers, statisticians, planners and regional scientists as a tool to support a variety of forms of spatial analysis using social, economic and demographic data. The new techniques and the expansion of applications from specific research oriented studies to broad based policy analysis raise questions of accuracy in spatial database operations that have received only limited attention in conventional spatial analysis. The problem of spatial database accuracy is often narrowly described as one of the cartographic domain. From a cartographic point of view it is the positional accuracy of the topological objects stored in the database that is the major concern, *e.g.* how well the lines in the database reflect the "true" lines on the earth's surface. Spatial database accuracy is, however, an equally important issue from the viewpoint of spatial modeling or regional science. For a number of reasons spatial database accuracy is a more serious problem in digital spatial data handling than in conventional cartography. Maybe the most important of these reasons is that GIS enables the creation of very large databases that draw on a variety of often heterogeneous data sources. Apart from increased technical feasibility, it is the move towards institutions sharing spatial information that makes it necessary to look more closely at the effects of integrating large spatial data sets.

A second, related issue is that GIS operations are essentially scale free. The ability to use data from various scales easily leads to the introduction of scale dependent error. If a spatial process is present at one scale, but is studied using data sets based on different inappropriate scales, the results of the analysis may be biased. Finally, GIS allows one to automate spatial analysis operations in a very flexible manner. This means that many more steps can be performed in any given spatial analysis project, and the error propagation issue might become intractable.

Even though this list is by no means complete (for a review of additional points, see Goodchild and Gopal 1989), it shows that the use of GIS in the management and use of spatial data actually aggravates the error problems

[1] U. Deichmann, M.F. Goodchild and L. Anselin, National Center for Geographic Information and Analysis, Department of Geography, University of California, Santa Barbara, CA 93106-4060, U.S.A.

that the spatial analyst faces. It is for these reasons that a research initiative was started at the National Center for Geographic Information and Analysis (NCGIA) on the topic of spatial database accuracy. The initiative started with a specialist meeting in December 1988 in Montecito, California, in which the specific research agenda was laid out. Although the initiative was formally closed in November 1990 with a series of special sessions at the GIS/LIS conference, many of the individual research projects are still continuing. In this paper we will briefly review some of the work conducted under this initiative and we will look at accuracy issues from the perspective of socio-economic data analysis. Finally, we present an example drawn from a multiregional modeling effort. Based on these experiences, it is argued that while GIS may lead to more problems in terms of spatial database accuracy, it could also provide for a flexible framework that enables the analyst to cope with these frequently encountered problems.

2. NCGIA INITIATIVE 1: ACCURACY OF SPATIAL DATABASES

From the perspective of spatial analysis, research that has been conducted under the umbrella of NCGIA Initiative 1 on the Accuracy of Spatial Databases can be broadly classified into three areas:

- Understanding the nature of errors in spatial databases.
- Understanding the effects of errors in spatial databases.
- Developing means to reduce or manage the error.

Understanding the nature of errors first of all requires us to develop a common classification of errors which considers the special implications of various data structures (for example, raster versus vector), as well as the different types of errors, such as positional versus attribute error (see Veregin 1989). A common classification is a prerequisite for developing reliable techniques for error modeling in the various data models, whereby the error model should not only include a consistent definition of the type of error but should also include measures and statistics for summarizing the uncertainty associated with a data set (see Goodchild 1990).

As an example, consider a digital soil map stored as a raster data set. Soil boundaries are usually not very well defined since they are the result of an interpolation based on point samples. This means that considerable uncertainty is associated with each boundary. One way of explicitly acknowledging this uncertainty is to assign to each raster pixel a vector of probabilities of belonging to a particular soil class (Goodchild, Sun and Yang 1992). An analogous application in the socio-economic field would be in market area research, where a gravity type model (*e.g.* the Huff model) might predict sharp boundaries while in reality these delineations are much less precisely defined. A probabilistic approach such as the one suggested adds a distinct measure of uncertainty to the GIS product, and at the same time makes it more realistic in a world of fuzzy relationships. Related work has been conducted on vector boundaries of area-class maps, where a process of generalization would essentially require a continuous view of space, and the surface represents the probability of belonging to a particular class (Mark and Csillag 1989).

As mentioned in the introduction, sequences of GIS operations lead to the propagation of errors through the spatial analysis process. A consequence is that error free products might be contaminated by combining them with data containing errors. Arbia and Haining (1992) developed a theoretical, mathematical model of the error propagation process. A more practical approach is to develop indexes of the accuracy of spatial data sets and to use these indexes to document the quality of derived map products by tracking the lineage of the derived product (Lanter and Veregin 1990). Two important implications of this work on error tracking are the analysis of the risk associated with basing decision making on GIS products of limited accuracy, and the development of accuracy standards for agencies (Amrhein and Shut 1990).

Once the nature and effects of errors in spatial databases are better understood, the next step is to develop means and methods that reduce or manage the error inherent in GIS operations. An example of how different data structures can improve the accuracy of how spatial objects are stored in a GIS database are models of surface representation. These were developed for terrain modeling but are also used to model socio-economic phenomena such as population density surfaces, or the probability surfaces used in crime modeling. A good understanding of the relative merits of these data structures can reduce the error introduced by using the wrong model specification (Kumler and Goodchild 1991).

Issues of spatial scale and spatial aggregation are a priority of research in the field of spatial analysis (Amrhein and Flowerdew 1989; Rogerson 1990). As mentioned before, scale dependence of model results is a major source of error in GIS modeling (Fotheringham 1989). Several research projects of Initiative 1 therefore dealt with the development of methods of analysis that are independent of scale (Tobler 1989), and the assessment of the effects of spatial aggregation.

We have previously argued that the integration of data sets from heterogeneous sources is one of the reasons why the accuracy issue in digital spatial databases is more prominent than in conventional analysis. Often different data sets that are needed for an analysis, are stored in incompatible formats. For example, a point sample might be used in combination with a polygon coverage. In order to use the data in analysis, an interpolation often has to be performed. This means that for spatial analysis in GIS robust spatial interpolation routines are required to deal with incompatible spatial data configurations (Flowerdew and Green 1989; Deichmann, Goodchild and Anselin 1990).

In this section we have given a brief summary of research that has been completed or is ongoing under the NCGIA's Initiative 1. More detailed information can be found elsewhere (Goodchild and Gopal 1989; NCGIA 1990; Goodchild 1990). In the following sections we will reflect on the specific spatial accuracy problems faced in the spatial analysis of socio-economic data.

3. IMPLICATIONS FOR SPATIAL ANALYSIS USING SOCIO-ECONOMIC DATA

Every user of socio-economic data knows about the problems involved in obtaining the necessary level of consistency of the data used in spatial analysis and modeling. These data problems are well summarized by Wassily Leontief (1986, p.424):

> *"The lack of effective coordination in the general area of policy formulation and implementation is matched by the absence of a clear overall design in gathering, organizing, and presenting the facts and figures on which both public and private decision making so critically depend."*

In practical work, the following broad classification of data problems emerges: sectoral/accounting inconsistencies; temporal problems; and spatial errors.

While all three sets of problems are a nuisance in regional modeling, the development of methods to address them has seen varying success. Sectoral and accounting problems are caused in the data collection process due to lack of coordination among (often competing) agencies. The major problem is that definitions of socio-economic data frequently change from one agency to another. For example, in the United States, the definitions of industrial output used in the input-output accounts of the Bureau of Economic Analysis differ from those used by the Bureau of the Census. Both agencies are part of the Department of Commerce. Often these problems can only be solved by heuristic or intuitive approaches.

Temporal issues have, in contrast, received large attention mainly in the field of econometrics and general time-series analysis. Irregular updates of data, time lags between collection and release of the data (*e.g.* the U.S. national input-output accounts), and missing values in time series belong to this set of problems. A large volume of work exists on forecasting of socio-economic data as well as on the missing value problem in time series (see Vandaele 1983; Granger 1986). Similarly, the issue of missing values in spatial cross-sectional data sets has been addressed recently by Griffith, Bennett and Haining (1989). Yet, a streamlining of the efforts by data collecting agencies could greatly improve the timely and complete dissemination of many data sets.

Spatial accuracy issues pertain to the nature of spatial errors which can be due to either specification error or measurement error, or to a combination of both (Anselin 1989). Specification error in a spatial context refers to the use of a model that does not account for location specific phenomena in the analysis, such as spatial dependence or spatial heterogeneity (see Anselin and Griffith 1988 for a review). Of specific importance in this context is the question of spatial error autocorrelation (*e.g.* whether the errors on a spatial surface are evenly distributed or clustered). In the remaining sections of this paper, however, we will concentrate on measurement error. In its simplest form, measurement error is present if a variable is referenced at the wrong geographic

location. Most often this is due to errors during the data capture process. Obviously, positional inconsistencies will have an effect on spatial operations such as interpolation, aggregation, or buffering. Also, they will have an impact on results of tests for spatial effects in the modeling stage. Besides positional errors, errors in the attributes that are stored with the spatial objects occur frequently. These attribute errors might be due to classification errors or simply to data input errors.

A further aspect of measurement error is frequently called conceptual error (Chrisman 1989). It relates to the process of transferring real world features into spatial database objects, or in other words, to the way in which data are recorded as spatial units of observation. Socio-economic data are usually collected for mostly arbitrary administrative units which do not consider the actual distributional properties of the data. Inconsistent or inappropriate sets of spatial objects used for gathering and organizing data often hinder meaningful spatial analysis. This problem occurs where different agencies collect data for different sets of regions, or where data are collected for one spatial scale but are needed for another, more disaggregated scale. For example, data might be available only for the national level, but are needed for counties or districts. In regional economic modeling this is the case with, for example, gross regional output at the substate level, or with the published input-output accounts.

This second type of measurement error touches on some fundamental issues of spatial analysis that have received a lot of attention without being satisfactorily solved. Regional socio-economic data are usually generated by aggregating the characteristics of individual economic agents (individuals or firms) in a portion of space (Arbia 1989), or by tabulating national samples by areal units. The aggregation relies on sampling procedures whereby conclusions for the entire region are derived from a number of observations. Since the possibilities of aggregating individual observations to groups of observations are very large, one could argue that the specific conventions used in aggregation will have an impact on the results of analysis. This problem is termed the *modifiable areal unit problem*, which according to Anselin (1988) arises from the fact that data are collected for spatial units that have arbitrary and irregular boundaries. This is contrary to the general notion of homogeneous space, which implies that aggregation is only meaningful if the key characteristic is evenly distributed across space. Not only the spatial partitioning chosen for the aggregation, but also the level of aggregation has an effect on the results of spatial analysis. This is termed the *ecological fallacy problem* (see Openshaw and Taylor 1979). According to Arbia (1989), if units are aggregated by summing into larger units, then mean, covariance and correlation vary. Tobler (1989), however, argues that problems with modifiable areal units and scale dependence should not be attributed to the spatial data configuration but to the wrong method of analysis.

In practice, the analyst rarely has the chance to either repartition his or her study area, or to aggregate to a scale where artifacts in the data set are filtered out; the analyst therefore has to work with whatever data are at hand. Therefore, it is important to develop a theoretical notion of the problem that helps to assess the uncertainty of results at different levels of aggregation. This could lead to the development of spatial analysis methods that are truly independent of the spatial data configuration, or in Tobler's words to "*frame independent spatial analysis*" (Tobler 1989). In order to obtain an idea about the problems that face the analyst in socio-economic impact analysis, it is useful to consider a practical example. In the next section we therefore present the case of a multi-regional modeling effort in which the problem of modifiable areal units occurred.

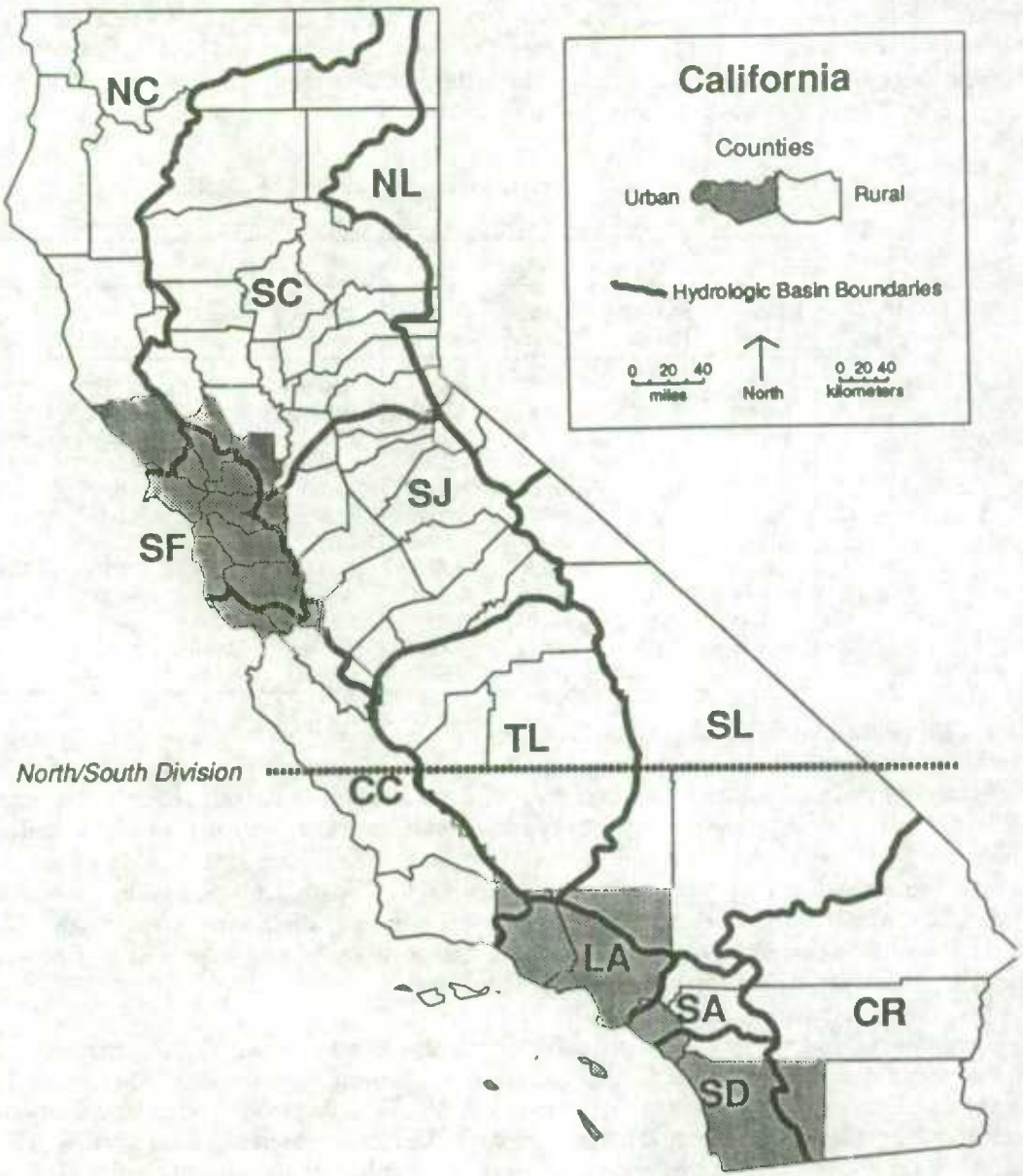
4. AN EXAMPLE: INTEGRATED REGIONAL MODELING IN CALIFORNIA

An important part of work in regional science is the development of integrated regional and multiregional models for socio-economic impact analysis (*e.g.* Isard 1986). An example of such a multiregional modeling effort related to water resources policy in California is outlined in Anselin, Rey and Deichmann (1990). The four regions that form the basis of this study are aggregates of the 58 counties in California representing an Urban and a Rural region in the Northern and Southern parts of the state respectively (Figure 1). The general objective of the model is to assess how population dynamics translate into changes in final demand from industrial sectors. These in turn lead to changes in water use in different regions and thus to a shift in the pattern of water flow among the regions. Emphasis in the project was given to the determination of the impact of different modeling strategies on model results. More specifically, the focus was on the choice of a single region versus a multi-region spatial scale, the selection of a linking versus an embedding approach to the various modules, and on the problem of the spatial aggregation of incompatible zonal data (for details see Anselin, Rey and Deichmann 1990).

In the course of data collection and model building, a number of the problems outlined above were encountered. Data by county for the state of California were needed for seven major variables: employment; payroll, wages, earnings and income; output; value added; population; transportation and commodity flows; and water supply, use and transfer. A detailed review about data availability and problems has been published elsewhere (Rey 1988). Here, we will concentrate on the spatial problems encountered.

As mentioned earlier, a major problem of working with spatially referenced published data is that different agencies publish data for different regions or zones. If the analysis is carried out on a very aggregate scale and the regions nest into each other, the problem can be solved by simple aggregation. If the level of analysis is fairly detailed, however, the more aggregate zonal data have to be disaggregated which requires an estimation procedure (*e.g.* from national to state level data). An example is the estimation of gross regional product using variants of the Kendrick-Jaycox method (*e.g.* Weber 1979), where national ratios of output to its components and regional data on the components are used to estimate regional output.

Figure 1. Economic Regions and Hydrological Study Areas



A rather more complicated variation of the problem occurs where the zonal systems are completely or partially incompatible. This might happen when data come from heterogeneous sources, or when district boundaries have changed over time. Since data for the California water model had to be integrated from a multitude of sources, this problem was encountered several times. For instance, the integration of water supply and transfer was complicated by the fact that data on these items are published for 12 hydrological basins which are largely incompatible with the county boundaries. Water use by industry, on the other hand, is available for counties. In the final activity analysis framework, a set of weights therefore had to be introduced that translated hydrological basin water supplies into economic region water demands.

Two different interpolation schemes were applied. The first is the straightforward areal interpolation (Goodchild and Lam 1980), in which two sets of polygons are superimposed and the area of overlap among each source and target region pair is computed. Arranged in a matrix and standardized, the source zone counts can then be redistributed over target zones in proportion to the areas of overlap. Clearly, the underlying assumption of this method is that the distribution of the variable is homogeneous within the source zones, and the method will mostly be applied in cases where no information on an auxiliary variable is available. In order to obtain an indication about the potential error introduced by using areal weights in a socio-economic application, a second estimate was derived, which is based on a fairly detailed map of 5,757 census tract centroids with their associated 1980 population totals. Reaggregation of these points for the economic and hydrological regions yields the share of the population of each basin that lives in each of the economic regions (see Anselin, Rey and Deichmann 1990 for details). The two weight matrices are shown in Table 1.

Table 1: Spatial Interpolation Weight Matrices

	Areal Weights				Population Weights			
	Urban South	Rural South	Urban North	Rural North	Urban South	Rural South	Urban North	Rural North
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

In a multiregional, multi-sector application neither of the two approaches is expected to be generally superior. For urban sectors (such as manufacturing or services), population provides a more meaningful approximation, while agricultural activity might be better approximated by area weights. Clearly, the actual flow patterns predicted by the linear programming/activity analysis framework are very sensitive to the different interpolation strategies (see Anselin, Rey and Deichmann 1990). In a policy context, the choice of the regional allocation scheme would have a large impact on the model results, and thus on potentially far-reaching management decisions. Given the frequency by which the problem of incompatible zonal arrangements is encountered, the development of robust interpolation schemes and sensitivity analysis of the results of various interpolation and modeling strategies should receive much more attention.

In cases where control totals for an auxiliary variable are not available, the derivation of estimates for incompatible zones has to rely on cartographic or mathematical approximation. Apart from the areal weighting method, several other methods have been proposed. A smooth distribution in which neighboring locations have similar values of density is assumed in the pycnophylactic (mass preserving) interpolation (Tobler 1979). In this approach the surface is approximated by a fine mesh of points draped over the study area. The total value of the variable for a region is then iteratively distributed over the points such that the total for each region remains constant and neighboring points have similar values. Another strategy is suggested by Flowerdew and Green

(1989) and consists of a statistical estimation procedure that takes account of additional information about the target zone that might be available.

The problems of incompatible zonal systems experienced with the California model outlined above has led to research on a generalization of the areal interpolation model. Initial results of this were presented in Deichmann, Goodchild and Anselin (1990). If the distribution in the target zones can be assumed constant and the number of target zones is smaller than the number of source zones, the target zone densities can be estimated as coefficients in a regression through the origin using the observed source zone populations as dependent variables and the areas of overlap of source and target zones as the independent variables. Simple ordinary least squares regression, however, does not guarantee that the estimated coefficients (densities) are positive and that the total estimated population for all target zones adds up to the known total. Ongoing research has therefore focused on the use of constrained regression techniques (*e.g.* Judge and Yancey 1986).

An extension of this approach is where one has access to a third set of zones - control zones - which have constant densities. If the number of control zones is less than the number of source zones, the control zone densities can be estimated as before. The target zone densities can then be derived by integrating the population density surface represented by the control zone densities over the area of each target zone. Preliminary results show that the control zones do not have to be defined very exactly to gain a considerable improvement in the estimation of target zone populations. For the California application, four control zones which were assumed to have constant densities were entered interactively as a GIS layer "by eyeballing", and could thus be regarded as expert opinion.

5. CONCLUSION: THE IMPORTANCE OF SENSITIVITY ANALYSIS

Among the problems faced in spatial modeling efforts - that of incompatible spatial arrangements of the data sources - is one of the most prominent. In the previous section approaches of dealing with the problem have been described; some have been developed under the umbrella of NCGIA Initiative 1 on the Accuracy of Spatial Databases. Given that spatial databases become larger with advances in spatial database technology and with a growing trend towards combining different databases, the need to integrate data from heterogeneous sources is becoming greater. On the other hand, GIS technology has the potential for providing a framework for the implementation of methods to deal with these data integration tasks. A large overlap exists among those issues that have been addressed during NCGIA Initiative 1 on the Accuracy of Spatial Databases and the research topics of Initiative 14, Geographical Analysis and GIS, which is due to start in Spring of 1992. Improving the analytical capabilities of GIS requires that data manipulation and exploration tools exist which allow the user to concentrate on confirmatory data analysis (see Anselin and Getis 1992; Goodchild, Haining and Wise 1992).

A major strength of GIS is that data are stored in a consistent manner which allows for great flexibility in setting up the spatial framework for analysis. A particular application of this functionality lies in cases where auxiliary information can be used to improve statistical estimations, similar to current trends of using GIS layers in a database to aid the classification of remotely sensed images (Davis and Simonett 1991). A step in the right direction is the use of classified Landsat Thematic Mapper images to aid cross area population estimation (Langford, Maguire and Unwin 1990). Thus, actual or subjective control zones stored in a GIS could be easily used in the general approach outlined in Deichmann, Goodchild and Anselin (1990), and Tobler's pycnophylactic interpolation could be modified to allow for areas in the study region for which it is known that the distribution is not smooth. In terms of actual implementation, a GIS could serve as the shell that provides an array of interpolation methods from which the analyst can choose the one that is most compatible with the nature of the data and the assumptions about its distribution.

Yet, we are still a long way from a generic data manipulation system that will give us the means of working with spatially referenced socio-economic data without having to worry about the scale and arrangement of the spatial units. It is therefore argued that a much larger emphasis in spatial analysis has to be put on sensitivity analysis that should accompany all stages of the modeling process. This involves a careful analysis of the statistical properties of the methods used, and testing a number of possible spatial arrangements and a number of analysis methods, such as measures of cross-correlation or interpolation procedures. The goal should be to provide the decision maker, who is the user of the results of spatial analysis, with some form of confidence limits, which give

a clear indication of the uncertainty associated with the analysis product. Clearly, the provision of such measures has been hindered so far by the large technical expertise and computational effort required in spatial modeling. The developments in GIS technology and the drive towards truly integrated spatial database and analysis systems, however, should allow us to obtain the flexibility necessary for robust forms of spatial analysis.

ACKNOWLEDGEMENT

This paper represents work that is related to research being carried out at the National Center for Geographic Information and Analysis/NCGIA. Support by the U.S. National Science Foundation (Grant SES-88-10917) is gratefully acknowledged.

REFERENCES

Amrhein, C.G., and Flowerdew, R. (1989). The effect of data aggregation on a Poisson model of Canadian migration, in *The Accuracy of Spatial Databases,* M. Goodchild and S. Gopal, Eds., London: Taylor and Francis, 229-238.

Amrhein, C.G., and Schut, P. (1990). Data quality standards and geographic information systems, *Proceedings, GIS for the 90s*, Ottawa: Canadian Institute of Surveying and Mapping, 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., and Getis, A. (1992). Spatial statistical analysis and geographic information systems, *Annals of Regional Science*, 26 (in press).

Anselin, L., and 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., and Deichmann, U. (1990). The implementation of integrated models in a multi-regional system, in *New Directions in Regional Analysis: Multi-Regional Approaches*, L. Anselin and M. Madden, Eds., London: Belhaven, 146-170.

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

Arbia, G., and Haining, R.P. (1992). Error propagation through map operations, *Technometrics*, 34 (in press).

Chrisman, N.R. (1989). Modeling error in overlaid categorical maps, in *The Accuracy of Spatial Databases*, M. Goodchild and S. Gopal, Eds., London: Taylor and Francis, 21-34.

Davis, F.W., and Simonett, D.S. (1991). GIS and remote sensing, in *Geographical Information Systems: Principles and Applications*, D.J. Maguire, M.F. Goodchild, and D.W. Rhind, Eds., New York: Wiley and Sons, 191-214.

Deichmann, U., Goodchild, M.F., and 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., and Green, M. (1989). Statistical methods for inference between incompatible zonal systems, in *The Accuracy of Spatial Databases*, M. Goodchild and S. Gopal, Eds., London: Taylor and Francis, 239-248.

Fotheringham, A.S. (1989). Scale-independent spatial analysis, in *The Accuracy of Spatial Databases*, M. Goodchild and S. Gopal, Eds., London: Taylor and Francis, 221-228.

Goodchild, M.F. (1990). Modeling error in spatial databases, *Proceedings, GIS/LIS 90,* Anaheim, 1, 154-162.

Goodchild, M.F., and Gopal, S. (Eds.) (1989). *The Accuracy of Spatial Databases*, London: Taylor and Francis.

Goodchild, M.F., and Lam, N.S. (1980). Areal interpolation: A variant of the traditional spatial problem, *Geoprocessing*, 1, 297-312.

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

Goodchild, M.F., Haining, R., and Wise, S. (1992). Integrating GIS and spatial data analysis, *International Journal of Geographical Information Systems*, 6 (in press).

Granger, C.W.J. (1986). *Forecasting Economic Time Series*, Boston: Academic Press.

Griffith, D.A., Bennett, R.J., and 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., and Yancey, T.A. (1986). *Improved Methods of Inference in Econometrics*, Amsterdam: North Holland.

Kumler, M.P., and 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., and Unwin, D.J. (1990). Cross area population estimation using remote sensing and GIS, *Proceedings, 4th International Symposium on Spatial Data Handling,* Zuerich, 1, 541-550.

Lanter, D.P., and 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*, 2nd Ed., New York: Oxford University Press.

Mark, D.M., and Csillag, F. (1989). The nature of boundaries on area-class maps, *Cartographica*, 21, 65-78.

NCGIA (1990). NCGIA 18 Month Report, Technical Paper 90-7, National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA.

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

Rey, S. (1988). Data availability and data problems for an integrated multiregional model of California water resources, Report W-700-88.2, Community and Organization Research Institute, University of California, Santa Barbara.

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, in *The Accuracy of Spatial Databases*, M. Goodchild and S. Gopal, Eds., London: Taylor and 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.

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

PRECISION OF RETAIL PRICE INDICES IN FRANCE AND OPTIMIZATION OF SAMPLES

P. Ardilly[1] and F. Guglielmetti[2]

SUMMARY

The purpose of this study was to calculate estimates of precision for indices at different geographic levels and levels of product nomenclature, on the one hand; and on the other, to optimize the sizes of agglomeration and survey samples, taking into account the information available. In this way, we hoped to be able to reduce survey costs when revising the retail price index, while maintaining the current level of quality.

KEY WORDS: Price index; Variance estimators; Optimization of samples.

1. INTRODUCTION

The price index calculated in France by INSEE is defined as an "index of consumer prices of urban households whose heads are workers or employees." The index covers all goods and services consumed by the reference households.

This coverage is represented by a stratification of the goods and services. Each stratum is called an "item" and there are about 300 items. Representative units, known as "varieties" are chosen from each item, on the basis of their share of consumption in the item, as well as previous knowledge of the representativeness of their price movements within the context of the item. There are about 1,000 varieties. The items may be grouped into sectors, which are 4 in number (Food - Clothing - Other Manufactured Goods - Services).

The sampling of price quotes is done as a two-degree stratified sample. Geographic strata are constructed by cross-classifying major areas (ZEATs) and categories of agglomerations known as *CC*s, which are defined in accordance with the number of inhabitants enumerated in the agglomeration. There are 8 ZEATs and 5 *CC*s, namely:

CC 2: Agglomerations of 2,000 to 10,000 inhabitants
CC 4: Agglomerations of 10,000 to 100,000 inhabitants
CC 6: Agglomerations of 100,000 to 200,000 inhabitants
CC 8: Agglomerations of over 200,000 inhabitants
CC 9: The Paris agglomeration.

For *CC*2s, *CC*4s, and *CC*6s, in each *CC*-ZEAT combination, sampling is similar to a simple random sampling of agglomerations. For *CC*8s and *CC*9s, the agglomerations are exhaustively sampled.

A sample of price quotes is obtained from each agglomeration. Here again, the process is similar to that used to obtain a simple random sample. Obviously, there is no sampling frame at this level, and the investigator draws the sample, under certain constraints (one of the most important of which is to respect the structure of consumption in the type of sales outlet: individual business, supermarket, hypermarket, *etc.*).

---

[1] P. Ardilly, Division Méthodes Statistiques et Sondages, INSEE, Paris, France.

[2] F. Guglielmetti, Division Prix de Détail, INSEE, Paris, France.

We can then construct elementary variety-agglomeration indices. Two types of varieties are distinguished: the so-called "homogeneous" varieties, where the prices of the component products can be added (for example, a loaf of bread); and the so-called "heterogeneous" varieties, where such an operation would not make any sense (for example, children's dolls).

If we let $p^t(j,i,v)$ represent the j^{th} price quote for variety v, in agglomeration i, at time t; and if $n(i,v)$ is the sample size, the unknown estimates of the true food indices are defined by:

$$\hat{I}(i,v) = \frac{\bar{p}^t(i,v)}{\bar{p}^{to}(i,v)},$$

where

$$\bar{p}^t(i,v) = \frac{1}{n(i,v)} \sum_{j=1}^{n(i,v)} p^t(j,i,v)$$

if the variety is homogeneous; and by:

$$\hat{I}(i,v) = \frac{1}{n(i,v)} \sum_{j=1}^{n(i,v)} \frac{p^t(j,i,v)}{p^{to}(j,i,v)}$$

if the variety is heterogeneous.

Aggregate indices are then defined at different geographic levels and for different product nomenclatures. We are essentially interested in annual indices by *CC* and for the whole of France, as they evolve from December to December, at the level of the varieties, items, sectors, and finally, product categories.

2. PRECISION OF THE INDICES

2.1 Method

The first level considered is that of variety indices by *CC*.

We use:

$$\hat{I}(cc,v) = \frac{1}{m(cc,v)} \sum_{i=1}^{m(cc,v)} \hat{I}(i,v),$$

where $m(cc,v)$ is the total number of agglomerations in the *CC* where variety v is sampled. This number is almost always much lower than the agglomeration sample size. In fact, and this is one of the complex aspects of the problem, we begin with the number of quotes to be surveyed per item (calculated on the basis of estimates of consumption derived from the Family Budget Surveys, and National Accounts data), and we distribute this number among the varieties proportionally to the share of consumption (or "weight") of each variety. Often, the weight of the variety is small enough that it is impossible to obtain quotes from each agglomeration. On the other hand, since the total number of quotes is about 160,000 each month, to be distributed among 1,000 varieties and a little over 100 agglomerations, we can easily see that, on the average, the expected number of quotes for each variety-agglomeration combination is less than 2. Under these conditions, as long as the variety is limited to a small number of sales outlets, or even unavailable in the agglomeration, no price quotes may be obtained there (product sold out, sales outlet closed, *etc.*).

The estimated precision takes the classic form:

$$\hat{V}\,[\hat{I}(cc,v)] = \underset{INTER}{\hat{V}}\,(cc,v) + \underset{INTRA}{\hat{V}}\,(cc,v),$$

where

$$\underset{INTER}{\hat{V}}\,(cc,v) = \frac{1}{m(cc,v)}\left(1 - \frac{m(cc,v)}{M(cc)}\right)\left[s^2(cc,v) - \frac{1}{m(cc,v)}\sum_{i=1}^{m(cc,v)}\frac{s^2(i,v)}{n(i,v)}\right]$$

and

$$\underset{INTRA}{\hat{V}}\,(cc,v) = \frac{1}{[m(cc,v)]^2}\cdot\sum_{i=1}^{m(cc,v)}\frac{s^2(i,v)}{n(i,v)},$$

with:

$$s^2(cc,v) = \frac{1}{m(cc,v)-1}\sum_{i=1}^{m(cc,v)}[\hat{I}(i,v) - \hat{I}(cc,v)]^2$$

and

$$s^2(i,v) = \frac{1}{[\bar{p}^{to}(i,v)]^2}\cdot\frac{1}{n(i,v)-1}\cdot\sum_{j=1}^{n(c,v)}\left(p^t(j,i,v) - \hat{I}(i,v)\,p^{to}(j,i,v)\right)^2$$

if variety v is homogeneous; and

$$s^2(i,v) = \frac{1}{n(i,v)-1}\sum_{j=1}^{n(i,v)}\left(\frac{p^t(j,i,v)}{p^{to}(j,i,v)} - \hat{I}(i,v)\right)^2$$

if variety v is heterogeneous.

M (CC) is the number of agglomerations in the CC.
In $CC8$ and $CC9$, the between variance is nil.

Given the small number of agglomerations involved, changes in strata that affect the agglomerations over time are not taken into account.

To obtain the precision at the national level, it is necessary to weight the variances of the indices by CC-variety. If we let W (CC/v) represent the "economic" weight; that is, in fact, the share of consumption attributed in the CC to a given variety, v, we get:

$$\hat{I}(v) = \sum_{cc} W(cc/v)\,.\,\hat{I}\,(cc,v),$$

namely:

$$\hat{v}\,(\hat{I}(v)) = \sum_{cc} W^2(cc/v)\,.\,\hat{V}\,[\hat{I}(cc,v)].$$

If variety v is not surveyed in the CC, the W (CC/v) weight used to calculate the index is nil. This situation is said to be abnormal (since all varieties are necessarily consumed in all CCs), in which case variance is not calculated at the variety level, but taken into account at the item level through imputation.

If we wish to obtain an estimate of precision by item, we weight the precision of the variety indices by the square of the weights of the varieties in the item. There is, however, a problem of co-variance between indices: although we can assume *a priori* that the elementary indices of two varieties of the same item in a given agglomeration

are independent (since the quotes are independent); the same is not true for the total agglomeration rate of two varieties of the same item. However, co-variance calculations indicate that, in view of the other terms, this can be numerically ignored.

Thus, it does not seem clear that an agglomeration that is more inflationary than the average for a variety of an item, would also be more inflationary for the other varieties of the same item.

Under these conditions, if the item includes varieties for which precision cannot be calculated at the variety level, we decide either to impute the average precision of the varieties of the item for which calculation was possible, or to limit ourselves to the latter, and re-weight them so that the sum of their weights is always one. This gives two estimates of precision, based on two different approaches. By weighting, we then proceed to the sectors and overall precision.

2.2 Principal Results

The calculations show that:

- The precision of the variety indices is poor, even very poor. This is explained by the large number of varieties with low weights; that is, few quotes (often under 100). Varieties for which the calculated standard deviation is over 2 represent 11% of the varieties, but only 2% of the weights. The median standard deviation is 1.1.

- The precision of item indices is slightly better, but remains mediocre overall: for 50% of the item indices the standard deviation calculated was over 0.7. This result is much more important than in the case of the varieties, because item indices are published. Items for which the estimators were over 1 represent 25% of the items, but only 8% of the weighting.

- The results for 1987 to 1990 were very similar from the item level up, and the standard deviation of the estimated general annual index was close to 0.05.

The table below shows national precision by sector in 1988, excluding fresh products:

	Standard Deviation	Weight	Number of Price Quotes (x 1,000)
Food	0.061	21	43
Clothing	0.140	10	23
Other Manufactured Goods	0.070	35	45
Services	0.082	34	19
Overall	0.042	100	130

3. OPTIMIZATION OF THE AGGLOMERATION SAMPLES

This is done on the general index, taking into account all varieties and weights selected.

Since a sample of agglomerations is drawn only once, an attempt is made to minimize costs while retaining precision, by not including quotes that may change from year to year. The cost function is simply equal to the total number of agglomerations drawn, since it was impossible to attribute an approximate cost by agglomeration.

In this solution, we are interested in the number of agglomerations $m\ (CC,Z)$ to be drawn by cross-classifying major areas with CCs (the major area is represented by Z).

Thus, we solve:

$$\text{MIN} \sum_{cc,z} m\,(CC,Z)$$

under the following constraints:

$$\begin{cases} \sum_{cc,z,v} W^2\,(CC,Z,V)\ (1 - \dfrac{m\,(CC,Z)}{M\,(CC,Z)} \times \dfrac{S^2\,(CC,Z,V)}{m\,(cc,z)} = V_{ref} \\ 0 \le m\ (CC,Z) \le M\,(CC,Z). \end{cases}$$

V_{ref} is the variance calculated with the current sample and the same model. Measures of dispersion S^2 are estimated, so that the criterion includes the inter-agglomeration variance plus "part" of the intra-agglomeration variance. If the values are missing because the total number of quotes for variety v in the CC–Z combination is insufficient, we impute a dispersion equal to the average of the dispersions of the indices calculated for all the major areas with a set variety-CC combination.

If the CC–Z-Variety weight is nil, no corrections are done, which is equivalent to re-weighting the varieties for which information exists.

On this occasion, we replaced the agglomerations in their current CCs (more exactly those of the 1982 census) in order to approximate the conditions that will prevail when the new sample is actually put to use.

From an algorithmic point of view, we deal with the constraint of inequality by not taking it into account initially. If the size required is larger than the total number of agglomerations available in the CC–Z, the inequality is saturated, and we start again without taking into account the CC–Z involved. This only occurs for large agglomerations in the Midi-Mediterranean area.

The calculations made for 1981 to 1990 show great numerical stability:

Optimal Distribution of the Agglomeration Sample by Agglomeration Category While Maintaining the Precision of the Current Year's Index

Category Of Agglomeration	1981	1982	1983	1984	1985	1986	1987
2,000 - 10,000 inhabitants	14	12	14	10	14	14	11
10,000 - 1,000,000 inhabitants	30	26	30	29	29	29	33
1,000,000 - 200,000 inhabitants	13	14	13	14	15	16	13
+ 200,000 inhabitants	24	22	24	23	23	23	22
Paris	1	1	1	1	1	1	1
Total	82	75	82	77	82	83	80

Details for 1989 in each *CC-Z* are shown in the next two tables, the first shows the current sample, and the second the optimal sample.

Current sample (1989):

		MAJOR AREAS								
		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

Optimal sample:

		MAJOR AREAS								
		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

Optimal sample: average of the optimal index distributions obtained from 1981 to 1987

MAJOR AREAS

1 Paris area
2 Paris basin
3 North
4 North-East
5 West
7 South-West
8 Centre-East
9 Midi-Mediterranean

CC: Category of Agglomerations

2 Agglomerations of 2,000 to 10,000 inhabitants
4 Agglomerations of 10,000 to 100,000 inhabitants
6 Agglomerations of 100,000 to 200,000 inhabitants
8 Agglomerations of over 200,000 inhabitants
9 Paris agglomeration

4. OPTIMIZATION OF SAMPLES OF PRICE QUOTES

Optimization is done on a fixed sample of agglomerations. Initially, an attempt is made to minimize intra-agglomeration variance while controlling costs. The main variables are the number of quotes $n(CC,v)$ by category of agglomeration CC for variety v. A chart of unit costs for collection $c(CC,v)$, as well as an overall budget C, are introduced.

Thus, we solve:

$$\text{Min} \sum_{cc,v} W^2(CC,v) \frac{s^2(CC,v)}{n(CC,v)}$$

under the following constraints: (4.1)

$$\sum n(CC,v)\, c(CC,v) = C.$$

When the estimate of dispersion $s^2(CC,v)$ is unknown, but the $W(CC,v)$ weight is not nil, the average of the dispersions of the different CCs is imputed to a given variety v. If the weight is nil, only two quotes are carried out.

Moreover, there are particular treatments for certain varieties. Unfortunately, this overall perspective does not allow for monitoring precision by item.

To remedy this situation, the cost may then be minimized by imposing certain limitations on the standard deviation of the items.

We then draw up a list P of the items with standard deviation greater than a given threshold $\sigma(p)$, and then solve:

$$\text{Min} \sum_{p \in P} \sum_{\substack{cc \\ v \in p}} n(cc,v) \cdot c(cc,v)$$

under the following constraints: (4.2)

$$\forall p \in P: \sum_{\substack{cc \\ v \in p}} \frac{w(cc,v)}{w(p)} \cdot \frac{s^2(cc,v)}{n(cc,v)} \leq [\sigma(p)]^2.$$

The cost related to the items in list P is determined (namely the optimal value of the function), and subtracted from the overall budget C. Finally, a program of type (4.1) is solved for the other items; that is, those that are not part of list P. This results in a deterioration of the standard deviations of the "good" items, but this is the choice that has been made. To be rigorously exact, we should repeat the process, since it is possible that, as the result of using (4.2) and then (4.1), a few items may be attributed a standard deviation that is above the imposed threshold $\sigma(p)$.

To avoid fluctuations in size at a level as detailed as the CC-Variety combination, we smooth the results by calculating the average of the optimal $n(CC,v)$ over the last three years.

With the following cost chart:

Sector	*CC*				
	2	4	6	8	9
Food	3	2	2	1	1
Clothing	9	5	5	3	3
Manufactured Goods	5	3	3	2	2
Services	7	5	4	3	3

For a program of type (4.1), we get

	Category of Agglomeration						Standard Deviation
	2	4	6	8	9	T	
CURRENT DISTRIBUTION							
Food	5	10	5	10	13	43	0.066
Clothing	2	6	3	5	7	23	0.149
Other manufactured goods	3	15	7	11	11	43	0.065
Services	2	5	3	4	5	19	0.100
Overall	12	32	18	30	36	128	0.044
OPTIMAL DISTRIBUTION							
Food	4	10	6	12	18	50	0.045
Clothing	1	5	2	6	7	21	0.104
Other manufactured goods	2	9	5	11	17	44	0.038
Services	2	4	3	6	11	26	0.047
Overall	9	28	16	16	35	53	0.025

The optimal distribution ensures a cost that is (almost) the same as the current cost of collection. Again, the standard deviations are intra-variety standard deviations only.

Using the (4.2) and then the (4.1) programs, while maintaining a constant total number of quotes, we get the following results for 4 values of $\sigma(p)$:

	Lack of Constraint (recall)	Intra (item) Standard Deviation Under:			
		0.9	0.8	0.7	0.6
STANDARD DEVIATION BY SECTOR					
Food	0.052	0.055	0.057	0.060	0.087
Clothing	0.095	0.110	0.111	0.115	0.114
Other manufactured goods	0.039	0.047	0.046	0.051	0.062
Services	0.043	0.055	0.054	0.055	0.077
Overall	0.025	0.030	0.030	0.040	
STRUCTURE OF THE SAMPLE (thousands of quotes)					
Food	35	36	35	30	0.087
Clothing	24	24	26	37	0.114
Other manufactured goods	40	40	41	45	0.062
Services	29	28	26	16	0.077
Overall	128	128	128	128	

Finally, regardless of the option, we attempt to distribute the $n(CC,v)$ quotes among a maximum number of agglomerations in the CC, taking into account existing conditions as much as possible.

5. CONCLUSIONS

Considering the sizes of the samples involved, and the hypotheses formulated throughout the calculation process, it would seem that we could be quite confident of the precision of the sector indices, and the overall index, as well as the optimal number of quotes per sector-agglomeration category.

Finally, within the framework of the revision of consumer price indices, these results have two aspects: on the one hand, a new understanding of the sensitive point of index quality; and on the other hand, an operational aspect. In fact, optimization of the samples, combined with estimates of precision, makes it possible not only to update the numbers annually, but also to revise the nomenclature of the items and the choice of varieties on a statistical basis.

However, the precision of the items can only be taken into account for optimization if data are available for several years.

The only operation that could be carried out on data previous to 1987 was a bootstrap estimate on estimated agglomeration indices for all products between 1981 and 1987. This estimate was too rough to produce results comparable to those obtained with the analytical method, but sufficient to show that the precision of the overall index does not depend upon the level of inflation. This result, which is interesting in itself, allows us to hope for some future stability in the results of optimization.

REFERENCES

Cochran, W.G. (1977). *Sampling Techniques*, New York: Wiley.

INSEE (1987). *Pour comprendre l'indice des prix* (Second Edition).

INSEE. *Bulletins mensuels de statistique.*

U.S. Bureau of Labor Statistics. *Item-Outlet Sample Redesign for the 1987 US Consumer Price Index Revision.*

Various Studies presented at Seminars on Consumer Price Statistics, GENEVA, (June 1986).

SESSION 7

Medical Geography

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

A SURVEY AND CRITIQUE OF DISEASE ATLASES FROM AROUND THE WORLD

S.D. Walter[1], S.E. Birnie[1] and L.D. Marrett[2]

ABSTRACT

Forty nine disease atlases were surveyed to characterize their mapping methodology with respect to population covered, diseases represented, mapping techniques, and statistical methods. Little consistency was found concerning choice of data function to be mapped, minimum event frequency requirements, method of age standardization, or map colour systems. Many atlases did not include basic descriptive information, emphasized statistical significance rather than rates, and concentrated on high rather than low risk. Few included environmental data or etiologic interpretation. As a result of methodologic differences, comparisons between atlases are difficult. A set of guidelines for use in future atlases is proposed.

KEY WORDS: Disease mapping; Mortality; Morbidity; Geographic variation; Methodology.

1. INTRODUCTION

1.1 Background

There has recently been renewed interest in the geographic distribution of disease, which has led to a proliferation of disease atlases. This is probably in part due to the recognition of the potential role of descriptive epidemiology in elucidating the causes of disease, monitoring the effects of changes in exposures, and planning of health services, and in part due to the ready availability of both the data and the computer software needed to prepare the maps and analyse geographic data.

During an investigation of the spatial aggregation of cancer incidence in Ontario, a number of disease atlases were examined, and substantial variation was noted in what was being mapped and how. As a result, it was decided to try to identify all recently published disease atlases, to examine them in a systematic way, and to summarize their contents and methods. We believe that all national and international atlases published within the 15 years or so prior to closure of the survey (early 1990) were examined. Atlases published more than 15 years earlier were not always included, and were generally not actively sought. Likewise, atlases which have only recently appeared are not covered.

A summary of the results of this survey are presented herein. More detailed methods and results have recently been published (Walter and Birnie 1991).

1.2 Objectives

The objectives of the survey were threefold, namely:

1. To describe mapping techniques currently in use;

---

[1] S.D. Walter and S.E. Birnie, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada L8N 3Z5.

[2] L.D. Marrett, OCTRF Epidemiology Research Unit, Department of Preventive Medicine and Biostatistics, University of Toronto, Toronto, Ontario, Canada M5S 1A8.

2. To assess the potential for atlases to provide interpretable and useful information to epidemiologists; and

3. To assist in the development of future atlases, by proposing a set of methodologic guidelines.

2. METHODS

Atlases were identified in a variety of ways: library searches, personal contacts, and informal advertising of our interest. The data extracted from the atlases included the basic descriptive information on the regions and diseases mapped (*e.g.*, population size, land area, diseases mapped, and case load); mapping techniques and statistical methodology employed; and provision of additional data and analysis. Two of us (SW and SB) independently reviewed all atlases and resolved discrepancies through discussion and re-examination. Whenever possible, an atlas was studied in its entirety. However, it was sometimes possible to review only photocopies of selected pages.

Where a needed piece of data, such as numbers of cases, population size or land area, was not provided in the atlas directly, it was either inferred from other data present in the atlas or obtained from other sources when possible.

3. RESULTS

3.1 Basic Descriptive Data

Forty-nine atlases were included in the survey (see complete list, Appendix A). Table 1 lists all the countries or regions covered by these. North America, Europe, and other developed regions of the world such as Japan, Australia, and New Zealand, are generally well represented. Some areas are even covered by more than one atlas, most usually for different diseases or time periods. Noticeable gaps exist, however, for countries in Africa, the Middle East, Central America, South America and Asia; these are areas where there is much to be learned, but where data of the type required may not be readily available.

Table 1: Areas covered by surveyed atlases

National and International

Australia	Finland	Portugal
Austria	France	Scotland (2)[1]
Belgium	Italy	Spain
Brazil	Japan	Switzerland (3)[1]
Canada (3)[1]	Netherlands	Taiwan
China	New Zealand	United Kingdom (2)[1]
Denmark	Nordic Countries	United States (4)[1]
England and Wales (3)[1]	Norway	Uruguay
European Economic Community	Poland (2)[1]	West Germany (2)[1]

Intra-national

Firenze (Italy)	Navarra (Spain)	Slovakia (Czechoslovakia)
Friulia-Venezia-Giulia (Italy)	Quebec (Canada)	Uppsala (Sweden)
Isère (France)	Saskatchewan (Canada)	Victoria (Australia)

[1] Numbers in brackets indicate numbers of atlases included in that area

From Table 2 it is evident that 38 of the 49 atlases surveyed concerned cancer, while a small number included all causes of mortality and an even smaller number covered selected causes other than cancer. Also, most were based on disease mortality, rather than incidence.

Table 2: Data mapped in surveyed atlases

Diseases mapped	No. of atlases
Cancer	38
All causes	8
Other	3
Disease measure mapped	
Incidence	13
Mortality	35
Both	1

There was a considerable range in the number of disease groups mapped (Table 3), even for cancer atlases (4 to 36, not shown). There was also considerable variation in the size of populations covered by the maps (from 500,000 to 825 million), and in the numbers of cases of disease occurring in the mapped areas (from 7,000 to 6.4 million).

Table 3: Characteristics of populations and diseases mapped in surveyed atlases

	Minimum	Median	Maximum
No. of disease groups[1]	3	18	59
Population (millions)	0.5	16	825
No. of cases (thousands)	7	235	6,406

[1] Excludes "total' or "other" causes. Each cause group is counted only once even if two maps (*e.g.* sex-specific maps)

3.2 Provision of Basic Data and Characteristics of Regions Mapped

From Table 4, it is evident that several fairly basic descriptive variables were not always provided. For example, only 51% of atlases indicated population size and only 29% land area, while between 70% and 80% gave numbers of disease cases and numbers of cases mapped.

Table 4: Frequency of specification of selected variables in surveyed atlases

Variable	Specified	Not Specified
Population size	25 (51%)	24 (49%)
No. of disease cases in population	39 (80%)	10 (20%)
No. of cases mapped	35 (71%)	14 (29%)
Land area	14 (29%)	35 (71%)

Table 5 describes the geographic divisions employed in the maps. Some atlases mapped very few regions, while some had a large number. The atlas of China (A8, Appendix A), for example, is at the upper end with 2,392 regions, while many of the intra-national atlases (A41-A49, Appendix A) were at the lower end. Perhaps surprisingly, there is more than a 2,000-fold difference between the largest and smallest average regional population (from 23 million in the Brazilian atlas to 10,000 in Finland (A4 and A14, Appendix A, respectively)), and even greater variation in the average area per region. For the smallest disease group mapped, there was considerable variation in the average number per region. The variation in definitions of regions and disease groups employed in the atlases results in very different coefficients of variation for the regional values shown in the maps.

Table 5: Characterization of mapping regions employed in surveyed atlases

Variable	Minimum	Median	Maximum
No. of regions mapped	4	75	3,072
Population per region (thousands)	10	240	23,800
Area per region (thousands of square kilometres)	0.1	3	1,700
No. of cases per region	235	2,804	136,307
No. of cases per region for least frequent event mapped	0.1	4.0	56

The issues of how many regions to map and how large they should be are obviously related and are non-trivial: regions must be large enough to produce stable rate estimates, yet small enough to provide meaningful information and, where possible, to represent a relatively homogeneous population. While one might expect there to be some recommended optimal region size, it does not exist. Perhaps the wide variation observed is related to issues beyond the control of the mapper. For example, regions may be determined politically; numerator or denominator data may be coded or available only in a certain way; or data quality may dictate the level of aggregation. In fact, the observed tendency in the surveyed atlases was for areas with larger populations to have more regions (not surprising), but also for their regions to have **smaller** land areas and **smaller** populations, on average, than the regions employed for atlases covering smaller populations.

Few atlases reported on the criteria employed to select disease groups or regions for mapping. Only one atlas had acknowledged criteria for both, 16 atlases specified criteria for one or the other, and 45 stated no criteria. When specified, the criterion for mapping a disease group was usually based on its frequency, whereas mapping a region was determined on the basis of its population.

3.3 Mapping and Statistical Methods

There was also considerable variation in the choice of data function or functions to be displayed in the maps (Table 6). (Atlases may be counted more than once in this table if they mapped different data functions, depending, for example, on the rarity of a disease.) The majority of atlases (65%) chose to map relative rates, either alone or in combination with other functions, while a sizable number (27%) showed absolute rates. A few atlases (13%) displayed only significance levels or relative frequencies.

Table 6: Data function plotted displayed in surveyed atlases

Data Function	No. of atlases[1]
Absolute rate (R)	12
Relative rates (RR)	20
Significance (S)	6
Relative frequencies (F)	2
R and S combined in same map	2
RR and S combined in same map	15
R and RR combined in same map	1
RR and S in separate maps	2
R and RR in separate maps	1
R,RR and F in separate maps	1
	Total: 62

[1] Atlases may be counted more than once

Most atlases displayed their data through use of chloropleth mapping, i.e., regional shading or colouring, which requires reduction of the data function to a single dimension in the colour/shade scale. This was most often accomplished by dividing the data function into either percentiles or categories of relative risk on a multiplicative scale. The percentile approach ensures predetermined numbers of regions in each of the percentile groups, as used, for example, in the cancer atlas of Scotland (A26, Appendix A), where the lowest 5%, next 10%, 20%, 30%, 20%, 10%, and highest 5% of function values are mapped. The multiplicative risk scale is able to equally emphasize positive and negative departures from the "average", but may have few regions represented in some groups and many in others. The latter scaling is employed, for example, in the cancer atlas of Finland (A14, Appendix A). It is of interest that when both significance and rates are illustrated, many atlases give precedence to statistical significance over rate values, and to risk elevations over risk deficits. Some atlases further employed smoothing techniques to avoid problems associated with small numbers, as was done for Finland (A14, Appendix A), but regional entities then become difficult to identify.

Just over one-half of the atlases surveyed employed colour. While there is colour theory that suggests some natural orderings of colour, few atlases followed them, resulting in a wide variety of colour schemes. Some common patterns of use were evident, however (Table 7). For example, red or orange was most often used to indicate high risk, while green was most often used for low risk and white for neutral. However, it is of interest that the same colours were used, in different atlases, to represent different categories of risk. For example, blue was used to designate the low risk category in 7 atlases, high risk in 4, and the neutral category in 2 atlases.

Table 7: Colours used for mapping in surveyed atlas[1]

Colour	Frequency used to indicate:		
	Highest risk category	Neutral category	Low risk category
Red	26	0	0
Orange	6	1	1
Brown	2	2	1
Blue	4	2	7
Green	1	4	21
Yellow	0	8	6
White	0	10	5

[1] Only includes atlases employing colour. Atlases can be counted more than once if a variety of colour schemes are used.

The majority of atlases used indirect standardization for age with an internal standard population (Table 8). In general, those employing indirect standardization had smaller total populations, although the average population and number of cases per region were actually larger than in the areas employing direct standardization. This is rather surprising, given that the major indication for employing indirect rather than direct standardization is small regional frequencies.

Table 8: Method of age standardization and standard population employed in surveyed atlases[1]

Type of standard population	Method of age standardization	
	Direct	Indirect
Internal	10	24
External	9	3
Not specified	1	0

[1] Some atlases are not included because the method of standardization was not specified.

3.4 Provision of Additional Data and Analysis

Most atlases provided regional frequencies and/or rates as either printed tables or microfiche (Table 9). A number of atlases also estimated the number of persons or regions according to risk level. Only the minority of atlases provided regional data on variables which may be related to health, such as population density, climatic factors, and socio-economic status. While the most commonly stated purpose for an atlas was related to identification of etiologic agents (usually through hypothesis generation), data were rarely provided to make this task easy - or even possible - for the reader, without recourse to supplementary data from other sources. Fewer than 50% of the atlases surveyed provided any interpretion of the maps, and only 3 included formal spatial analysis.

Table 9: Supplementary tables or maps included in surveyed atlases

Tables	No. of atlases (% out of 49)
Case frequencies	39 (80%)
Regional rates	45 (92%)
No. of regions or persons by risk level	16 (33%)
Maps	
Population density	12 (24%)
Climate	5 (10%)
Physical features	5 (10%)
Socio-economic status	4 (8%)
Ethnicity	3 (6%)
Geology	3 (6%)
Other	5 (10%)

4. CONCLUSIONS AND RECOMMENDATIONS

It is generally agreed that atlases have the potential to provide interpretable and useful information to epidemiologists. This is supported by the fact that in-depth studies of etiologic hypotheses generated by atlases

have been conducted. However, atlases have not yet attained their maximum potential, in part because of the lack of standard and high quality methods and presentations, and in part because sufficient information and analysis has generally not been provided along with disease maps to enable and encourage interpretation. As a step towards achieving increased comparability and utility of disease atlases to epidemiologists, a set of guidelines is proposed in Table 10.

While this list is rather long, the last four items ("Supplementary information or analysis") are less important than the others. Additionally, it is recognized that not all atlases can provide the volume of information indicated in the "Data" section: these items could be considered as desirable, but not essential. General acceptance of standard guidelines and provision of more analysis and correlate information could substantially increase the utility of disease atlases to epidemiologists.

Table 10: Proposed list of inclusions for disease atlases

A. Basic information:

1. A clear statement of purpose;
2. A commentary on the quality of the data, including its regional variation;
3. Justification of regional units mapped, and information about the stability of regional data;
4. Specification of criteria for mapping regional values (*e.g.*, minimal frequencies);
5. Specification of the criteria used to select and group diseases for mapping;
6. Specification of the age standardization method (and age grouping) employed. The direct method of standardization with an internal standard population is recommended, to facilitate comparisons between regions within an atlas. The adoption of standard age groups across atlases is also recommended.

B. Data:

7. Region-specific frequencies of mapped diseases and populations;
8. Regional values for the mapped data function, and its precision;
9. Age-specific frequencies and populations by region (to permit computations which would facilitate between-atlas comparisons).

C. Maps:

10. A key map to permit location and identification of specific regions;
11. Maps of event rates rather than frequencies. Significance levels should not be mapped to the exclusion of rates;
12. Clear distinction between rates and significance when a combined data function is used;
13. A logical colouring scheme;
14. Symmetric plotting groups, with equal emphasis to high and low risk, and to significance in either direction.

D. Supplementary information or analysis:

15. Time trend analysis;
16. Formal analysis of spatial structure;
17. Maps of relevant environmental and lifestyle correlates of health, preferably using the same regions as disease maps;
18. Analytic commentary on results, even if only to postulate hypotheses to explain the observed regional variation.

ACKNOWLEDGEMENTS

SDW is supported by a National Health Scientist Award from Health and Welfare, Canada. This project was also supported in part by a grant from the Ontario Ministry of Health.

The authors would like to thank the following individuals who assisted us with locating several of the atlases and in translation activities: Mr. Tom Broz, Ontario Cancer Foundation; Dr. Jacques Estève, IARC, Lyon; Dr. R. Frentzel-Beyme, Deutches Krebsforschungszentrum, Heidelberg; Dr. Yang Mao, Health and Welfare Canada, Ottawa; Fumihisa Matsumoto, Chiba, Japan; Dr. Chrisoph Minder, Universitat Berne, Switzerland; Dr. C. S. Muir, IARC, Lyon; Dr. R. Semenciw, Health and Welfare Canada.

REFERENCES

Walter S.D., and Birnie, S.E. (1991). Mapping mortality and morbidity patterns: An international comparison. *Int J Epidemiol*, 20, 678-689.

APPENDIX A. List of atlases surveyed

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 of cancer mortality in Belgium (1969-1976). Brussels: Institute of Hygiene and Epidemiology, Ministry of Public Health and Environment.

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. Health and Welfare Canada (1980). Mortality Atlas of Canada. Volume 1: cancer. Hull: Canadian Government Publishing Centre.

A6. Health and Welfare Canada (1980). Mortality Atlas of Canada. Volume 2: general mortality. Hull: Canadian Government Publishing Centre.

A7. Health and Welfare Canada (1984). Mortality Atlas of Canada. Volume 3: urban mortality. Hull: Canadian Government Publishing Centre.

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. International Agency for Research on 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. Bern: 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. Bern: Office Fédéral de la Statistique.

A30. Bisig, B., Paccaud, F. (1987). Répartition géographique des principales causes de décès en Suisse 1969/1972, 1979/1982. Bern: 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 al.* (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 al.* (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). Cancer Incidence: Atlas of the Province of Quebec 1982-1983. Montreal, The Cancer Research Society 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.

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

AN OVERVIEW OF ANALYTICAL METHODS & PRESENTATION TECHNIQUES IN MEDICAL GEOGRAPHY

G.J. Sherman[1]

ABSTRACT

Early successes in epidemiology were often in the study of spatial distribution of disease. An early, and often cited, analysis was that made by John Snow, who plotted cholera cases in 19th century London and succeeded in identifying a contaminated well as the source of the epidemic. Although analytic techniques have advanced enormously since that time, much of the medical geographic work done by epidemiologists today is essentially the same. Dot maps of cases and area maps of rates are still very popular, due in part to the visual appeal of any picture compared with a table, but should they continue to be considered sufficient? Is the ability of the human mind overrated when confronted with visual pattern recognition problem? How successful have disease maps been in "providing clues to etiology"?

A variety of mapping tools and techniques are discussed, as are a number of ancillary and follow-up strategies.

KEY WORDS: Spatial distributions of disease; Cartogram; Relative spaces; Clustering analysis.

1. INTRODUCTION

Some initial successes in epidemiology can be traced to the study of spatial distributions of disease, *e.g.*, that of John Snow who plotted cholera cases in 19th century London and thereby identified a contaminated well as the source of the disease. Although mapping analytical techniques have changed greatly since then, the idea is still much the same: cases identified as occurring in proximity are considered as an indication of the existence of a common risk factor. Such a finding might help to understand a disease etiology or at least provide clues for additional investigations. The study of spatial distributions is, if not easy, relatively straightforward and provides an good starting point in the explication of the exposure-outcome pathway.

The emphasis in this presentation is on techniques rather than on study results. Furthermore, due to obvious limitations of time and the vehicle of oral presentation, many specific techniques will be barely mentioned or not mentioned at all. The techniques that are mentioned in this presentation are covered because of their practicality, usefulness or acceptance. This is not to be taken as an endorsement for their use in any setting; technique is important but is only one part of the investigative process.

A small, but growing number of epidemiologic investigations have focused on the geographic distributions of disease. Most of this work uses geopolitical maps that identify areas of high and low outcome-, race-, sex-, specific, possibly age-adjusted rates. Further epidemiologic interpretation or statistical analysis has usually not been attempted. That is not to say, however, that an impressive array of techniques have not been developed and used. Gesler, in a 1986 paper which I have found useful for organizing my thoughts, divided spatial analytic techniques into six categories: those that dealt with points, lines, areas, surfaces, map comparisons and relative spaces. I will borrow heavily from his paper and use his organizational scheme.

---

[1] G.J. Sherman, Division of Childhood Diseases and Injuries, Health and Welfare Canada, Tunney's Pasture, LCDC, Building #6, Room 28C, Ottawa, Ontario, Canada K1A 0L2.

Areas

SPATIAL ANALYTIC TECHNIQUES

Figure 1: Areas

- Location Quotients
- Standardized Mortality Ratios
- Poisson Probability
- Space Clustering
- Space-Time Clustering
- Autocorrelation Measures
- Hierarchical Clustering

LCDC/HPB

Maps of areal units of disease can be constructed in many ways and represent one of the most common presentations. The simplest of these are based on natural discontinuities in discrete distributions or on the mean and standard deviations of distributions and are essentially descriptive. Methods such as location quotients and standardized mortality (or morbidity) ratios are commonly used simple forms of analysis and may be more useful in pattern assessment. There have been several instances in which the Poisson distribution has been used to identify sub-areas that have significantly high or low disease rates. One well-known example is the now four-volume Mortality Atlas of Canada series produced jointly be Health and Welfare and Statistics Canada.

A major limitation in plotting disease rates on geopolitical maps is the fact that geographic subunits such as counties are not defined in terms of the population-at-risk of the endpoint under study. Large, sparsely populated areas tend to visually dominate the map, whereas interest should be focused on areas with high populations. Rigorous statistical analysis of maps based on geopolitical boundaries is complicated for the same reason - the geographic subunits often contain greatly different populations-at-risk. Statistical analysis of geopolitical maps has usually consisted of a significance test to identify those high rates that were likely to have occurred from influences other than random fluctuations. Another way to deal with random fluctuations associated with small populations-at-risk is to combine sparsely populated areas into larger geographic units but this strategy decreases the specificity of a geographic analysis.

Ideal data for geographic analysis should contain the location of the cases of the disease under study. Data with exact locations are rarely available, the complete postal code being the most favourable possibility. Naturally, age- and sex-specific population denominators are also required on the same geographic basis.

Selvin *et al.* have described a computerized mapping technique in which the original geopolitical map (they used the census tracts of the City/County of San Francisco) is transformed to produce a cartogram, *i.e.*, a map which distorts location and geographic area in order to equalize density of some other quantity. Then statistically rigorous methods are used to identify non-random spatial clustering. In their example, the cartogram was defined so as to equalize density of the population-at-risk. On an ordinary geopolitical map, most of the inequality of risks among different geographic areas is due to differing population densities. The transformation equalizes population density across the entire map so that other factors influencing disease distribution can be more easily identified and analyzed.

The cartogram technique has two advantages over conventional methods: (1) complete geographic detail is preserved in the analysis, with no need to arbitrarily combine areas having small populations-at-risk; (2) the cartogram transformation preserves adjacency relationships so that information from adjacent areas can be visually or mathematically integrated in the interpretive model.

There are a number of published algorithms which can be used to transform maps. Many are not available in a "user-friendly" form.

Studies of the spatial distribution of cases of disease are generally easy to perform and relatively inexpensive and provide a helpful starting point toward the understanding of a disease etiology, particularly when environmental agents are suspect as risk factors. Potential confounding factors such as smoking, socio-economic status, and access to medical care (among others) may prevail as possible explanations for observed non-random patterns. A density equalized map eliminates only the distribution of the resident population as a confounding influence.

Tests for spatial clustering of disease and space-time clustering methods have been developed for areal data. To overcome problems of latency and mobility in the study of chronic disease outcomes such as Hodgkin's disease,

some researchers have used a case-control approach; *i.e.*, the formation of odds ratios which are tested for significance in standard ways in each of the areas under study.

Various measures of spatial autocorrelation, especially Moran's I statistic have been used to study disease patterns. Glick has formulated several ways in which Moran's autocorrelation statistic for interval data can be used to examine spatial patterns of diseases and to look for biologic, chemical, physical, cultural and ethnic factors that might be associated with these patterns. The weights used to calculate Moran's I can be based on simple adjacency of geographical units, proportions of common boundaries, whether or not two areal units have identical values on some categorizing variable and distance between the centres of the units. Spatial correlograms can be constructed to measure autocorrelation at different spatial lags. Glick has carried this approach to the extent of examining trends in the autocorrelation function across linear transects and examining residuals from trend models.

Grimson and co-workers have developed a method for determining hierarchical clusters of high risk areas by examining adjacencies between pairs of ranked rates. The observed number of adjacencies are compared to the results of Monte Carlo simulations that establish probabilities for the occurrence of joins among the appropriate number of units. Monte Carlo simulation is favoured by Grimson for situations where spatial data are not independent and areal units are irregular in shape: two very common occurrences.

Points

SPATIAL ANALYTIC TECHNIQUES
Figure 2: Points

- Mean Centre Standard Distance
- Standard Deviational Ellipse
- Gradient Analysis
- Nearest Neighbour Analysis
- Variance Mean Ratio Test
- Quadrat Analysis
- Space Clustering
- Space-Time Clustering

LCDC/HPB

Point maps have probably been used as frequently as areal maps and certainly predate them in the epidemiological study of disease; I have already mentioned John Snow. Of all the analytic techniques developed for point maps (*e.g.*, standard distance, standard deviational ellipses, nearest neighbour, quadrat), epidemiologists have concentrated on examining point patterns for possible clustering. Several different cluster methods have been developed but the ones which take the passage of time into account have received considerable attention. Knox is generally given credit for the basic space-time clustering concept. He stated that the detection of epidemicity in a set of data depends on a distribution in time, a distribution in space and interactions between these two dimensions. To examine interactions, pairs of cases which are relatively close in time are examined for their proximity in space. Pairs are classified according to both criteria and used to construct a contingency table. Observed pair frequencies can then be compared to expected values which are based on a time interval distribution formula. As already noted, space-time clustering has also been applied to areal data.

Lines

SPATIAL ANALYTIC TECHNIQUES
Figure 3: Lines

- Random Walk
- Vectors
- Graph Theory
  - Nodality
  - Connectivity
  - Dispersion
  - Nodal Hierarchies
  - Flow analysis

LCDC/HPB

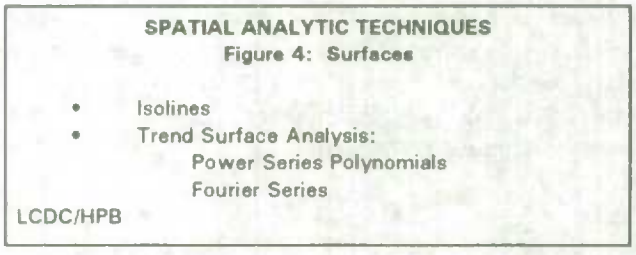
The collection of analytic techniques which Gesler refers to as "one-dimensional" or "line analysis" has been used very little for disease study. I do not know why this is except that the techniques are more purely mathematical than statistical and "user friendly" software to carry them out is definitely in short supply.

The concept of a random walk has been used to analyze the movement of the "clinical front" of a disease. The idea is to compare the actual direction of the disease movement with chance movements. Departures from expected directions could be suggestive of certain non-random constraints or environmental parameters which might be at work in certain locations.

Graph theory or network analysis has been used to some extent by medical geographers in both disease and health care delivery research. On the "disease side" networks have been developed in diffusion studies to indicate various types of relationships (called "joins") between the spatial units being investigated. The network of relationships is a convenient way of illustrating certain processes and need not be analyzed in terms of measures such as nodality or connectivity. This work is so foreign to me that I will depart from the form of the presentation to mention, as an example, Haggett's work in which seven alternative graphs were constructed to represent seven possible diffusion models of measles spread in south-western England: regional, urban-rural, local-contagion, wave-contagion, journey-to-work, population size and population density.

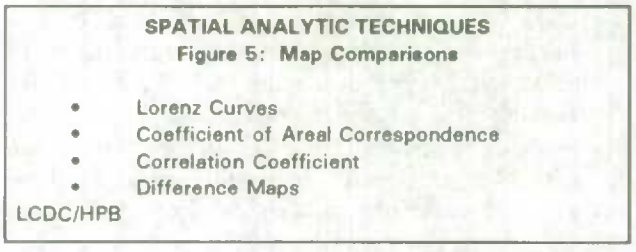
Graph theory has obvious applications to diffusing systems, thus it is not surprising that they have found more use in studies of hospital travel flows, disease progression, location/allocation modelling, *etc.*

Surfaces



A surface or scalar field can be constructed by using a "z" or "height" value in three dimensions. The same variable can be reduced to two-dimensional co-ordinates to draw isolines. Trend surface analysis is a well-known technique but has not been used extensively and then mostly in the analysis of disease diffusion processes, *i.e.*, in much the same setting as one-dimensional graph theory is used but with the advantage of a more (for most people) intuitive visual presentation.

Map Comparisons

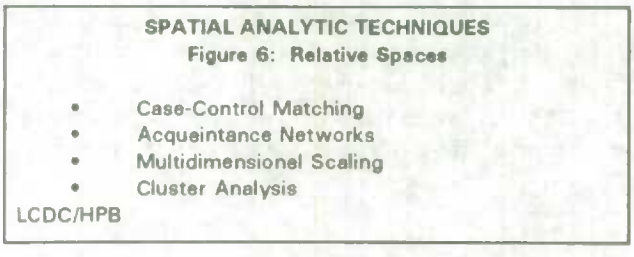


A technique which has become popular and accessible in recent years with the availability of powerful microcomputers and commercial software is map comparison. Of course, map comparison work was done long before microcomputers; dependent and independent variables were simply plotted on different maps of the same scale and visual comparisons were made. I believe that this is essentially the process most people try to go through when presented with an areal map of disease rates.

Probably the most used statistical method of map comparison is correlation analysis or "ecological correlation" using either a parametric or rank order correlation technique. This seems to be the approach of choice (other than simple overlays and thematic shading) of most microcomputer-based interactive mapping software available at this time. In fact, this approach is amenable to a great degree of statistical elaboration (*e.g.*, hierarchial clustering, factor analysis for the explication of underlying relationships between variables and for variable reduction) in the preparation of the measures or indices which are then compared.

Another type of map comparison technique that has been little used by medical geographers is based on the coefficient of areal correspondence which is the ratio of the area over which two phenomena are located together to the total area covered by the two phenomena (*i.e.*, a ratio of intersection to union of two phenomena).

Relative Spaces



Gatrell has called the dimensional analyses discussed so far the study of spatial arrangements because they are based on the metric properties of distance. Gatrell suggests that this is just the beginning of spatial analysis and what we should be thinking about is relative spaces and nonmetric relationships among sets of objects.

Multidimensional scaling is one of the tools in the "relative spaces" armamentarium. MDS is actually a class of techniques which use proximities among any kind of objects as input. A proximity is a number which indicates how similar or different two objects are, or are perceived to be, or any measure of this kind. The chief output is a spatial representation, consisting of a geometric configuration of points, as on a map. This configuration reflects the "hidden structure" in the data and often, after long reflection, makes the data easier to comprehend.

Clustering analysis (which is not the same as point clusters or space-time clustering) is a categorization technique that is based on a "taxonomic distance" between data items. The techniques of cluster analysis, of which there are many, are used in the exploration of data that arise from the measurement of a number of characteristics for each of an assorted collection of individuals or objects. The aim of the exploration is to determine if the objects can be subdivided into groups which can be shown to be relatively distinct or to belong together. The aim is clearly different from that of discriminant function analysis or similar assignment-type techniques which are used to allocate objects to known groups. Cluster analysis is concerned with the more difficult and intrinsically more interesting problem of finding the groups in the first place.

2. SUMMARY and CONCLUSIONS

I can agree, more or less, that a picture is worth a thousand words but I also believe that most people grossly overestimate their abilities to perform visual pattern analysis. Some people can play a dozen games of chess simultaneously in their head; maybe there are people who can mentally compute spatial autocorrelation at the 5th lag or do principal components analysis without a scratch pad. I've never met someone like that. My point is that county maps of disease rates; what I call "mapping data", have pretty well outlived their usefulness. As attractive as they are in six colours, they seldom tell us much. If the appropriate spatial analyses have been done and the results presented in well-designed map, we have probably gained something over viewing, say, a very large table of numbers. This isn't a trivial job as epidemiological, statistical, cartographic and artistic skills are all required. Others in this conference will have addressed these issues which are of fundamental importance.

I think it is time that more a sophisticated and, I believe, more informative treatment of disease outcome and covariate data should be used in our "atlas-type" publications. These are seen by a much wider and more general audience than that of the specialty journals to which they have been restricted. Our challenge is to mine the data for information and present it in an understandable way.

REFERENCES

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., Shaw, G., Schulman, J., and Merrill, D.W. (1987). Spatial distribution of disease: three case studies, *JNCI*, 79, 3, 417-423.

Selvin, S., Merrill, D., Schulman, J., Sacks, S., Bedell, L., and Wong, L. (1988). Transformations of maps to investigate clusters of disease, *Soc Sci Med*, 26, 2, 215-221.

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

RESEARCHING CANADA'S MEDICAL GEOGRAPHY - THE USE AND ABUSE OF FEDERAL SURVEYS AND DATA

M.W. Rosenberg and A.M. James[1]

ABSTRACT

Medical geography can be divided between studies of the spatial and temporal distribution of diseases and studies of access to health services. Both types of study lead medical geographers to employ federal surveys and data in the development of spatial statistics and statistical models. In addition, there are increasing numbers of researchers who are trying to link environmental data to health issues. These efforts raise questions about the appropriateness of scale and technique. To illustrate these themes, we review a selection of spatial statistics and models to identify some of the problems encountered in researching Canada's medical geography.

KEY WORDS: Spatial distribution; Temporal distribution; epidemiological waves; Spatial autoregressive models.

1. INTRODUCTION

Although medical geography is one of the smaller sub-disciplines of modern geography, its practitioners are extremely active nationally (*e.g.*, see Rosenberg 1990) and indeed internationally (*e.g.*, see Earickson 1988, 1990). Canada's medical geography, however, is to a great extent *terra incognita*. For all of our activities, we know relatively little about the links between health and the environment, about the spatial diffusion of diseases, about health behaviour and access to health services, or the rationale for the spatial organization of health services.

One answer to this paradox might be to argue that the explanation is simply one of numbers; there are very few medical geographers, and the number of issues that might define Canada's medical geography is enormous. We will argue, however, that a second answer is to be found through an examination of the kinds of questions medical geographers would like to answer, the spatial statistics medical geographers have developed and the data required to answer those questions.

The argument is organized in the following manner. In the next section, the interests of medical geographers are briefly explained. Then some examples of the use of spatial statistics in medical geography are described. In the third section, the role of federal surveys and data in examining Canada's medical geography is discussed. Finally, we conclude with a gaze into the future to answer where we need to go in the use of spatial statistics and federal government surveys and data to study Canada's medical geography.

2. THE STUDY OF MEDICAL GEOGRAPHY

Historically, medical geographers have mainly been interested in two broad themes: the study of the geography of the spatial and temporal distribution of diseases and the study of the geography of health care delivery (See Akhtar 1991; Meade *et al.* 1988; Jones and Moon 1987). These themes have been linked conceptually by a more general interest of geographers about the relationships between humanity and the environment in which we live. In medical geography, this concern might be expressed as the links between health and the environment.

---

[1] M.W. Rosenberg and A.M. James, Department of Geography, Queen's University, Kingston, Ontario, Canada K7L 3N6.

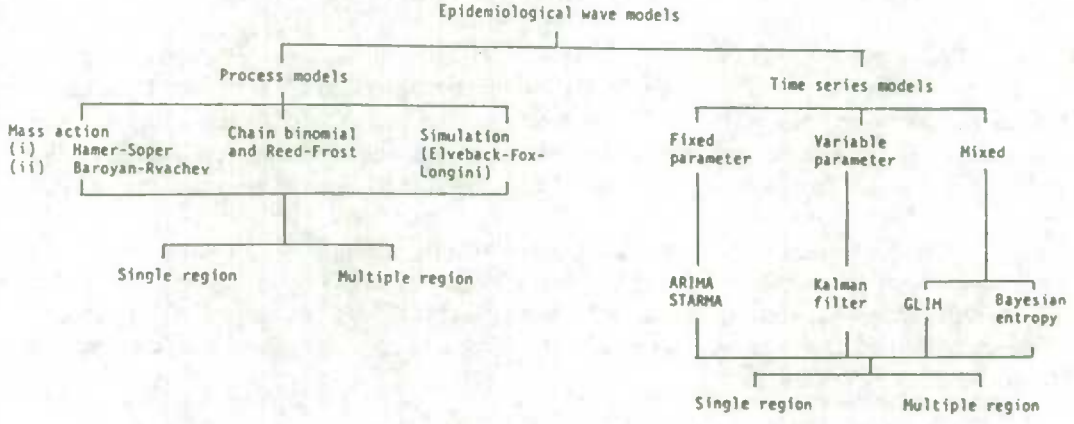
The first step in almost any medical geographer's research is to map morbidity and mortality rates, the location of populations at risk, or the location of physicians, health care services and potential users of medical services. It is the visualization of spatial patterns of disease, mortality rates, or the location of health care delivery which has been the incentive for medical geographers to develop spatial statistics. They allow medical geographers to link the parameters of health, life and indeed death with socio-economic characteristics of the population affected and the environment in which they live to understand the complex links among spatial patterns, nature, and cultural, economic, political and social processes.

3. SPATIAL STATISTICS

Three classes of spatial statistical modelling which have been particularly important in medical geography are: epidemiological wave models; spatial autoregressive models; and spatial ecology models. Time and space do not allow a discussion of these models in any detail, so we briefly comment on each.

Cliff and Haggett (1986) provide a useful review of epidemiological wave models. Figure One summarizes epidemiological wave models as either process models or time series models. Process models "generate or simulate the biological process by which susceptibles become infected with a disease" whereas time-series models incorporate the historical record of a disease to identify "the shape of the generating forces which underpin it" (Cliff and Haggett 1986, p. 85). Whether one chooses to model disease diffusion as a process or time series model, the central concerns are the actual modelling of the waves, wave threshold and wave shape.

Figure 1: Epidemiological Wave Models



Source: Cliff and Haggett (1986:87).

Using the Hamer-Soper model as an example, Cliff and Haggett (1986, pp. 86-89) illustrate one way modelling epidemiological waves has progressed over time. In Equation One, the number of infectives, I, in time $t + 1$ is equal to the number of infectives in time t plus a diffusion coefficient, b, multiplied by the number of susceptibles, S, multiplied by the number of infectives in time t minus a recovery coefficient, c, multiplied by the number of infectives in time t.

$$I_{t+1} = I_t + bSI_t - cI_t \qquad (1)$$

To make the model more realistic, births need to be taken into consideration in calculating the number of susceptibles. In Equation Two, susceptibles in time $t + 1$ are equal to susceptibles in time t minus the diffusion coefficient multiplied by the susceptibles multiplied by the infectives in time t plus the number of births, a.

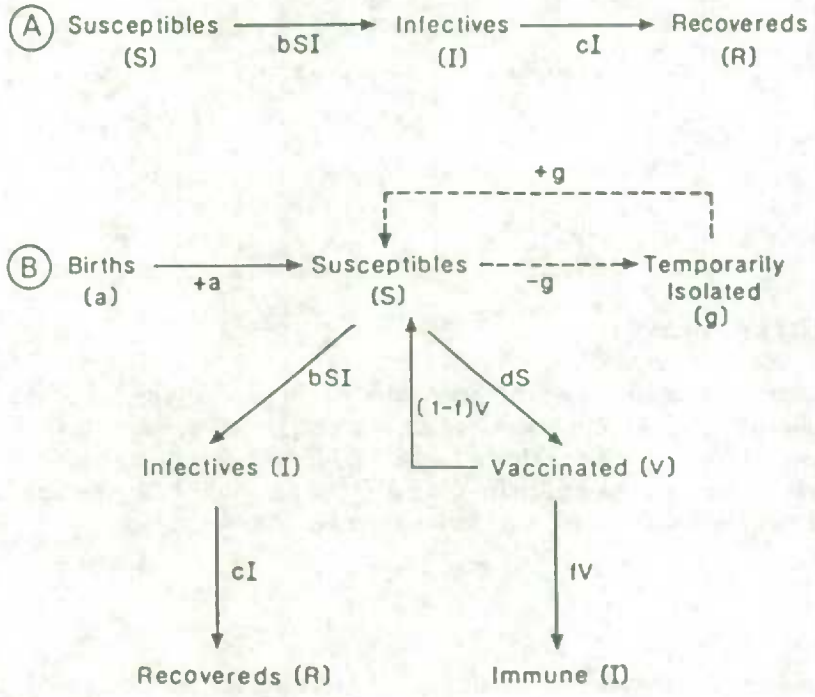
$$S_{t+1} = S_t - bSI_t + a \quad (2)$$

A third modification to Equation One can be found in Equation Three. Added to Equation Two are expressions to define persons vaccinated, V, the vaccination rate coefficient, d, the number who become immune, I, and the successful immunization rate, f.

$$S_{t+1} = S_t - bSI_t + a - dS_t + (1 - f)V_t \quad (3)$$

The development of Equations One to Three is summarized in Figure Two.

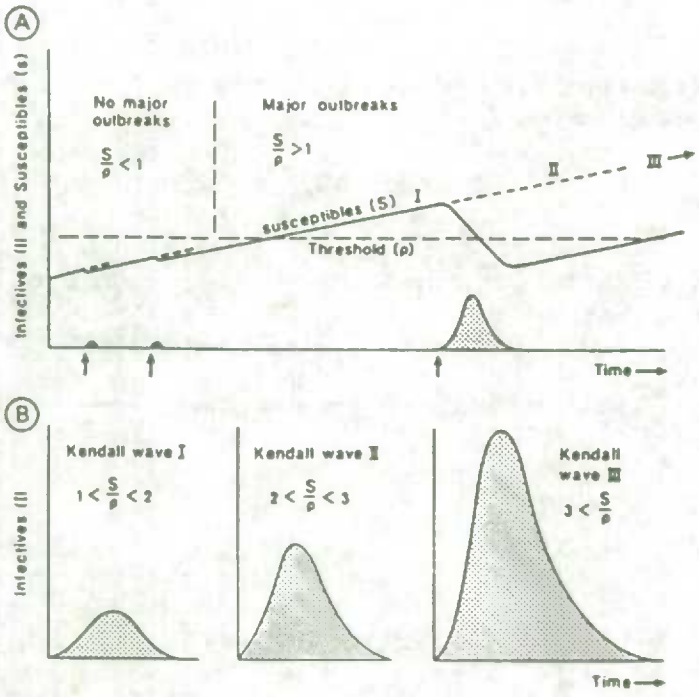
Figure 2: Wave Generating Models



Source: Cliff and Haggett (1986:88).

Wave threshold can be measured using a simple ratio, $p=c/b$, where p is the threshold, c is the recovery coefficient and b is the diffusion coefficient (Cliff and Haggett 1986, p. 98). Threshold measures can be linked back to wave models as illustrated in Figure Three.

Figure 3: Relationships between the Slope of Epidemic Waves and Wave Threshold



Source: Cliff and Haggett (1986:99).

Building on the measures of wave modelling and threshold, wave shape measures seek to characterize the various aspects of the spread of a disease through measures such as average time lag, wave duration, wave velocity, wave kurtosis and diffusion coefficients (Cliff and Haggett 1986, pp. 103-106). Equation Four defines a simple measure of average time lag in the waves of an epidemic where T is the duration of the epidemic in months, x_t is the number of reported cases in month t and n is the total number of cases reported.

$$\bar{t} = \frac{1}{n} \sum_{t=1}^{T} t x_t \tag{4}$$

Equations Five and Six define standard wave duration, s, and from these equations, wave kurtosis, b_2, is defined in Equations Seven and Eight.

$$\text{Standard Wave Duration, } s = \sqrt{m_2} \tag{5}$$

where:

$$m_2 = \frac{1}{n} \sum_{t=1}^{T} (t - \bar{t})^2 x_t \tag{6}$$

and

$$\text{Wave Kurtosis, } b_2 = m_4 / m_2^2 \tag{7}$$

where:

m_2 is specified as above;

$$m_4 = \frac{1}{n} \sum_{t=1}^{T} (t - \bar{t})^4 x_t \tag{8}$$

The diffusion coefficient, b, is defined in Equations Nine and Ten as a logistics model where P_t is the cumulative proportion of cases up to time t, e is the base of natural logarithms and a and b are parameters to be estimated so that in Equation Ten, b represents the average rate of growth.

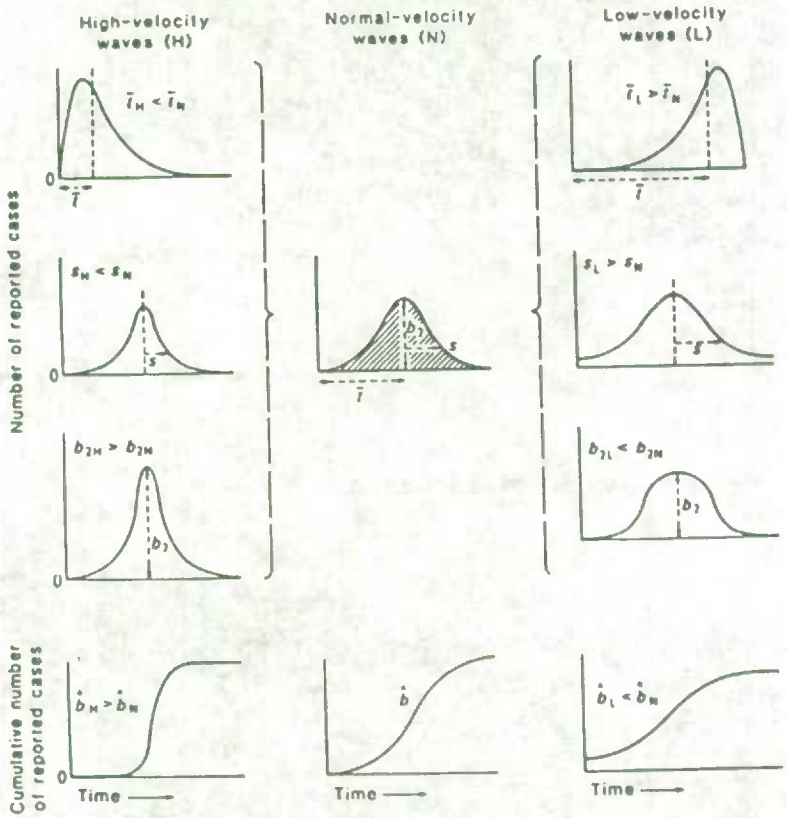
$$P_t = 1/(1 + e^{a-bt}) \tag{9}$$

where:

$$\ln\left(\frac{1}{P_t} - 1\right) = a - bt \tag{10}$$

The relationships between average time lag, standard wave duration, and actual and predicted diffusion coefficients are illustrated in Figure Four. Note that the key to this kind of spatial modelling is incidence data which are georeferenced consistently over time and space.

Figure 4: Geographical Definitions of Wave Shape and Measures



Source: Cliff and Haggett (1986:104).

Spatial autoregressive models have been used to identify the importance of regional and neighbourhood effects in the geographic distribution of mortality data where spatial dependency is common. Kennedy (1988) provides an example of how one might specify such a model. Equation Eleven (adapted from Kennedy (1988)) defines a spatial autoregressive model where the regional and neighbourhood effects are simultaneously estimated.

$$Y_i = \underbrace{B_1 h_i + B_2 p_i}_{\text{Regional trend component}} \underbrace{+ B_3 Z_{3i} + \ldots + V_7 Z_{7i} + e_i}_{\text{Neighbourhood effect component}} \tag{11}$$

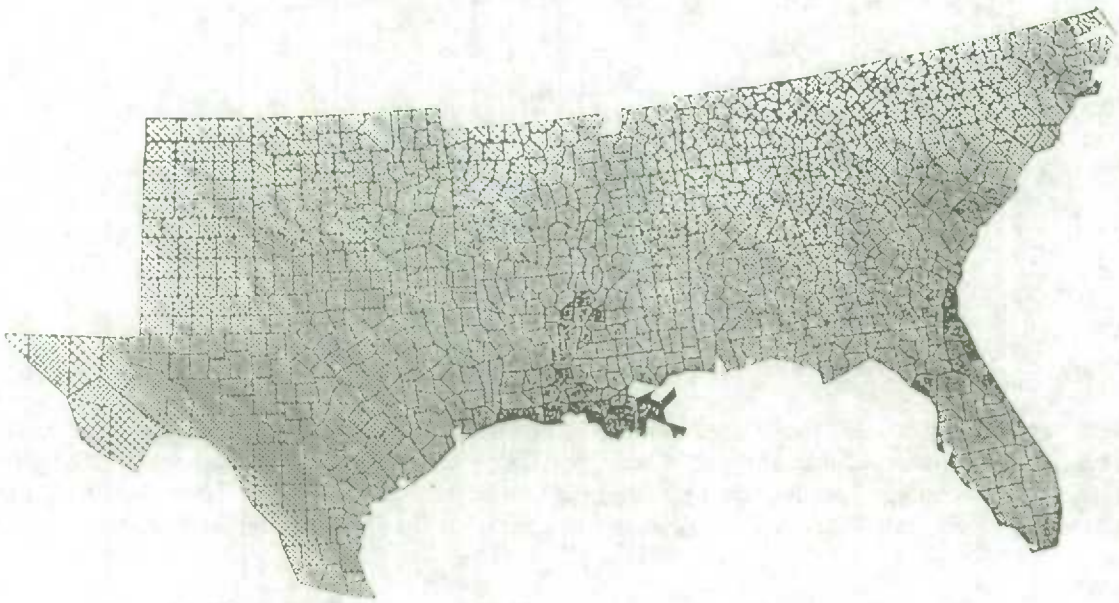
Y_i is the age standardized death rate in county $_i$. h_i is $\mathrm{Cos}(p_i)*h_i$ where h_i is the longitude coordinate of the centroid of county $_i$ and p_i is the latitude of the county centroid. Z_{ik} is equal to $W_{ij}W_j$ where W is the weighting matrix for the j th neighbour of county $_i$ and V is the matrix of death rates for the neighbours $_j$ of county $_i$, and k is the order of neighbour. e_i is the error term.

Figure Five shows the actual distribution of male lung cancer mortality by county in the Gulf of Mexico and Southeastern United States and Figure Six shows the predicted distribution of male lung cancer mortality by county generated using the spatial autocorrelation model (Kennedy 1988).

Figure 5: Male lung cancer by county in the Gulf of Mexico and southeast Atlantic Coast

Source: Kennedy (1988:121).

Figure 6: Predicted Male Lung Cancer Mortality



Source: Kennedy (1988: 125).

To paraphrase Susan Kennedy (1988, p. 120), the scale of spatial variation plays an important role in the geographic approach to modelling disease. A disease might exhibit a great deal of variation at one scale of aggregation, but show little variation at another scale. Therefore, it is important to analyze data at the scale at which it exhibits spatial variation. We return to this theme later in the paper.

Spatial ecology models are by far the most commonly presented in medical geography. Typically, they are in the form of a multiple regression equation where the dependent variable is a count or rate per spatial unit of mortality or a measure of access to health care and the independent variables are measures of the socio-economic or physical conditions in the corresponding spatial units. One suspects that at least part of their popularity is a function of a lack of data on individuals which can be accessed at a geographic scale below the provincial level.

4. THE USE AND ABUSE OF FEDERAL SURVEYS AND DATA

Where do medical geographers in Canada look for their health data when they do not choose to collect it for themselves? Most notably, they look to *Vital Statistics* and *the Health Reports*, and less often to *the General Social Survey* and *the Health and Activity Limitation Survey*.

Almost all of the published examples of research by medical geographers take variables from *Vital Statistics* or *the Health Reports* and use these as dependent variables, and take data from the Census or environmental databases as their independent variables to fit spatial ecology models. For example, Foster and Norrie (1988) used cancer data from *the Mortality Atlas of Canada* and correlate mortality rates for specific cancers at the census division level with water quality information taken from a database created by the University of Ottawa and Health and Welfare Canada. Godon *et al.* (1989) used *the Quebec Cancer Registry* and agricultural and socio-economic data from *the Census* to show the ecological links between some cancers and pesticide use in rural areas of Quebec. One of their conclusions, however, is that to demonstrate causality would require research at the individual level.

What of data at the individual level? We know of no published examples where medical geographers have used *the General Social Survey*. Moore *et al.* (1990) and Moore and Rosenberg (1991) have used *the Health and Activity Limitation Survey* extensively, but this research has been mainly descriptive in drawing out the basic demography of disabled persons in Ontario.

This is not to say that medical geographers do not analyze data at the individual level. There are many published examples of research where data have been collected by medical geographers in their own surveys (*e.g.*, Shannon *et al.* 1988) or individual data have been combined with socio-economic data from *the Census* (*e.g.*, Liaw *et al.* 1989). Methodologies have ranged from simple correlation to logit models.

What is also striking from a survey of the recent published literature, however, is that no examples can be found of epidemiological wave modelling or spatial autoregressive modelling of diseases in Canada by Canada's medical geographers.

5. THE FUTURE USE OF SPATIAL STATISTICS AND FEDERAL SURVEYS IN STUDYING CANADA'S MEDICAL GEOGRAPHY

What do the previous sections so far tell us about the use of spatial statistics and federal surveys in studying Canada's medical geography? Medical geographers in Canada are a small group where the preponderance of their research has depended on spatial ecology models, and their data have come mainly from *Vital Statistics*, *the Census*, some environmental databases and their own surveys.

Are medical geographers incapable of employing some of the more sophisticated epidemiological wave modelling techniques and spatial autoregressive models? In all honesty, some probably are, but others are indeed capable of employing these techniques. What is also obvious to us, but may be less obvious from the literature which has

been reviewed, is that medical geographers want to seek explanations which link people's health to where they work and live, to their lifestyles and their behaviour. In this respect, they find themselves faced with a dilemma.

No one who has used *the General Social Survey* or *the Health and Activity Limitation Survey* can criticize the quality of the data or the breadth of its information. The irony is that it contains almost no geography to allow medical geographers or indeed other medical or social scientists to answer questions about *whether geography matters* to borrow a phrase made famous by Doreen Massey (1984).

To return to Susan Kennedy's comment about the importance of spatial variation in scale. We are not likely to see more examples of spatial autoregressive models or epidemiological wave models in the study of Canada's medical geography unless data are made available at the census subdivision level or even smaller geographic scales. The corollary of this is that the development of spatial statistics to understand Canada's medical geography will be slowed because the data needed to test the models either is inaccessible to medical geographers in practical terms or does not exist at all.

At this point, it would be easy to conclude by saying that if Statistics Canada would simply provide more data at smaller geographic scales everything else would solve itself. We do not believe this. Instead, we would like to make the following suggestions. Closer collaboration is needed between medical geographers and the statisticians and demographers of Statistics Canada who are particularly interested in health and health care. The collaboration might come in the form of working together in-house using data at small geographic scales to do some of the kinds of modelling described. This might solve some of the problems of accessibility which are connected either with issues of confidentiality or cost. Second, on custom surveys, consult medical geographers to develop questions about people's geographies so that analysis can be carried out to test the links between health, lifestyle, workplace, home and the physical environment. Third, a topic which we have not discussed is how medical geographers and demographers and statisticians at Statistics Canada might collaborate on developing and testing methods for estimating rates at larger geographic scales which could be used to make reliable estimates at smaller geographic scales.

So what is the abuse of federal surveys and data in studying Canada's medical geography? The abuse is how little research is being carried out on the links between health and the environment at a time when there is increasing public concern over this issue. Our belief that spatial statistics might contribute to this study will, however, only be validated when medical geographers in Canada begin to exploit the techniques discussed in this paper and in other papers in this proceedings and concomitantly, when federal surveys and data provide the types of information needed to make it worthwhile to exploit spatial statistics in the study of Canada's medical geography.

REFERENCES

Akhtar, R. (1991). *Environment and Health*. New Delhi: Ashish Publishing House.

Cliff, A.D., and 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., and Norrie, I. (1988). Water quality and cancer of the digestive tract: the Canadian experience. *Proceedings of the Third International Symposium of Medical Geography*, M. Anderson, M. Rosenberg and R. Tinline, eds. Kingston, Canada: Queen's University, Department of Geography, 91-101.

Godon, D., Thouez, J.-P., and 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., and 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., and 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!* D. Massey and J.Allen, eds. Cambridge: Cambridge University Press, 1-11.

Meade, M., Florin, J.W., and Gesler, W.M. (1988). *Medical Geography*. New York: Guilford Press.

Moore, E.G., and Rosenberg, M.W. (1991). Disability in the adult population living in Ontario Institutions. A report prepared for the Ontario Ministry of Community and Social Services. Kingston, Canada: Queen's University, Department of Geography.

Moore, E.G., Burke, S.O., and Rosenberg, M.W. (1990). The disabled adult residential population of Ontario. A report prepared for the Ontario Ministry of Community and Social Services. Kingston, Canada: Queen's University, Department of Geography.

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., and Sackett, D.L. (1988). Lung cancer and air pollution in an industrial city - a geographical analysis, *Canadian Journal of Public Health*, 79, 255-259.

Statistics Canada (1985). *General Social Survey - Health and Social Support*. Ottawa: Supply and Services.

Statistics Canada (1986). *Health and Activity Limitation Survey*. Ottawa: Supply and Services.

Statistics Canada. *Vital Statistics: Causes of Death*. Ottawa: Supply and Services (Comes in various years and volumes).

Statistics Canada. *Health Reports*. Ottawa: Canadian Centre for Health Information (Comes in various years and volumes).

SESSION 8

Spatial Analysis of Survey Data

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

HEALTH DIFFERENCES BY NEIGHBORHOOD CHARACTERISTICS

Russell Wilkins[1]

ABSTRACT

Through use of the Statistics Canada Postal Code Conversion File and postal code finder software, health data files for which address data are available can be coded to the census tract, enumeration area or block-face level with a moderate amount of manual intervention. When the health data thus coded are related to census or other data on neighborhood characteristics for the same areas, various types of analyses are possible. Questions of equity, environment and administrative effectiveness can thereby be addressed at least indirectly, without resort to collection of additional data. Examples of such work include studies of trends in infant mortality and low birth weight by neighborhood income in urban Canada. For the non-institutional population living in urban areas, simpler analyses using only the first three digits of the postal code are also possible, even in the absence of full address data.

KEY WORDS: Postal codes; small area data; socioeconomic status.

1. INTRODUCTION

Health data by local area are needed for purposes of planning, surveillance, evaluation and research. Data on births, deaths, and hospitalization are routinely coded to the municipal level, but within large cities where much of the population resides, more precise geographic coding is required to attribute health events to service districts. Health data coded to census tract or smaller units can also be used to study socioeconomic inequalities and environmental hazards, based on neighborhood characteristics.

In the past, small-area geographic coding had to be done manually using street indexes and map look-ups. Recently, however, through use of the Statistics Canada Postal Code Conversion File (PCCF) and postal code finder software, much of this work can now be automated. Increasingly, it will be feasible to produce and use small area data on a routine basis.

After briefly explaining how this coding can be done, and from which files, I will give a few examples of the kind of findings which can be drawn from such studies, based on work I have been involved with over the last few years.

2. METHODS

2.1 Sources of data

Health-related event data amenable to analysis by local area include vital statistics on live births, deaths, and stillbirths, hospital morbidity statistics from admission-separation records, data from disease registries (such as for cancer or renal failure), census disability data from the 20% sample, data on persons in long-term care, and *ad hoc* or special purpose data collections on a variety of subjects.

---

[1] Russell Wilkins, Canadian Centre for Health Information, Statistics Canada, 18-N, R.H. Coats Building, Ottawa, Ontario, Canada K1A 0T6.

Denominators for the calculation of rates for these events may sometimes be available in the same file (such as for rates of low birth weight calculated from the live births file), although the usual source of denominator data is from census population counts by local area. Census population data by local area may be obtained for any standard census geographic units or aggregates thereof, or in the case of the non-institutional population living in urban areas, for non-standard "forward sortation areas" (corresponding to the first three digits of the postal code). Alternate sources for denominator data include population registry-type information generated from specially compiled lists of provincial health care beneficiaries (such as in Québec and Saskatchewan), or from tax filer data compiled by Statistics Canada for urban forward sortation areas and rural postal code service areas.

2.2 Types of analyses based on the coded data

Health data coded to local areas can be used to describe health conditions geographically (in terms of the smallest units for which statistically reliable data can be produced), aggregated to local administrative or program service areas (such as health and social services areas, units, districts and regions), or aggregated according to neighborhood characteristics such as the percentage of population under the low-income cut-off. In the latter case, area of residence is used as an indicator of socio-economic status. At a rather cruder level of approximation, both geographic and socio-economic oriented analyses can be performed using only the first three digits of the postal code (FSAs) rather than the more precise census tracts or enumeration areas based on the full 6-digit postal code.

2.3 Use of the Postal Code Conversion File (PCCF)

The PCCF (Statistics Canada 1991) is used to relate postal codes to local area geography. In urban areas of Canada, approximately 30 persons are served by a typical postal code, and those persons usually reside within a single block face and enumeration area. In such cases, the PCCF will show that the postal code is related to a unique map centroid, which is in turn related to the entire hierarchy of census geographical units: from blockface and/or enumeration area to census tract, census metropolitan area or census agglomeration; or from blockface and/or enumeration area to census subdivision, census division, province and region.

Outside of urban areas, however, each rural postal code refers to the entire area served by a rural post office, which may include several thousand persons. In the PCCF, almost all rural postal codes are thus related to more than one enumeration area, and frequently to more than one census subdivision (or municipality). Rural postal codes in Canada are easy to identify because they always have a zerp in the second position.

A detailed description of methods used for coding postal addresses to small areas was presented in a Statistics Canada workshop following the symposium, and will be published in a future issue of *Health Reports* (Statistics Canada catalogue 82-003). SAS routines to match postal codes to precise geographic location using the PCCF are available on request from the author at the Canadian Centre for Health Information.

2.4 Analyses based on the first three digits of the postal code

Where address data are not available for verification of questionable postal codes, or where a simpler and quicker method of analysis is desired, it is frequently possible to do a somewhat cruder, but nevertheless meaningful analysis based on only the first three digits of the postal code, which Canada Post refers to as the Forward Sortation Area (FSA). In urban areas, a typical FSA may contain 20,000 to 40,000 persons, which makes it 5 to 10 times larger than a typical census tract of 4,000 persons. Note, however, that the size of FSAs are based on the volume of mail received, so an FSA in a downtown area with many businesses will have a smaller population than an FSA in a suburban area with few businesses. Although less socially homogeneous than census tracts, FSAs provide a convenient way of dividing large urban areas into relatively small sub-municipal units which can be used for mapping rates of occurrence of health events, grouping into approximations of health and social services districts, or aggregating according to socioeconomic characteristics.

Census profile data are available from Statistics Canada for the non-institutional population living in urban FSAs (from CANSIM or regional offices). Rural FSAs are excluded as they overlap with urban FSAs, and the institutional population is excluded because the postal code data on the census questionnaires was only captured from households in the 20% sample, which excludes institutional residents. Since rates of institutionalization are

high at ages 75 and over (the highest age group shown in the FSA profiles), I recommend limiting analyses of FSA-based data to the population aged less than 75, in which rates of institutionalization are very low.

3. RESULTS

3.1 Coding data to census tract

Based on data prepared for a study of unfavourable birth outcomes and infant mortality by neighborhood income (Wilkins, Sherman and Best 1991), Table 1 provides an example of local area data coded to census tract. For

TABLE 1: BIRTHS BY CENSUS TRACT (CT): CHARACTERISTICS OF MOTHERS AND NEWBORNS, BY CENSUS METROPOLITAN AREA (CMA), URBAN CANADA, 1986 (ROW PERCENTAGES)

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 À LA NAISSANCE = INSUFFISSANCE 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: * INDICATES COEFFICIENT OF VARIATION FROM 16.7% TO 33.3%: USE WITH CAUTION.
. INDICATES COEFFICIENT OF VARIATION OVER 33.3%: DATA SUPPRESSED.
-- INDICATES ZERO PERCENT.

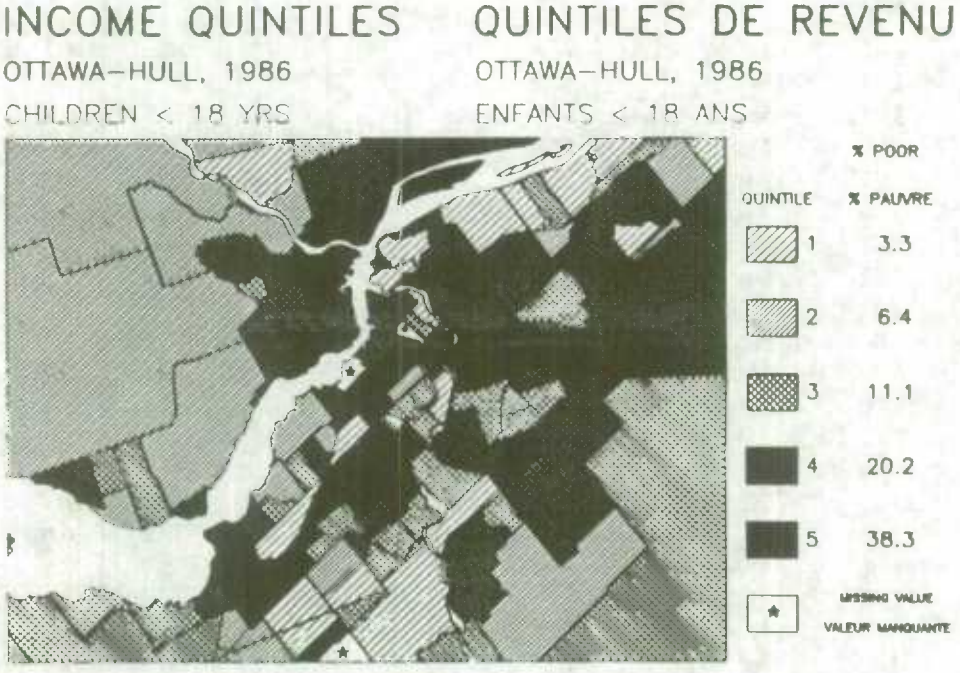
SOURCE: CANADIAN CENTRE FOR HEALTH INFORMATION, STATISTICS CANADA.
FILE=SUM1CTS (COUNTS), SUM2CTS (%); PROGRAM=BBILON2 1990-03-30

each census tract, the total number of live births is shown, as is the percentage of births by the age, marital status and place of birth of the mother, and the percentage of births which were of low birth weight, premature, or small for gestational age. More stable rates could have been obtained by grouping data for several years. Health data coded to census tract can be used to identify target populations, and/or aggregated to correspond to the areas served by local and regional health and social services districts.

3.2 Grouping census tracts by socio-economic characteristics

An alternate way of using small area data is to first group census tracts according to socioeconomic characteristics, and then calculate event occurrence rates. Map 1 shows how census tracts in the Ottawa-Hull Census Metropolitan Area (CMA) were aggregated into quintile groups based on the percentage of population under 18 years of age living in families whose income was below the Statistics Canada low income cut-off (according to special tabulations of the 1986 census for the Canadian Centre for Health Information). In this case, the cut-points were chosen so as to allocate one-fifth of the total births to each quintile group. This is not the same as allocating one-fifth of the census tracts to each group, since for historical reasons, central-area census tracts tend to be both smaller and poorer, while suburban census tracts tend to be larger and less poor.

Map 1



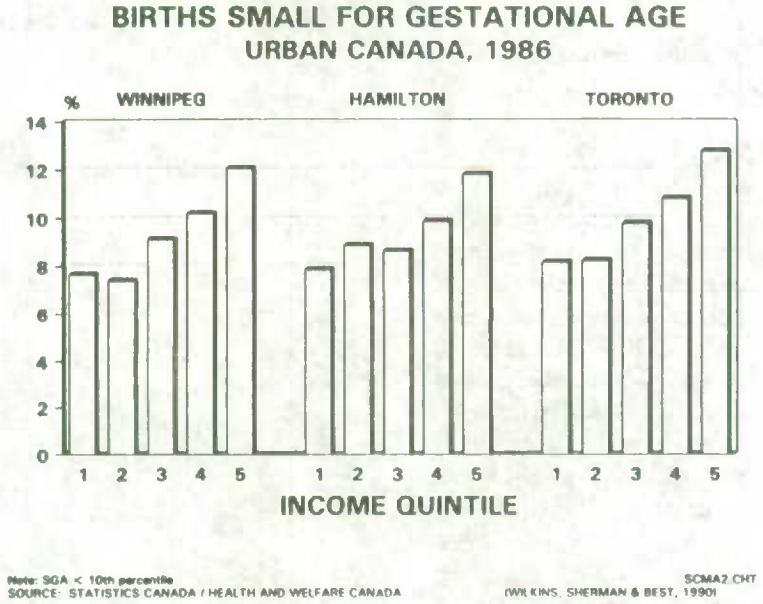
3.3 Rates of births small for gestational age in three metropolitan areas

Figure 1 shows the results of analyses for three CMAs after grouping census tracts as described above (Wilkins, Sherman and Best 1991). In Winnipeg, Hamilton and Toronto, the percentage of births small for gestational age was about 50% higher in the poorest areas (Quintile 5) compared to the least poor areas (Quintile 1). In the study from which these data were taken, other analyses made use of the full range of individual data available on the birth records, including the birth weight and gestational age of the child, and the age, parity (number of previous births), marital status and birth place of the mother. The only "ecologic" variable employed was the percentage of low income children in the neighborhood, which was used to combine similar census tracts into quintile groups.

Note that this type of analysis--where neighborhood of residence is used as a proxy for socioeconomic status--is not appropriate outside of fairly large urban areas. This is because in small towns and rural areas, the smallest residential areas which can be reliably defined in terms of postal codes or postal addresses is the municipality or group of municipalities sharing a common post office, in which case all socioeconomic classes will be lumped

together. Where all classes share the same geographic code, ecologically-based studies tend to obscure, rather than manifest any real differences by socioeconomic status which might exist at the individual level.

Figure 1



3.4 Lung cancer mortality by region and district

Indirectly standardized mortality ratios (SMRs) for lung cancer were calculated for each of the main health regions of Québec (Hoey, Wilkins, Gagnon and O'Loughlin 1987). The difference in SMRs between the highest and lowest of the regions was about 50%, and the Metropolitan Montreal health region (which includes the entire Island of Montreal census division) was ranked just slightly better than average. However, when SMRs for each of the 7 community health districts (DSCs) within the Montreal region were calculated, it was found that the difference between the DSCs in Montreal was much greater than was the difference between the regions across the province. Note that DSCs in Montreal have an average population of about 200,000 persons, and are by no means completely homogeneous in terms of the socioeconomic status of their populations. Nevertheless, the lung cancer mortality rates were over twice as high in the relatively poorer DSCs (Verdun and St-Luc), compared to the relatively affluent DSCs (West Island and Ste-Justine). Moreover, when the smaller and relatively more homogeneous local community service centre (CLSC) districts were used as the unit of analysis, the differences in health outcome indicators were even more pronounced.

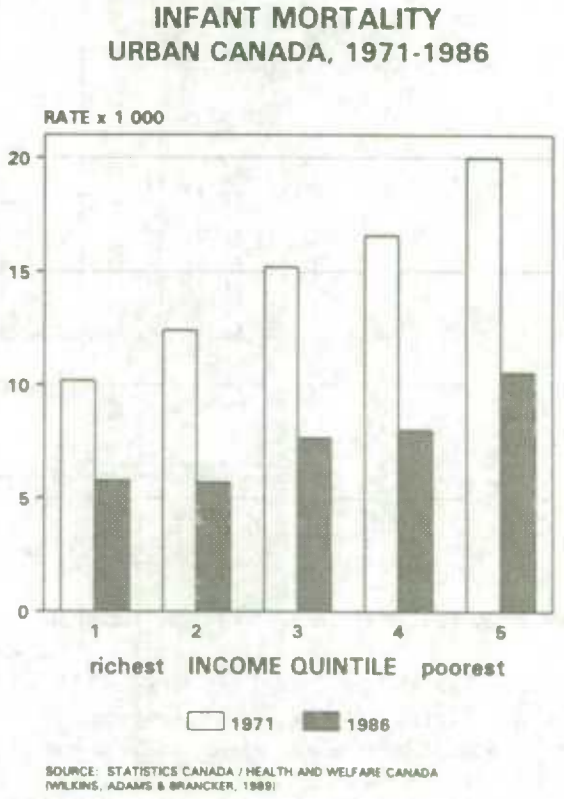
For comparisons across several areas, I prefer to use directly rather than indirectly standardized rates, so that the same set of weights will be applied to the specific rates in each area. (In the previous example, indirectly standardized weights were used for the DSCs in Montreal only because the results were to be shown alongside previously published regional data which had been indirectly standardized.) Local populations may sometimes contain disproportionate numbers of men (or women), in which case standardization by sex as well as age is important for purposes of comparability.

3.5 Trends in infant mortality by income from 1971 to 1986

Figure 2 shows infant mortality rates by income quintile group in 1971 and 1986 (Wilkins, Adams and Brancker 1989). In 1971 as in 1986, census tracts in each Census Metropolitan Area were combined into five quintile groups of approximately equal population size, based on the percentage of population below the low income cut-offs applicable at that time. Note that while the same census tract was not necessarily classified into the same income quintile in each period, the principle of classification was the same in both cases. This sort of

analysis provides the only evidence we have of long term trends in the evolution of health inequalities in Canada--even though the reduction of health inequalities is officially recognized as one of the fundamental goals of Canadian health policy. In this case, we see that although in both periods the infant mortality rates in Quintile 5 were nearly twice as high as in Quintile 1 (the rate ratios were similar), the absolute differences in rates was only half as great in 1986 as in 1971.

Figure 2



3.5 Rates of injury to child pedestrians and bicyclists in Montreal

Other locally collected and *ad hoc* datasets can also be analyzed in this fashion--for example, data on injuries to child pedestrians and bicyclists which resulted in hospital treatment or police reports (Dougherty, Pless, and Wilkins 1990). After coding the census tract of place of residence of the victims from their postal codes and addresses, it was found that rates of injuries were nearly 5 times higher in the poorest areas of Montreal compared to the least poor areas. In this case, unsafe local neighborhood environments would be expected to affect all residents of an area, regardless of their family income, so that geographic location should be important in its own right, and not merely as an indicator of neighborhood socioeconomic status. Both individual (or family) and neighborhood characteristics would need to be available simultaneously before one could hope to disentangle those two effects.

3.6 Other ways of making use of local area data from health files

Analyses making use of only the first three digits of the postal code may also provide very useful information at the submunicipal level in larger urban areas. This may be necessary, for example, when the full six-digit postal code is not available (Choinière 1991), when address data are not available to permit dealing with the incorrect, missing or problematic postal codes (Gentleman, Wilkins, Nair and Beaulieu 1991) or when a "quick and dirty" analysis of a dataset with complete addresses and postal codes is desired. In this kind of analysis, the observed relationships are typically less strong than those seen in analyses based on more precise geographic coding.

In the previously mentioned study of births by income (Wilkins, Sherman and Best 1991), the ratio of births to the census population under one year of age was used to estimate the extent and distribution of the census undercount relating to families with young children, and to do this by Census Metropolitan Area and neighborhood income. The results of this operation compared well to estimates based on the reverse record check (Germain 1988; Statistics Canada 1988), but were available at a much more detailed level of aggregation.

In the mortality by income study (Wilkins, Adams and Brancker 1989), postal codes were also used as an aid in identifying institutional residents. Moreover, when mortality rates were compiled by place of birth and neighborhood income, it appeared that the relationship of income to mortality was confounded somewhat at the neighborhood level by the generally lower mortality of immigrants, who were also more likely to live in relatively poorer neighborhoods of big cities.

4. CONCLUSION

Health data coded to local areas can be used to describe health conditions geographically, aggregated to local administrative or program service areas, or aggregated according to neighborhood characteristics such as the percentage of population under the low-income cut-off.

When health data coded to local area are related to census or other data on neighborhood characteristics for the same areas, various types of analyses are possible. Questions of equity, environment and administrative effectiveness can thereby be addressed at least indirectly, without resort to collection of additional data. Examples of such work include studies of trends in infant mortality and low birth weight by neighborhood income in urban Canada. For the non-institutional population living in urban areas, simpler analyses using only the first three digits of the postal code are also possible, even in the absence of full address data.

REFERENCES

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., and Wilkins, R. (1990). Social class and the occurrence of traffic injuries and deaths in urban children, *Canadian Journal of Public Health*, 81, 204-209.

Gentleman, J. F., Wilkins, R., Nair, C., and Beaulieu, S. (1991). An analysis of frequencies of surgical procedures in Canada, *Health Reports*, 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, Social Survey Methods Division, Statistics Canada.

Hoey, J., Wilkins, R., Gagnon, G., and O'Loughlin, J. (1987). L'état de santé des Québécois: un profil par région socio-sanitaire et par département de santé communautaire, in 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, 135-158.

Statistics Canada (1988). Undercoverage rates from the 1986 Reverse Record Check (1986 Census of Canada, User Information Bulletin Number 2), Ottawa: Statistics Canada.

Statistics Canada (1991). Postal Code Conversion File (January 1991 Version), Ottawa: Geography Division, Statistics Canada.

Wilkins, R., Sherman, G. J., and Best, P. A. F. (1991). Birth outcomes and infant mortality by income in Urban Canada, 1986, *Health Reports*, 3, 1, 7-31.

Wilkins, R., Adams, O., and Brancker, A. (1989). Changes in mortality by income in Urban Canada from 1971 to 1986, *Health Reports*, 1, 2, 137-174.

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

SPATIAL AND STATISTICAL APPLICATIONS OF ENVIRONMENTAL GEOCHEMICAL DATA TO HUMAN HEALTH ISSUES

D.R. Boyle[1]

ABSTRACT

The plotting of geochemical data to outline the distribution of elements and geochemical factors known to possess some component (deficiency, toxicity) of risk to human health, can be termed Geochemical 'Sensitivity' Mapping. Such elements or factors can be measured on a variety of media associated directly or indirectly with the human-food-water-air chain. The emphasis in this type of study is in characterizing, both spatially and statistically, areas having anomalously high or low levels of a particular parameter suspected (or proven) to be implicated in the pathology of a given disease.

By 'casting' regional geochemical data for directly imbibed media, such as water, into contour levels related to known dose-response effects, the environmental geochemist can better describe the impact of geochemical data to those officials responsible for health and environmental planning.

KEY WORDS: Dose-Response concepts; Geochemical 'sensitivity' mapping.

1. INTRODUCTION

Geochemistry involves the study of the abundance of elements and the nuclides in the earth, the distribution of the elements among the geochemical phases of the earth, and the laws governing these abundances and distribution relationships. This is accomplished mainly by detailed geochemical analyses of the factors controlling the abundance and distribution of the elements within and between the five geochemical spheres of the earth (lithosphere, pedosphere, hydrosphere, biosphere, atmosphere).

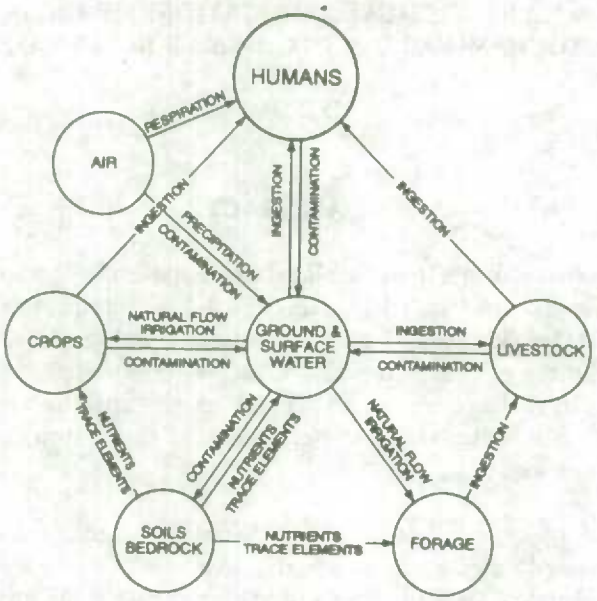
Humans interact with all aspects or spheres of their natural environment (Figure 1) and receive their nutrients from various sources (crops, livestock, waters etc.). It can be seen in Figure 1 that if waters receive an abnormal concentration of a given element or compound, whether through natural processes or contamination, humans can receive such elements or contaminants from a much greater variety of sources than just water. It also follows that if an element is deficient in a particular geochemical environment then humans ,being on the top of the animal chain, will find that such deficiencies may occur in their local food and water supplies.

The geochemical environment can, therefore, play a major role in influencing the geographic distribution of pathological and nutritional problems relating to human health. Examples of such influence are manifest in: a) regional deficiencies of such essential elements as I (goitre, hypothyroidism, cretinism), Fe (anemia), Ca-Mg-Na (cardiovascular disease, hypocalcaemia), Cr (glucose regulation, diabetes), Co (pernicious anemia), Zn (perakeratosis, enzyme related disorders, impaired wound healing), F (dental caries, osteoporosis ?), and Se (cardiomyopathy, cancer prevention); and, b) in regional over-abundance, as the result of certain geochemical processes, of non-essential toxicants such as Cd (renal dysfunction, hypertension, heart disease), Pb (neuropathy, psychotic disorders, hypertension), Hg (neuropathy), Al (Alzheimer's disease ?), As (cancer), and radioactive elements (cancer).

---

[1] D.R. Boyle, Geological Survey of Canada Energy Mines and Resources, 601 Booth Street, Ottawa, Ontario, Canada K1A 0E8.

Figure 1: Human linkage to the environment. All routes eventually lead to humans (after Boyle 1991).



Drinking water composition, which is controlled by certain chemical, climatic and geological factors within the geochemical landscape, may be causally related to such conditions as cardiovascular disease (water hardness), osteoporosis (Ca and F availability), infant mortality (high nitrate, copper or magnesium), neurological and psychotic disorders (acid rain mobilization of Cd, Pb, Hg, Tl, Al), hypertension (high Na, Ba, Cd or Pb) and cancer (high As, Ra or Rn).

When contoured as abundances, ratios or 'factors', regional geochemical data for a variety of media (*e.g.* water, soil, rock, vegetation) can be used both directly (dose-response concepts) and indirectly ('sensitivity mapping')to aid health officials in carrying out:

a) Exposure Assessments:-to outline percentages of a population that may develop diseases/disorders related to element/compound deficiency or toxicity;

b) Risk Assessments:-developing models for predicting possible health effects of particular elements or compounds over similar geological environments which have not been geochemically surveyed,

c) Epidemiological Surveys:-may be directed at determining either associative or causal relationships between elements or compounds in the geochemical landscape and certain diseases (*e.g.* radium in groundwater and bone cancer), or they may be carried out to determine the possible health effects of human perturbations on a given geochemical environment (*e.g.* acid rain and Alzeimer's disease),

d) Rural to Urban Migration Studies:-changes in health status of a rural group moving from a specific geochemical/geological environment to an urban setting,

e) Land Development Studies:- possible health effects of population settlements in geochemically unfavorable areas (*e.g.* land development in the high fluoride regions of the Maritime Carboniferous Sedimentary Basin),

f) Ethnic Migration Studies:-movement of an ethnic group from one geochemical environment to another (*e.g.* Ukrainians from Ukrainian Steppes to Canadian Prairies).

g) Pathways Analyses:-studies of the abundance and changes in speciation of an element and its possible health effects (*e.g.* mobilization, concentration and methylation of mercury through the rock-soil-water-vegetation-human pathway)

h) Background Corrections:-normalization of exposure data by addition or subtraction of natural levels of an environmental parameter (*e.g.* addition of natural background radiation levels to industrial radioactive dose studies).

In all of the studies mentioned above, confounding factors other than those attributable to the geochemical environment must also be taken into account (*e.g.* life style, diet, and smoking habits).

2. STATISTICAL AND SPATIAL METHODS OF GEOCHEMICAL ANALYSIS

Numerous statistical and spatial plotting methods have been applied to the interpretation of geochemical data; most of these have been used in mineral exploration. Statistical and spatial methods of presenting environmental geochemical data are still in their infancy. The spatial presentations used in this paper were generated by the UNIRAS[2] system using the GEOINT interpolation package for gridding of data and the GEOPAK plotting package for final map presentation. This system also contains a kriging interpolation subroutine (KRIGPAK) and an optional user interface (UNIMAP) for interactive mapping. Many other similar systems are in use by environmentalists. Regardless of the plotting package used, the geochemist should ensure that environmental data presented in symbol or contour plotted forms incorporate, depending on application, "Maximum Acceptable Concentration Levels", "Aesthetic Limits", "Dose Response Levels" or any other regulatory limits as distinct levels of a plotting scheme so that health or environmental officials can have easily recognizable and sometimes enforceable limits with which to outline areas of high or deficient elemental abundances and thus populations at risk. This has been the approach taken in the present paper for elements or parameters to which the above mentioned limits may apply. For other elements, suitable contour intervals have been chosen to emphasize areas of both depletion and anomalous enrichment within a region.

3. REGIONAL GEOCHEMICAL 'SENSITIVITY' MAPPING

Generally, before cause and effect can be determined for the etiology of a disease/disorder, associations or correlations between the disease/disorder and suspected health risk factors must be determined. In order to do this, study designs and survey areas need to be established. Also, it is imperative to know the background values of the health risk factors in question. As mentioned earlier, at least some component of the etiology or overall risk associated with many diseases can be attributed to major/trace element deficiencies or excesses in the environment. Methods of mapping the distribution of health risk factors, in this case the distribution of elements in toxic or deficient quantities within the environment, are required in order to give focus to detailed health studies.

Regional geochemical surveys, which portray as contoured levels or symbol plots, the distribution of trace elements or factors known to possess some component (deficiency, toxicity) of risk to human health, can be given the term Geochemical 'Sensitivity' Mapping. Such elements or parameters can be measured on a variety of media associated directly or indirectly with the human-food-water-air chain. The emphasis in this type of study is in outlining populated areas characterized by anomalously high or low levels of a particular parameter suspected (or proven) to be implicated in the pathology of a given disease. Populations at risk in such areas (general or age-sex-ethnic groups), together with adjacent background populations, can be incorporated in an environmental epidemiology sampling frame to investigate the etiology of the disease in question. It will be important, when determining the magnitude of risk to health of a particular parameter, that the pathways by which the element might reach humans and the forms in which it is available to them be clearly understood. Examples of geochemical 'sensitivity' mapping are given below to illustrate the usefulness of this type of survey.

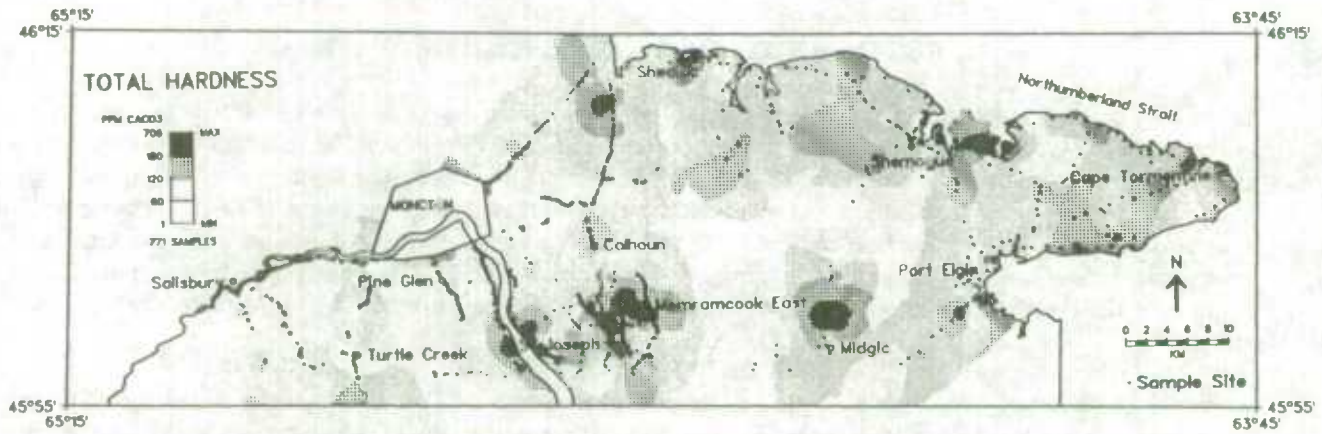
---

[2] European Software Contractors A/S, Norregade, Denmark.

Considerable research on the relationship between cardiovascular disease and drinking water composition has culminated so far in the acceptance that there is a 'water factor(s)' in hard waters that offers some degree of protection from heart disease and that people on soft waters have a slightly higher risk of cardiovascular disease than those on hard waters (Shaper *et al.* 1980; Calabrese *et al.* 1980; Lacey and Shaper 1984). Isolation of this 'water factor(s)' would be one step in preventing heart disease, well behind such factors as smoking and diet, but still important.

The chemical elements most involved in causing or preventing heart disease are Ca, Mg, Na, Ba, Cd, Pb, and Se. Calcium and Mg are essential 'heart building' elements and a considerable percentage of daily intake of these elements can come from drinking water; especially Mg (Hopps and Feder 1986). These two elements are the greatest contributors to water hardness and maps, such as Figure 2 showing water hardness of groundwaters in the Moncton area of New Brunswick based on low (0-60 meq/l), medium (60-120 meq/l) and high (>120 meq/l) degrees of hardness, can be used by epidemiologists to select exposed and non-exposed population groupings for studies of the impact of this parameter on disease outcome. Sodium, Ba, and to a lesser degree, Cd and Pb are 'hypertensive elements'; elevated intakes of one or a combination of these elements can lead to increased blood pressure and possible heart disease. Both the Ba distribution in groundwaters of the Moncton area (A, Figure 3) and, perhaps more effectively, the Na/Ca+Mg ratio for these waters (B, Figure 3) can be used in conjunction with the water hardness distribution (Figure 2) to give a more comprehensive interpretation of the exposure of population groups to certain 'water factors' possibly related to heart disease. Other approaches to this problem, such as the combined effects of Na+Ba+Cd+Pb (normalized data), could be undertaken to give a 'hypertensive index' for drinking waters for population groups under study. Selenium, which is considered to be protective against some forms of heart disease, especially those related to the myocardia (Chen *et al.* 1980), can also be incorporated into an investigative study of trace element intakes and heart disease.

Figure 2: Hardness of groundwaters in the Moncton area, Maritime Carboniferous Basin (unpublished data).



When the atmospheric loading of acids exceeds the base production from the weathering of overburden and bedrock, both the pH and quality of groundwater can be severely affected. Acid groundwaters can have a detrimental effect on human health through greater solvency and mobility of toxic heavy metals (*e.g.* Pb, Cd, As, Cu, Al). These metals can be dissolved from both the lithosphere (weathering) and household plumbing systems (plumbosolvency). Contoured maps of the pH of groundwaters, such as the one presented for part of the Maritimes in Figure 4, can be used to outline areas that might be sensitive to acid rain loadings and, together with population density maps, can be used to determine the percentage of the overall population that might be at risk from greater mobility of toxic elements. Element mobility studies would have to be carried out before a full judgement could be made.

Figure 3: Sodium/calcium + magnesium ratio (A) Barium (B) distributions in groundwaters of the Moncton area, Maritime Carboniferous Bassin (unpublished data).

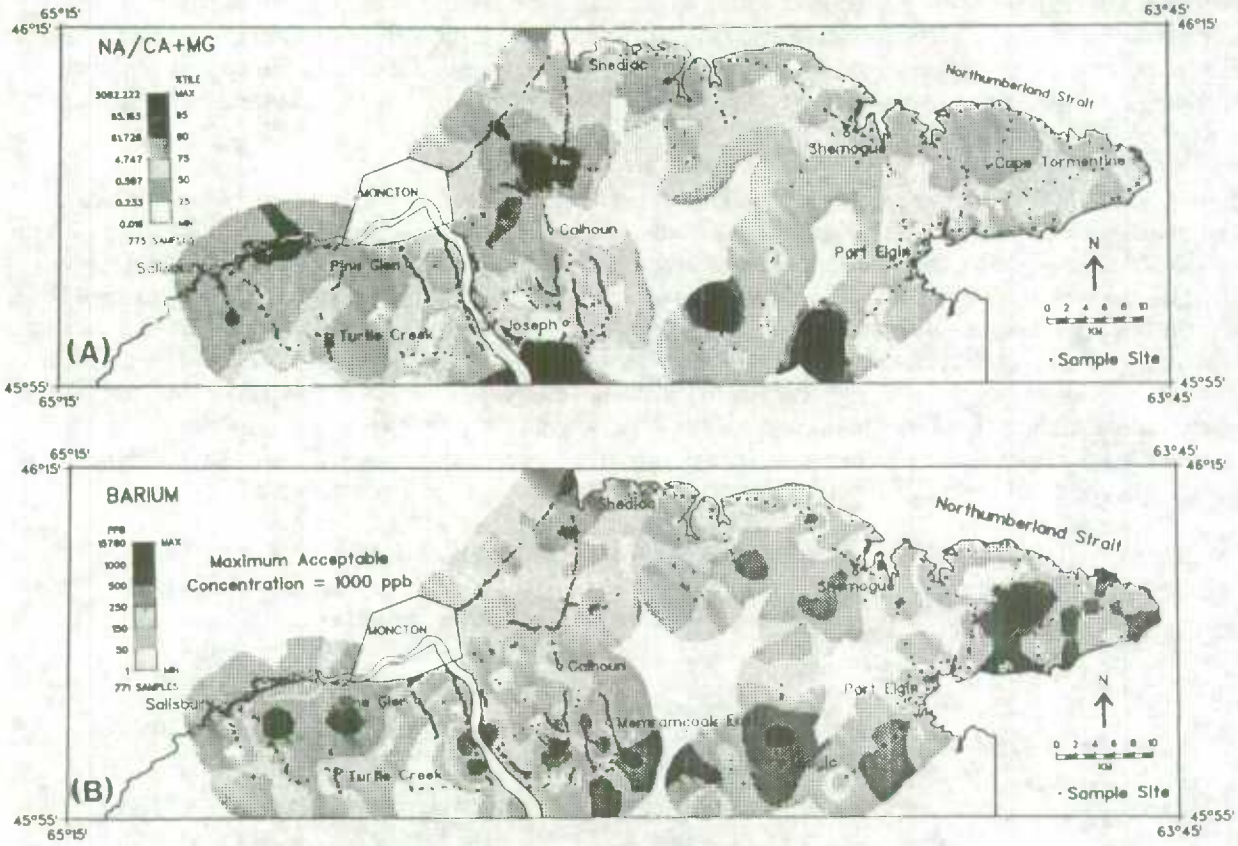
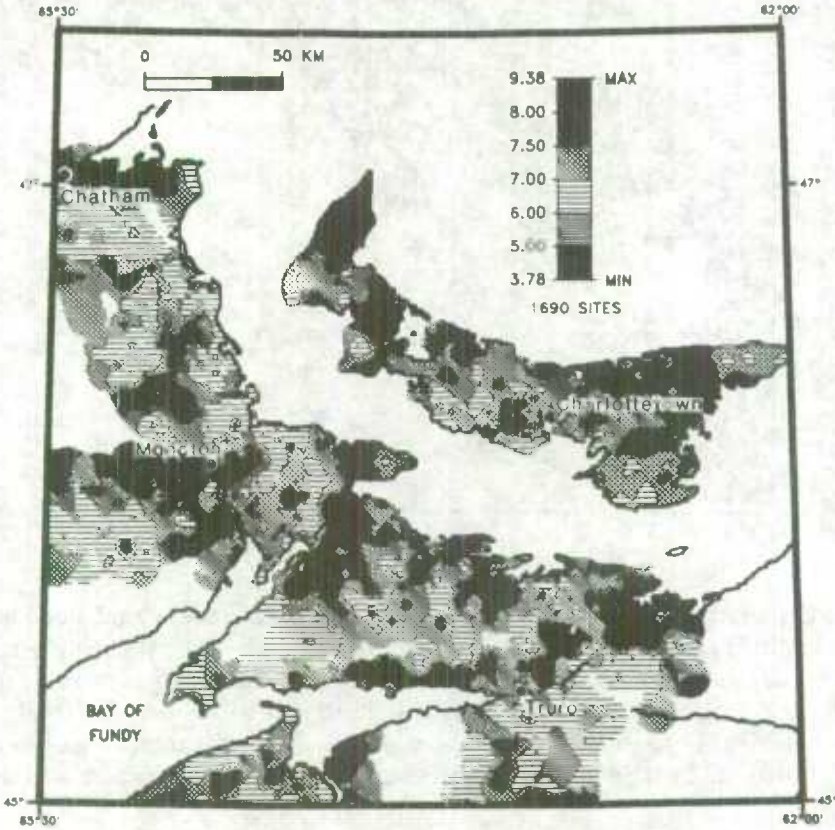


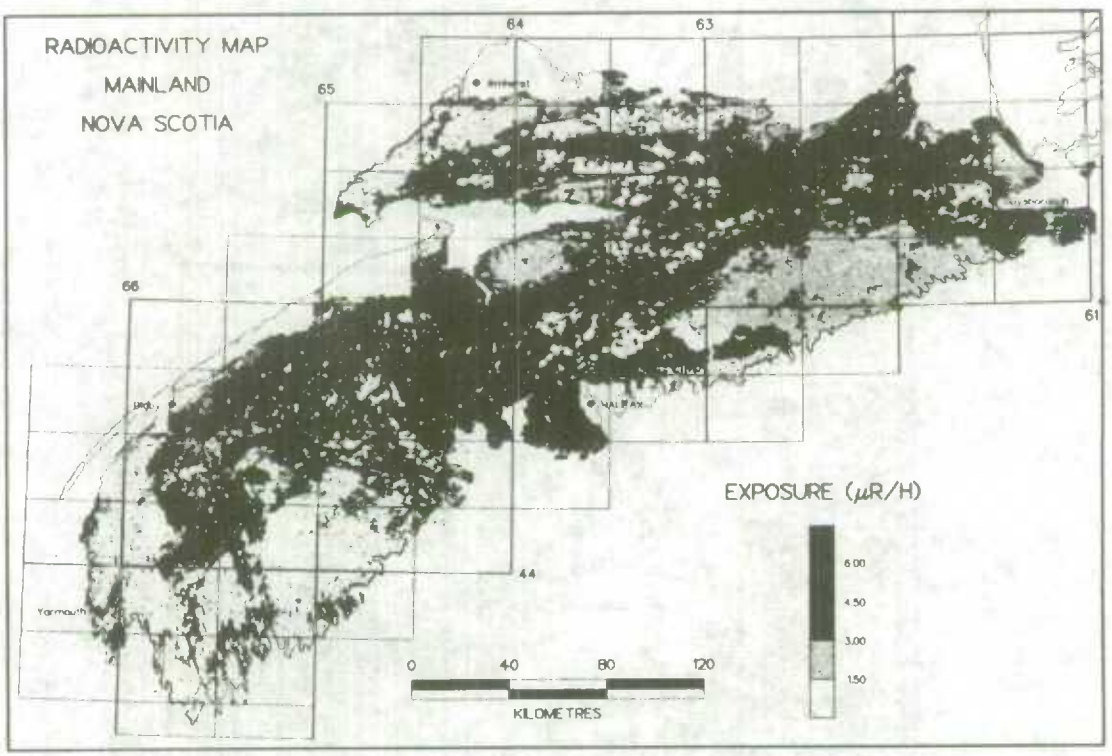
Figure 4: pH of groundwaters in the East Central region of the Maritime Carboniferous Basin (data from Dyck 1976).



Most of the radiation exposure humans receive comes from the natural environment and the vast majority of this is gamma radiation. For this reason, maps showing natural radiation levels are important for determining populations at risk with regard to radiation exposure to the external body and internal inhalation of radioactive components such as radon. The Geological Survey of Canada has carried out airborne gamma radiation surveys over almost 30% of the Canadian landmass. Details of natural background radiation in Canada, together with a methodology for calculating average summer and annual outdoor radiation doses to Canadians, can be found in Grasty *et al.* 1984.

For those health officials and epidemiologists working in the field of radiation and its effects on human health, exposure maps such as that shown in Figure 5 can be used to determine background radiation loadings on a given population group under study or they may be used to select population groups for study. Furthermore, such data can be used by health officials and building code planners to determine whether populations may be at risk from increased household radiation levels and what measures must be used to alleviate such risk. Those health officials carrying out research on radiation exposure in the work place, on either a regional or national basis, can use exposure maps such as Figure 5 to calculate total gamma radiation exposure (natural plus workplace). For example, those people in Nova Scotia living on granitic terrane receive two to four times more gamma radiation per hour than those living on sedimentary rocks (see caption Figure 5); such large differences in natural exposure could seriously affect interpretations given to work place exposure data.

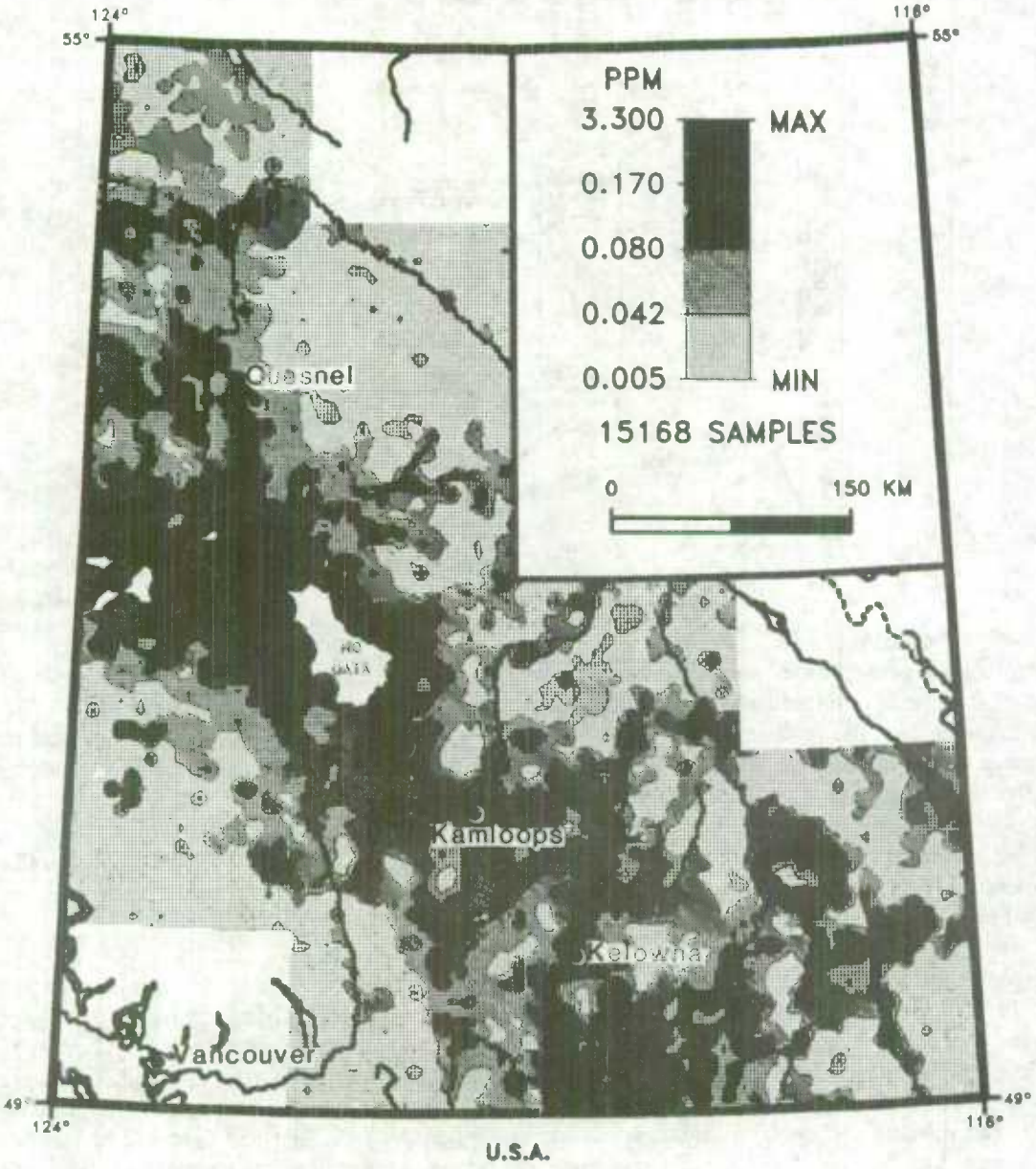
Figure 5: Gamma radiation exposure (uR/H) from bedrock and overburden in mainland Nova Scotia. Areas of high and low exposure correspond generally to regions underlain by granitic and sedimentary rocks, respectively (after Boyle 1991).



The above examples demonstrate a reasonably direct influence that element or radiation loadings might have on human health issues. However, distribution maps of elemental concentrations in various types of surficial geochemical media not ingested by humans can also act as 'sensitive' indicators of possible toxic or deficiency related diseases. For example, the distribution of fluoride in stream waters for most of the southern half of British Columbia (Figure 6) shows a number of regions with high fluoride contents. Because the surface water geochemical characteristics of fluoride are similar to those in shallow groundwater regimes, surface water

concentration patterns for this element are generally highly reproducible in groundwaters. In most cases, shallow groundwaters used for drinking purposes are more concentrated in fluoride than associated surface waters by a factor of about ten. It is possible, therefore, to apply such a factor to the data in Figure 6 and thus outline regions where fluoride toxicity or deficiency related health problems may occur. When used with population density maps, surveys such as this can be effectively used in Exposure Assessment studies. The above mentioned approach to interpreting surface water fluoride data can be compared to that described below for dose-response interpretations of fluoride in rural groundwater drinking supplies.

Figure 6: Distribution of fluoride in stream waters of South Central British Columbia (fluoride data from Geological Survey of Canada National Geochemical Reconnaissance data base).

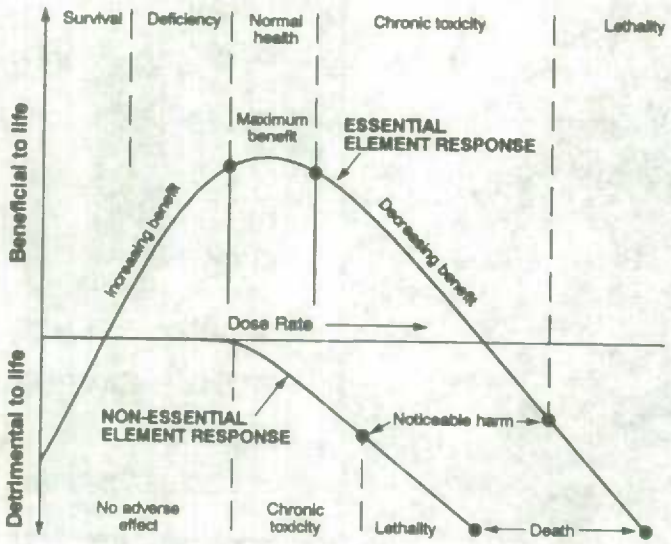


The examples given above are but a few of the geochemical 'sensitivity' maps that could be presented. Regional geochemical data of this nature could also be used to map additive and multiplicative factors considered to be a risk to human health.

4. APPLICATION OF DOSE-RESPONSE CONCEPTS TO GEOCHEMICAL DATA

Dose-response curves (Figure 7) describe the beneficial and detrimental health attributes affecting human populations based on dose per unit time (usually 24 hours) over a specific duration. The beneficial attributes may relate to general state of health (low morbidity) or to relative freedom from certain diseases/disorders (*e.g.* dental caries, goitre, cardiovascular disease). The detrimental attributes are manifest in diseases and disorders related to elemental deficiency or toxicity. Such curves may be used to describe the potential state of health of human populations with respect to certain ingested or absorbed concentration levels of an element or compound (essential and non-essential).

Figure 7: schematic representation of dose-response curves for essential and non-essential elements (after Boyle 1991)



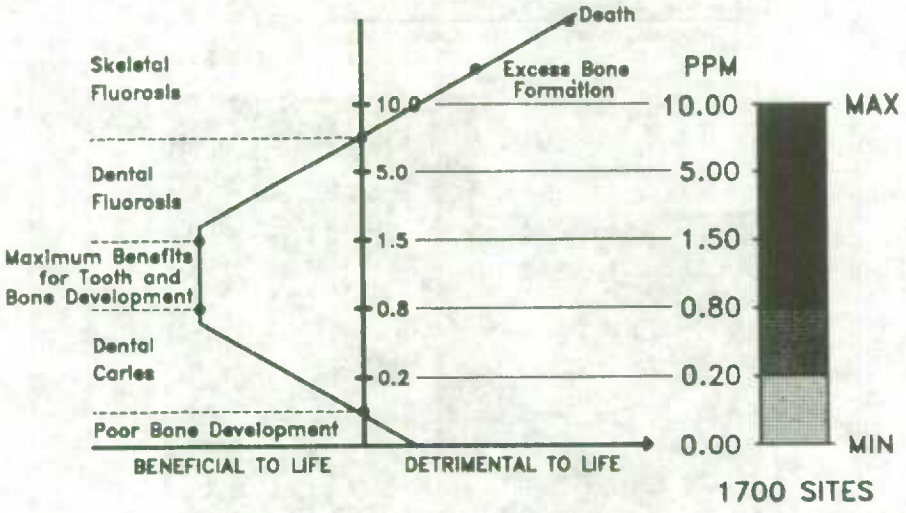
The interpretation of data based on dose-response concepts must take into account the model(s) on which the dose-response curve is generated and any assumptions made in doing so. Curves can be based on total or partial ingested amounts of a particular element or compound per unit time. Partial ingested levels (*e.g.* fluoride in drinking water) may be used for interpretive purposes where an average daily dietary intake from other sources can be determined (*e.g.* food and beverages).

By 'casting' regional geochemical data into contour levels relating to known dose-response effects, the environmental geochemist can better describe the impact of geochemical data to those officials responsible for health and environmental planning.

An example of the application of dose-response concepts to regional geochemical data can be seen in Figures 8 and 9. In both these areas domestic well waters were collected at a density of approximately one sample per 10 km2 and analyzed for fluoride and a number of other elements and water parameters (Dyck *et al.* 1976 and Dyck 1980). The fluoride data are represented as a patterned bar scale corresponding to a schematic dose-response curve showing potential health outcomes at various concentration intervals. The beneficial plateau represents the fluoride concentration interval considered to be adequate for good dental (and possibly bone) development at the latitude of these two areas based on a consumption of 1.5 l of water per day. It must be emphasized when interpreting these data that dietary information on the dose rate intake of fluoride must be either measured or assumed. Generally an average dietary intake would be applied to the population group for which the contoured dose-response maps for fluoride in waters is to be applied (*e.g.* infants, children to age 15, elderly people, certain predisposed groups). When combined with population density maps, the dose-response interpretations in Figures 8 and 9 can be used to determine the possible impact that fluoride deficiencies and excesses will have on the health status of specific regional populations. Areas where the total fluoride intake

(water, food, dental supplements) is considered to depart severely from the norm (high or low) may be considered as regions for more detailed health studies. The dose-response interpretation for fluoride in the Maritime groundwaters (Figure 8) displays large regions characterized by very low fluoride concentrations and a number of distinct areas where the fluoride concentrations are very high; there are very few wells within the beneficial interval. These data may be contrasted with those of the southern Saskatchewan region (Figure 9) where most of the fluoride concentrations are at, or just below, the lower limit of the beneficial interval with very few areas displaying high concentrations. For both these areas, the majority of the population (100% for P.E.I.) are on groundwater well supplies.

Figure 8: Dose-response map for fluoride in groundwaters of East Central region of the Maritime Carboniferous Basin (fluoride data from Dyck 1976).



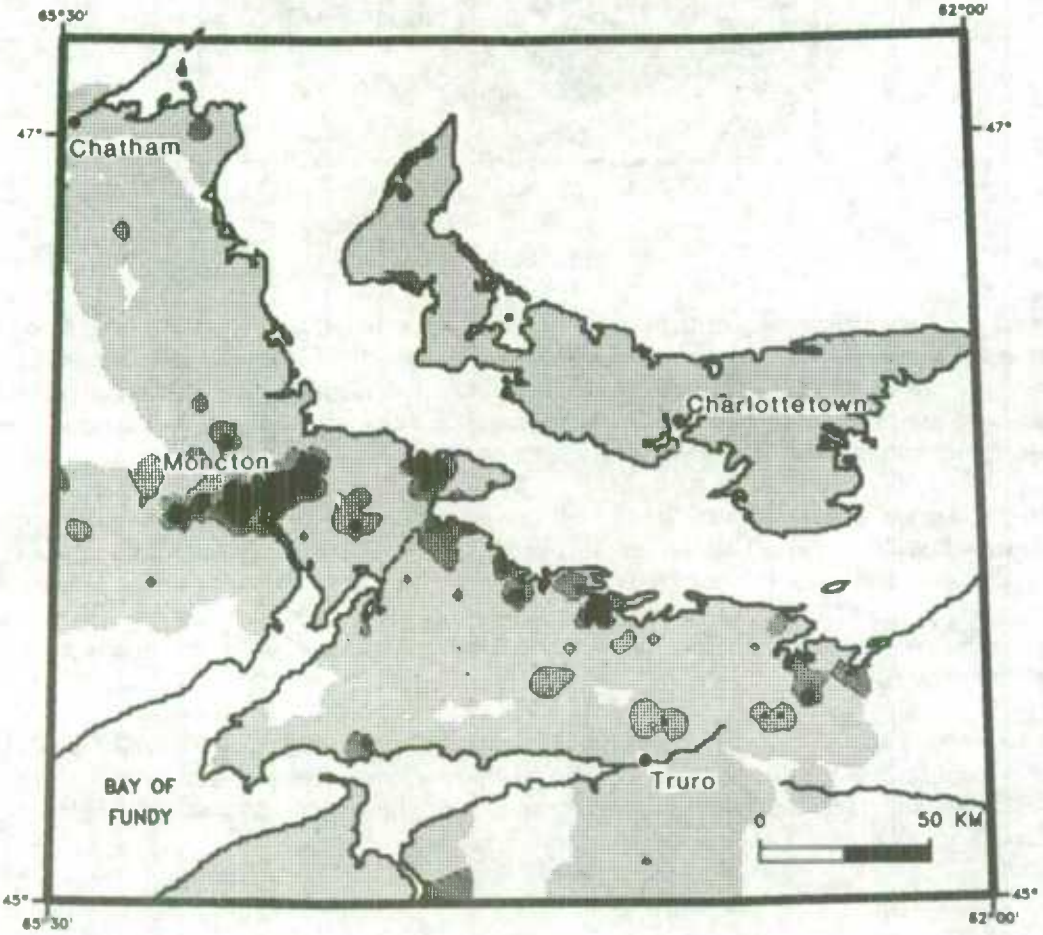
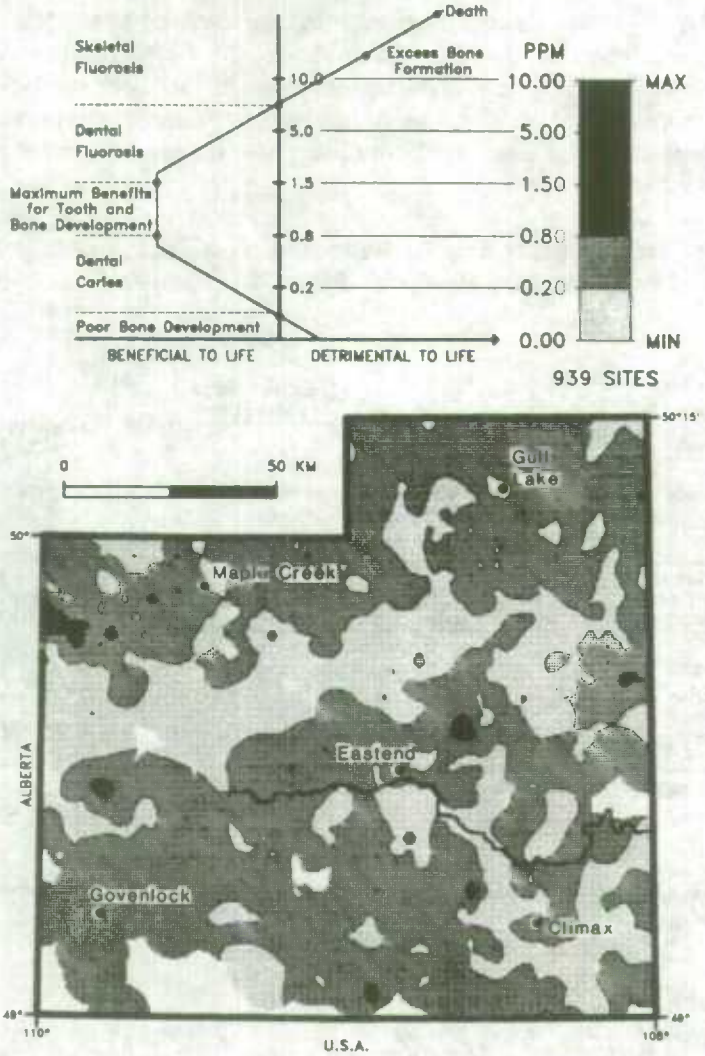


Figure 9: Dose-response map for fluoride in groundwaters of southwestern Saskatchewan (fluoride data from Dyck 1980).



5. DISCUSSION

There is a great need for environmentalists to quantify those parameters that may have a bearing on human health outcome and to present such parameters in a form in which health officials can pass judgment as to the degree of risk to human health each parameter or group of parameters might display. The overall distributions of these parameters will of course be important in determining the size of population at risk and the magnitude of the risk associated with element/compound deficiencies or toxic excesses.

Research in medical geography and medical geology has shown that no two geochemical environments are equal with regard to overall health status or risk of a specific disease. This has been emphasized in studies of high and low longevity (National Research Council 1981), dental caries (McClure 1970), and cardiovascular disease (National Research Council 1979; Calabrese *et al.* 1980; Shaper *et al.* 1980), to name a few. It is incumbent upon the scientific community, therefore, to characterize those environments that fail to give maximum quality of health to their inhabitants.

The examples described above of the applications of Geochemical 'Sensitivity' Mapping and Dose-Response Concepts serve to emphasize the fact that geochemistry can play a strong role in characterizing environments of high or low disease incidence, thus making it possible to compare health risk factors and formulate sampling designs for epidemiological studies.

6. ACKNOWLEDGEMENTS

I would like to thank the computer cartography unit of the Geological Survey of Canada for production of the schematic figures used in this paper. W. Spirito and K. Ford of the Mineral Resources Division, Geological Survey of Canada are thanked for producing the geochemical and radiometric maps respectively. Comments from C.E. Dunn of the Geological Survey of Canada on an initial draft of this paper are greatly appreciated. Opinions expressed in this paper are solely those of the author and do not represent any stated policies of the Geological Survey of Canada.

REFERENCES

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., and 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., and Ge, K. (1980). Relation of selenium deficiency to the occurrences of keshan disease. *Selenium in biology and medicine*, J.E. Spallholz *et al.* (eds.), AVI Publ., Westport, Conn., 171-175.

Dyck, W., Garrison, E.W., Godoi, H.O., and 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 and 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., and Holman, P.B. (1984). Natural background radiation in Canada. Geological Survey of Canada Bulletin 360, 39p.

Hopps, H.C., and 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., and 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., and Pocock, S.J. (1980). The British regional heart study: Cardiovascular mortality and water quality. *Jour. Environ. Pathology and Toxicology*, 4-2, 3, 89-111.

FIGURE CAPTIONS

1. Human linkage to the environment. All routes eventually lead to humans (after Boyle 1991).

2. Hardness of groundwaters in the Moncton area, Maritime Carboniferous Basin (unpublished data).

3. Sodium/calcium + magnesium ratio (A) Barium (B) distributions in groundwaters of the Moncton area, Maritime Carboniferous Basin (unpublished data).

4. pH of groundwaters in the East Central region of the Maritime Carboniferous Basin (data from Dyck 1976)

5. Gamma radiation exposure (uR/H) from bedrock and overburden in mainland Nova Scotia. Areas of high and low exposure correspond generally to regions underlain by granitic and sedimentary rocks, respectively (after Boyle 1991).

6. Distribution of fluoride in stream waters of South Central British Columbia (fluoride data from Geological Survey of Canada National Geochemical Reconnaissance data base).

7. Schematic representation of dose-response curves for essential and non-essential elements (after Boyle 1991).

8. Dose-response map for fluoride in groundwaters of East Central region of the Maritime Carboniferous Basin (fluoride data from Dyck 1976).

9. Dose-response map for fluoride in groundwaters of southwestern Saskatchewan (fluoride data from Dyck 1980).

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

THE IMPACT OF GEOGRAPHIC DISTORTION DUE TO THE HEADQUARTERS RULE

R. Burroughs[1]

ABSTRACT

The Census of Agriculture collects information about the operation of a farm with reference only to the location of the farm headquarters. Since an increasing number of farms operate over several separate tracts of land, a certain amount of distortion is introduced into the results particularly for land based variables. This research activity attempts to measure the distortion in two different locations; one in Prince Edward Island and the other near Swift Current, Saskatchewan.

KEY WORDS: Headquarters rule; Positive distorsion; Negative distorsion; Total farm area.

1. INTRODUCTION

The most frequent type of error encountered in sub-provincial Census of Agriculture data can be traced to a convention referred to as the headquarters rule. There is likely to be some error of this type in any number tabulated below the province level. Despite the frequency, few users are aware of it nor are they likely discover any evidence of it without careful scrutiny of the data.

The careful user of the published data might notice in a couple of Saskatchewan census divisions that the total area of farms is slightly larger than the geographical area of the entire census division. This will be brought to the attention of Census of Agriculture staff perhaps twice during the five year census cycle. This paper will explain how this error is introduced into the data, outline the factors that influence its impact and examine the effects on the data for two study areas using data from the 1986 Census of Agriculture.

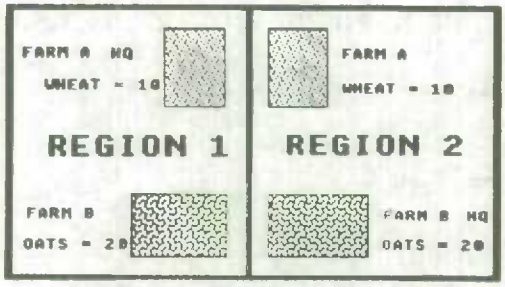
2. THE HEADQUARTERS RULE

It is common for farms in Canada to be composed of more than one parcel of land. The data collected on the questionnaire refers to the whole operation without allocating any of the variables among the various parcels of land. The only information collected with respect to each parcel of land is the legal description and the total area. The fundamental problem is posed when a geographic reference point is attached to this data. How are the data to be allocated among the parcels of land? In practise, the convention is to attribute all data to the parcel of land designated as the headquarters by the operator regardless of the location of any other parcels of land in the operation.

While this approach avoids the complexities of allocating data among the parcels of land, it introduces errors into the resulting tabulations. Figure 1 illustrates how this happens. Consider a simplified agricultural economy consisting of two farms (A & B), two regions (1 & 2), and two crops (wheat & oats). Farm A, with its headquarters in Region 1, operates two parcels of land, one in each region, and each parcel has an area of 10 acres all sown to wheat. Farm B has its headquarters in Region 2 and operates a 20 acre parcel of land in each region both sown to oats. If the land and crop areas were derived by observation of the diagram (land based approach as most users assume) then each region would have a total farm area of 30 acres broken down into

[1] R. Burroughs, Statistics Canada, Agriculture Division, 12-C2, Jean Talon Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

Figure 1



	TOTAL	BLÉ	AVOINE
REGION 1			
AREA LAND BASED	30	10	20
AREA HQ BASED	20	20	0
DIFFERENCE	-10	+10	-20
REGION 2			
AREA LAND BASED	30	10	20
AREA HQ BASED	40	0	40
DIFFERENCE	+10	-10	+20

10 acres of wheat and 20 acres of oats. If the land and crop areas are derived using the headquarters convention, somewhat different results are obtained. Since only Farm A has a headquarters in Region 1, then the only areas that can be attributed to Region 1 are the two 10 acre wheat parcels. Similarly, in Region 2, only the two 20 acre parcels of oats from Farm B can be attributed. The differences between the land based value and the headquarters based value are termed the distortion due to the headquarters rule. Note that the sum of the distortions over the two regions is zero for each variable (*i.e.* land, wheat and oats).

3. FACTORS THAT INFLUENCE THE AMOUNT OF DISTORTION

The following are the factors that determine the amount of distortion which will occur.

Boundary Location

The distortion illustrated in Diagram 1 resulted from the situation where the parcels of land within the farm operation were separated by the boundary between the two regions. If the boundary does not separate parcels of land within the same farm operation, then no distortion will occur. The Manitoba/Ontario border is an example where no distortion is likely to occur, while the Manitoba/Saskatchewan border is an example where distortion is likely to occur.

The Distance Between Parcels

The greater the distances between the headquarters parcel and the other parcels, the greater the risk that a boundary will separate them.

The Number of Parcels

The greater the number of parcels within the operation, the greater the risk that one or more will be separated by a boundary.

The Length of the Boundary

The longer the boundary, the greater the risk that it will separate parcels of land from their headquarters.

The Value of the Variable Associated with the Separated Parts

The amount of distortion in a given variable increases with the value of the variable associated with the separated parcel. If oats is growing on a separated parcel, then the greater the area of oats in that parcel, the greater the distortion of the oats variable. If there is no oats growing on the separated parcel, then the distortion of the oats variable will be zero.

4. CASE STUDY I - CHESTER, SASKATCHEWAN

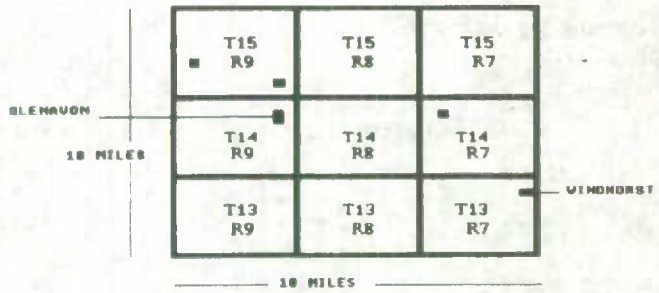
Chester is a rural municipality in south-eastern Saskatchewan. In the 1986 Census of Agriculture, the farm area published (under the headquarters rule) was slightly higher than the geographical area within its boundaries. For this reason it was selected as the starting point of the investigation.

Its legal description declares Chester to be Saskatchewan Rural Municipality #125 consisting of nine townships (Townships 13,14 & 15, Ranges 7,8 & 9 west) of the second meridian. Its physical dimensions are exactly 18 miles by 18 miles.

The approach to measuring the distortion in the total farm area requires two measurements. The first is termed the positive distortion. This is the area of parcels of land which are not located in Chester but are associated with farm operations with a headquarters in Chester. This measurement was obtained by recording the area of all such parcels as reported on the questionnaires with headquarters in Chester. The second measurement is termed the negative distortion and refers to the area of parcels of land which are located in Chester but are associated with farm operations with a headquarters located outside of Chester. This measurement could also have been taken from the questionnaires, however it would require looking at more than a thousand of them in all the neighbouring municipalities. It was much simpler to derive it from information already at hand. The method is explained in the Appendix.

Figure 2 presents the results. The area of Chester was 207,360 acres. It contained the headquarters of 220 farms with a farm area totalling 207,958 acres. This is the result of a positive distortion of 22,326 acres and a negative distortion of 20,608 acres on a land based area of farms of 206,240 area. The important things to notice are:

Figure 2



MAP OF CHESTER #125

	AREA ACRES	FARMS REPORTING
AREA OF LAND	207,360	-
AREA OF FARMS WITH HQ IN #125	207,958	220
AREA ATTRIBUTED BUT NOT IN #125	22,326	44
AREA IN #125 BUT NOT ATTRIBUTED	20,608	N/A
LAND BASED AREA OF FARMS	206 240	-
PERCENT CHANGE	1	

1) The area of farms is greater than the area of the municipality because the positive distortion is not completely offset by the negative distortion and the fact that virtually the entire area of the municipality is agricultural.

2) While the net impact of the distortion on farm area is relatively small, the size of the positive and negative distortions (more than 10% each) is significant.

5. CASE STUDY II - LOT 19, PRINCE EDWARD ISLAND

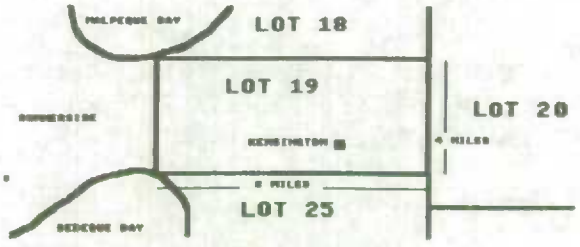
Lot 19 in Prince Edward Island represents a contrast to the situation in Chester. The type of agriculture is different. The cadastral organization is different. The farms are generally smaller in area. Would the distortions be as large in these circumstances?

The method used was similar to that described in Chester except that the negative distortion was measured using the questionnaires rather than the equations. The amount of work was considerably less than the case in Chester.

Prince Edward Island municipal data is published by lots. These lots are considerably smaller than RM's in Saskatchewan. Lot 19 is approximately 8 miles by 4 miles and is about 60% agricultural. It is bounded by the city of Summerside on the west and three other rural lots on the north, east and south.

The results of the study are presented in Figure 3. The total area of Lot 19 is estimated at 20,352 acres. It contains the headquarters of 62 farms with a farm area of 12,598 acres. The positive distortion was calculated at 3,355 acres and the negative distortion came out to 3,336 acres.

Figure 3



MAP OF LOT 19

	AREA ACRES	FARMS REPORTING
AREA OF LAND	20, 352	-
AREA OF FARMS WITH HQ IN LOT 19	12, 598	62
AREA ATTRIBUTED BUT NOT IN LOT 19	3,355	13
AREA IN LOT 19 BUT NOT ATTRIBUTED	3,336	19
LAND BASED AREA OF FARMS	12,579	-
PERCENT CHANGE	0	

Figure 4

MOVE THE HEADQUARTERS OF THREE FARMS FROM LOT TO NEIGHBOURING LOTS .....

	AREA ACRES	FARMS REPORTING
AREA OF LAND	20,352	-
AREA OF FARMS WITH HQ IN LOT 19	9,064	59
AREA ATTRIBUTED BUT NOT IN LOT 19	774	10
AREA IN LOT 19 BUT NOT ATTRIBUTED	4,289	22
LAND BASED AREA OF FARMS	12,579	-
PRECENT CHANGE	-28	

Since the number of farms associated with the distortions was relatively small, a deeper investigation was performed. It revealed that three farm operations have a large impact. These operations had large holdings of land in Lot 19 and in surrounding lots as well. Their headquarters were located in Lot 19. Figure 4 presents the land area data for Lot 19 as if the operators of these farms had chosen a parcel of land in another lot as the headquarters. The area of farms under the headquarters rule drops by 28%. The positive and negative distortions become considerably unbalanced.

Up to this point, the total farm area has been the only variable discussed as it is the only information on the questionnaire for each parcel. The study of distortion in other variables would require small area land-based information from another source. The satellite imagery in the Remote Sensing Unit of the Crops Section was suitable and readily available.

They had a satellite image of Prince Edward Island taken in the growing season of 1986. The unit also had considerable experience estimating the area of potatoes from these images in recent years. They outlined the boundaries of each potato field in Lot 19 and calculated the area. The result was a land-based estimate of potato area of 3,568 acres. The comparable headquarters-based value produced by the Census of Agriculture was 3,764 acres.

The important things to note from this case study are:

1) As in the case of Chester, the net distortion of farm area was relatively small. Likewise the positive and negative distortions were of significant proportions, and more than double the proportions calculated in Chester.

2) The impact of the three large farm operations in Lot 19 was enough to drastically alter the area of farms and the balance of the positive and negative distortions.

3) The 196 acre difference (5 %) between the land-based and headquarters-based estimates for the area of potatoes demonstrates that a measure of distortion for one variable is not a reliable indicator of the distortion of another variable.

6. FINDINGS

The findings of this research are as follows:

1) The amount of distortion at the small area level can be significant although the net impact may appear to be negligible.
2) Large operations can produce large distortions.
3) The distortion of one variable is not necessarily related to the distortion of another.
4) Distortion is a function of several factors not all of which are easily measured.

Some users may regard the small net distortions observed in this paper as of little consequence. Others may point to the relative size of the positive and negative distortions and consider them cause for serious concern. The recommended approach is to assess each situation separately.

For example, there is no need to be concerned for provincial level data. The risk tends to increase with each successive level of geographic disaggregation. Similarly, for variables which are commonly reported in an area, the tendency for positive and negative distortions to balance is to be expected. Also, in areas where farms are smaller than average in size, the tendency for farms to consist of multiple parcels is reduced. The user's assessment of the risk and its possible impact in each application should govern the use of the data.

APPENDIX

Method for Estimating Negative Distortion in Case Study 1

Negative distortion can be estimated in the areas covered by township plans in the Prairie Provinces using two equations as follows:

Equation 1

TLA = NALA + AGLA

where TLA = total land area
NALA = non-agricultural area
AGLA = agricultural area

Equation 2

AGLA = AGLA(HQ) - PD + NG

where AGLA(HQ) = agricultural area under the headquarters rule
PD = positive distortion
ND = negative distortion

Substituting the right side of Equation 2 for AGLA in Equation 1 and solving for ND gives:

Equation 3

ND = TLA - NALA - AGLA(HQ) + PD

Data sources for the right side components of Equation 3 are:

- TLA - legal description of the municipality
- NALA - census enumerator's township plan
- AGLA(HQ) - published farm area for the municipality
- PD - recorded from the microfilm records of the questionnaires

SESSION 9

Geographic Frameworks for Statistical Data

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

ALTERNATIVE FRAMEWORKS FOR RURAL DATA

A.M. Fuller, D. Cook and J.G. FitzSimons[1]

ABSTRACT

Maintenance of clearly defined, commonly held concepts of rurality are becoming increasingly difficult in the 'Arena Society'. Although spatial units and configurations remain relatively fixed, emergent human behaviours are dynamic and are best described as differences in 'lifestyles'. Three options to the redefinition of rurality are explored based upon (a) the CMA and CA concept (b) adjustment of existing population and density definitions and (c) a multivariate classification approach.

KEY WORDS: Rural; Arena society; Multivariate classification.

1. INTRODUCTION

"The transportation industry which made it possible to **work in** and **live out** has now enabled people to work anywhere and live nowhere."

Gilbert 1960

'Rural' must go. This cry has long been heard from the research community, yet the term, the old definitions and the concept itself remain deeply embedded in our psyche. As a general descriptive term it has obvious meaning and value. It conveys to most people conditions associated with countryside, small settlements, greenness and remoteness. These may be seen as land, people, ecology and space related. Yet the precise degree, combination or nature of these basic characteristics vary enormously according to objective reality and individual experience. Most importantly these conditions and perceptions change over time such that maintaining clearly defined, commonly-held concepts of rurality becomes increasingly difficult.

For public agencies which have to develop policy and administer programs, the labels which describe differences in the circumstances of people is a constant concern. Such agencies are driven by demands of equity and justice, special needs, fiscal responsibility and political efficacy. Policy analysts and researchers who provide public agencies with information on needs, costs and benefits, are required to use data that is publicly available and is seen to serve the common good. All these demands and constraints on public policy and program development place enormous emphasis on the requirement for good, accurate data that reflects reality. Popular images of 'rural' are not sufficient to inform policy in a consistent and meaningful way, although they are important for society to uphold.

The essential characteristic of 'rural' is that it is a geographical concept. It conveys a sense of space, whether measured in terms of low density of people, houses or activities. It invariably contrasts with high density urban or metropolitan centres in that it looks different. However, a major debate has evolved as to whether the people who occupy so-called 'rural' space are different, behave differently or have a different construction of reality. It is our belief that rural areas not only look different but that people have different lifestyles based on their environment and their values. The important thing however is that there are many different types of rural space and probably many different rural lifestyles, such that the simplistic comparison of rural with urban is insufficient and often misleading.

---

[1] A.M. Fuller, D. Cook and J.G. FitzSimons, University School of Rural Planning and Development, University of Guelph, Guelph, Ontario, Canada N1G 2W1.

It is the purpose of this paper to trace the growing disutility of the term 'rural' as used by Statistics Canada and to establish the need for better constructs based on contemporary reality and future probabilities. It is therefore largely a conceptual paper with only passing reference to data manipulation to provide clues as to the prospects and problems of changing definitions, measurements and units of analysis. In this way, it is hoped to stimulate further examination of the problem and to provide some ideas that might lead to more informed research and debate on what is 'rural'.

2. THE CHANGING NATURE OF 'RURAL'

2.1 'Rural' Must Go?

Dissatisfaction with the term rural is evident in the literature. Hoggart (1990), speaking mainly from the British experience makes the general point that,

> " . . . undifferentiated use of 'rural' in a research context is detrimental to the advancement of social science."

He goes on to observe the distinction between "'rural', meaning a particular kind of geographical milieu, and 'rurality' which refers to a particular behaviour style associated with such areas." As both these concepts are also vague and contain within them large variations in types and, as researchers acknowledging this, still use existing definitions and data because of convenience, he advocates "Lets do away with 'rural'."

The growing plethora of seminars on the future of rural areas in a period of global restructuring illustrates the paucity of suitable data to verify the emerging trends of rural change and dynamism. The seminar, Agriculture and Beyond, Rural Economic Development, held in the USA in 1987 produced international concern about measuring rural dynamism, beyond agriculture (Castle, Newby, Summers, de Janvrey and Deavers).

An international seminar in Scotland on Rural Policy Issues (1991) produced similar observations with regard to the difficulty of measuring the new dynamics of rural areas. The Agricultural and Rural Restructuring Group (ARRG) Seminar on Sustainable Rural Communities in Saskatoon, 1989 and the ARRG Preconference to the Agricultural Economics and Farm Management Society Meetings in Vancouver, 1990, echoed the data problem, particularly the papers by Fuller, Ehrensaft and Gertler; Ehrensaft and Freshwater: and by Fuller, Bollman and Ahearn. References to the dilemmas of definition were made at the Aspen Institute International symposium on Economic Change, Policies, Strategies and Research Issues, in 1990 by Bonnen (USA), MacDowell (USA), Capellin (Italy) and Fuller (Canada). This culminated in 1990, with the Statistics Canada conference on Rural and Small Town Canada, which illustrated through the papers presented both the constraints in the data as well as the paucity of useful constructs with which to identify 'rural'. In 1991, two further conferences have confirmed this view, among European rural development specialists who met in Galway (*e.g.*, Grohn, Finland; Henrichsmeyer, Germany) and rural geographers from the U.K., Canada and the USA who are concerned with rural restructuring who met in the U.K. in August (*e.g.* Munton, Hart and Bryant).

2.2 'Rural' Ain't What It Used To Be

What emerges from this debate is that 'rural' ain't what it used to be. Definitions that may have served well in the past have become redundant over time and because they are maintained for convenience and continuity, have actually led to poor research. A simple 3-stage evolution of 'rurality' can be used to illustrate the dynamics of change in Western Industrial societies over the last one hundred and fifty years.

The Short Distance Society is based on the primacy of primary economies. It captures the old reality where one settlement served to focus most of the activity of its surrounding hinterland. It would be a relatively short distance by horse drawn conveyance to the centre for goods, services and institutional needs such as church and school. The essential dynamic is centripetal, the economy is resource dependent and social organization relatively structured and closed. The unity of space and function in the 'short distance society' is high.

The Industrial Society depicts the broadening of the interactive space, but the focus remains on the central community which becomes industrial in its mode of organization and function, even when supplying farm needs and processing farm outputs. Social organization remains community-based despite the growing 'contractual' form of economic and social relations. Technology is the main motor of change and net labour out-migration characterises this phase in most rural systems. The single resource based community (the single industry town) which brings together the short-distance society and the industrial society, is an example.

The Arena Society reflects the emerging rural reality. The spatial context for activity widens appreciably to include several trade centres for the consumption of personal and household goods and services, as well as for socialisation. The economy may now be linked to the international production and capital markets, with a high segmentation of the labour market. Distance-shrinking technology overcomes isolation in terms of news and media, but the physical distances and the effects of space remain and are still made visible as high transportation costs. Great personal mobility based on the motor vehicle is a feature of the local multicommunity system.

It is essential to recognize that this 3-stage construct developed originally by Persson (1991) is a simplification of how rurality has evolved over time. Importantly, all three 'rural' societies can exist in one region or nation. The concept recognizes the roots of our present rurality, elements of which still remain, while identifying some of the essential features of the emerging reality. The sense of space and the physical configuration of landscape and the infrastructures such as road and settlement patterns remain largely the same, but the economic and social reality of human activity has altered considerably. **This suggests a fundamental shift in scale between spatial units and socio-economic functions.**

2.3 If 'Rural' Ain't What It Used To Be, Then What Is It?

The emerging rural reality is a complex mix of interrelationships that take place within and between the infrastructures that were laid down for the short distance societies in the nineteenth century. The individual need to identify with one place (home and community), to interact with several places (the multicommunity system) to achieve the quality of life expected today, and to be aware of global news and views in the arena society is the emerging norm. It is in this context that new and rural labour markets are emerging, high forms of mobility (vulnerable to fuel crises) become prominent and economies link or de-link with international markets. Although the spatial units and configurations remain relatively fixed, the emergent human behaviours are dynamic and may best be described as differences in 'lifestyles'.

3. REDEFINING 'RURAL'

In searching the literature for ideas which may be of value in reformulating the concept of 'rural', the findings are disappointing. Only four authors outside of Statistics Canada appear to have something to say on the rurality question in terms of measurement.

Paul Cloke developed a Rurality Index for England and Wales in the mid-1970's (Cloke 1977). It was based on a multi-variate analysis of factors which produced a gradient of values by which to map rurality. Updated in 1986, it shows the tendency for rurality and deprivation to increase in remote and upland areas such as Central Wales (Cloke and Edwards 1986).

Marvel Lang assessed alternative approaches to redefining urban and rural for the U.S. Census of Population in 1986. He observes that settlement patterns and socio-cultural lifestyles are the two changes of note in the American population and observes that " . . . one's residence in a rural environment no longer automatically typifies a rural lifestyle." Although he advocates using a household-aggregation approach, he recognizes that it does not account for the socio-cultural aspects of the population and we are left with another example of advocacy without a satisfactory means of measurement.

Beal on the other hand is very specific about measurement and divides geographical space in the U.S.A. into units based on population numbers (Beal 1978). He argues that there is a continuum from rural to metropolitan and that a division into ten categories captures the essential differences between rural and urban areas. It has the merit of being a relatively simple approach with a sound internal logic based on the continuum idea.

However, in testing the Beal codes for Canada, Ehrensaft found the population cut-offs used by Beal to be unsuitable in the Canadian context, although the concept of distance from major centres is a useful one (Ehrensaft and Beaman 1991). Ehrensaft in attempting to use the Beal codes in Canada has added an eleventh code for northern and native environments.

3.1 Measurement in the Public Domain

Two requirements dominate the variety of considerations that need to be taken into account when reformulating statistical and geographical concepts. One is the need to make new definitions consistent with the most valid of old definitions such that temporal continuity can be maintained. An important use of statistical record keeping is to assess societal development. Such analysis would be impossible if new definitions which suit only the current era are invented for every census period. One solution is to be able to adjust concepts and definitions as new phenomena of social importance arise by adapting them to the constructs that have already been established and used.

The second requirement is that all concepts and definitions need to be simple and universally applicable. Complexity is both costly and likely to lead to confusion on behalf of the users. Concepts and definitions need to have broad meaning and universal acceptance. With the globalisation of economies, new political and economic alliances and the recognition of cultural diversity, the need for concepts and definitions which have international currency is also becoming important.

The requirements of temporal consistency, simplicity and universality are constraining factors in the search for improved measures of rurality.

3.2 Reconceptualising Rurality

There are three contributions to this discussion: the evolving concept of rurality as the Arena Society, the ideas and information from other studies including the experience of those attempting to measure rurality and apply it over space; and the constraints imposed by statistical agencies for continuity and universality. The outcome may be formulated as three options: **to make the best** of what there is; **to adjust** what is there; or **to change** the concept and definition completely.

A. The New Status Quo

It is important to recognize that a valiant attempt at improving the situation in Canada has already taken place and that the Census Metropolitan Area (CMA) and Census Agglomeration (CA) constructs are the result. Instituted in the 1986 Census, the two population based, geographical constructs each have three tiers within them: the urbanized core, the urban fringe and the rural fringe. The CMA, as a whole has more than 100,000 population and the CA has 10,000+ population. Both geographical constructs permit the recognition of rural sub-units within them, and suggests that rurality will vary according to the differential in the order of magnitude of the population base.

The overall assumption is that the CMA and CA definitions involve the classification of space based on population interactions, that is, those used to describe the labour market. It attempts to recognize the dominant influence of major metropolitan centres over the social and economic structure of the surrounding hinterland. This classification is based on the traditional concept of a metropole/hinterland relationship and takes labour patterns as indicative of this relationship.

Although this objective in itself is valid, it also becomes the weakness of the system as it assumes that all human inter-actions are governed by metropolitan relationships. More importantly it assumes the dominance of **one** urban centre of agglomeration and this is not what we recognize in the emerging arena society where multiple centres and modes of interaction are common, especially in rural areas.

In essence, the urbanistic focus, the unipolarity, the arbitrary population size cut-offs, and the dependency on labour markets are weaknesses which render the CMA/CA classifications unsatisfactory from a rural perspective.

B. The Adjustment Approach

Option B is to improve the situation by making adjustments to the existing definitions without doing violence to the need for temporal continuity. It was decided to retain the geographical units (census sub-divisions) to maintain spatial continuity but to multiply the categories of rural, by searching for population cut-offs that showed significant difference on some key variables that were selected from our understanding of contemporary rural conditions in Southern Ontario.

It was decided to test this approach by using 2A/2B profiles of the 1986 Census for Ontario at the Census sub-division level. The data were transformed into percentages and then into quintile ranges for purposes of comparison between centres of varying sizes. Population, population density, and distance from urban population centres of specified size were selected as independent variables. The following were considered to be fundamental indicators of the social economy of any geographical unit and were selected as the dependent variables:

housing type
migration
employment
education
infrastructure
income.

A Kolmogorov-Smirnov test was run to test for differences in the selected indicators between the population categories. As the test was run on data transformed into quintile ranges, the results indicated levels of difference in the structure of the variables between units, not differences in the values of the variables.

It is interesting to note that when density was factored into the analysis using the official definition of <1,000 population and <400 population density to equal urban, 52% of the cases were not classified within the given parameters. Subsequent analysis on the 52% as a statistical group showed the validity of 1,000 population as a significant cut-off, but 400 population density as not so. This confirms that low population does not necessarily mean low density and vice versa and reflects the size and location of geographical units rather than rurality.

Figure One illustrates the outcome of the tests of difference between the selected indicators for 6 groups of population size (1-999, 1,000-2,499, 2,500-4,999, 5,000-9,999, 10,000-19,999, 20,000+). All indicators show a statistical difference between areas with less than 1,000 population from those of 1,000-2,499 population. Some differentiations occurred between units of populations 2,500 to 10,000, but without consistency. Units with between 10,000-20,000 population differed on two indicators and places above 20,000 were different again. The resultant pattern, shown in Figure One, is simple and revealing. To emphasize their characteristic differences, when examined in reality, we ascribed names to the population size groups.

Figure 1

Rural Canada

0 - 1,000	1,000 - 10,000	10,000 - 20,000
Rural Area	District Centre	Regional Centre
Adjacent Remote	Adjacent Remote	Adjacent Remote

A **Rural Area** is one with less than 1,000 population and can either be near or remote from a major population centre. A **District Centre** has between 1,000-10,000 population and may vary according to location (adjacent or remote from major urban centre). A **Regional Centre** has between 10,000-20,000 population.

All three groups of census subdivisions could be described as comprising Rural Canada in a collective sense, as most of the spatial, and population variations in rural areas are accounted for. We also know that the population size groups are different on income, migration and employment indicators which are fairly good surrogates of the labour market. Together with the proximity to urban centre variable, another labour market factor, we can reasonably assume that most rural situations are covered in the three part grouping.

C. The Multifactor Approach

The employment of a 2-stage process involving multivariate classification of the lowest level of available Census data as the basis of selected socio-economic indicators and the subsequent classification of larger census divisions on the basis of the proportion and combination of these socio-economic groups found within them, is an approach worthy of further consideration. It could be used in combination with the more general indicators such as population size, density and metropolitan proximity.

Equally important, the multi-variate approach permits us to describe rurality by means of changing lifestyles. Well chosen indicators may effectively capture the essence of the Arena Society while allowing the residual conditions of the short-distance and the industrial societies to remain in the calculus and to be relevant.

Some indication of what the first stage of this process might look like is found in the Lifestyles (TM) market segmentation classification developed by Compusearch using the 1981 Census and extensively re-worked using 1986 Census data.

The 1986 Lifestyles (TM) classification is based on a non-hierarchical cluster analysis of thirty-five variables which reflect income, education, age of head of household, household size, employment and occupation, household mobility, dwelling type and tenure, residential setting and mother tongue (Compusearch 1989). Separate analyses were conducted for areas over 25,000 ("urban") and the remaining small cities, towns and rural areas ("Rural"). The analyses generated forty-eight "urban" and twenty-two "rural" clusters representing comparatively homogeneous "Neighbourhoods" (enumeration areas) on the variables included in the analysis. The seventy clusters produced are further grouped into thirteen broad aggregated categories.

Although capturing the richness of the database, albeit for a market segmentation purpose, a more simplified general classification system for EA's might be developed using a more restricted list of variables without an a-priori distinction between "urban" and "rural" on the basis of the 25,000 population cut-off. Larger geographic census units could then be classified on the basis of both the diversity and relative proportion of the cluster types contained within them.

A multifactor classification approach overcomes some of the limitations of aggregation, classification and causation inherent in the present system. The problem of aggregation is addressed by permitting reporting space to be organized by social and economic activity rather than some existing predefined units. The multifactor approach allows for the classification of units on the basis of contemporary patterns of behaviour and hence permits the exploration of new and perhaps more relevant questions. Finally, by treating population as just one variable rather than as the independent variable, the approach permits research to move beyond the simplistic population behaviour causal relationship.

Conceptually, a multifactor classification approach is more in accord with the new spatial reality of the Arena Society, a reality which is more mobile, and in which social and economic activity is more dispersed. The Arena Society represents the broader trend towards globalisation, in which technological change in communications has dispersed values and information and in which improved transportation has reduced real distance. Population size per se is thus less of a factor in social and economic differences. The multifactor classification approach moves us beyond the implicit population size/behaviour relationship.

The principal limitations of the multifactor classification approach relate to its lack of temporal continuity and its subjectivity. Implementation of a multifactor approach in future censuses would result in a lack of comparability both with data reported in the past and also with data collected in future censuses, since to retain its contemporary utility the classification would be reworked following each census. The multifactor approach is thus more subjective than simplistic definitions which could remain constant over time.

4. SUMMARY AND CONCLUSIONS

From our preliminary review it is evident that changing the concept and definition of rurality so that it more accurately reflects present reality is a daunting task. The requirements of simplicity and temporal continuity have produced in all the western industrial nations, definitions of rural space based on population and related measures such as population density.

Although a classification system based purely upon population, population density and proximity to major metropolitan centres has the apparent advantage of simplicity, it also has a number of problems. These problems may be categorized as problems of aggregation, classification and causation. The limited number of indicator variables and the gross level of geographic representation fail to capture the considerable socio-economic diversity of larger geographic reporting units. Existing spatial reporting units, based upon administrative subdivisions, are also reflective of the short-distance/industrial society and the use of such administrative subdivisions may not adequately capture new emerging patterns of interaction. Moreover, the existing urban/rural definition presumes a correlation between population size and population density and socio-economic activity and interaction. Given that the nature and form of this relationship is clearly changing, such simplistic secondary indicators may no longer be adequate as descriptors.

Three observations emerge from this paper which suggest further research and deliberation.

1. The rural reality in Canada has evolved into a complex set of characteristics which differ as much from each other as much as they distinguish rural from urban. A thorough academic debate needs to be undertaken to reach consensus on just how many broad types of rural space there are and what the leading indicators of difference are among them.

2. A specific study should be made of the multifactor classification approach to classifying rurality and rural space. This approach offers the most promise as it includes a dynamic element in its conception and permits redefinition over time.

3. Further examination of the adjustment approach using population cut-offs below 20,000-25,000 population needs to be undertaken across Canada.

If such enquiries are undertaken in a serious and planned way, then the opportunity to fully comprehend the viable options for redefinition of 'rural' could be reached in time for a change of census definition before the twenty-first century.

REFERENCES

Beale, C. (1977). Quanti-dimensions of decline and stability among rural communities. In Richard Rodefeld (Ed.) *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. Paper presented at The Contemporary Social and Economic Restructuring of Rural Areas Conference, London, U.K.

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., and Beeman, J. (Eds.) (1991). Distance and diversity in non-metropolitan economies. L'Université du Québec à Montréal, unpublished document.

Gertler, M., and Baker, H.R. (Eds.) (1989). *Sustainable Rural Communities in Canada*. Proceedings of Rural Policy Seminar No. 1, Saskatoon, Saskatchewan, October 11-13.

Gilbert, E.W. (1960). The idea of the region. *Geography*, XLV, 157-175.

Grohn, K. (1991). Rural policy in Finland. Written to the OECD Rural Program Secretariat, Ministry of the Interior, Department of Municipal and Regional Development, Finland.

Hart, J.F. (August, 1991). Part-ownership and farm enlargement. Paper presented at the Contemporary Social and Economic Restructuring of Rural Areas Conference, London, U.K.

Henrichsmeyer, W. Sustainable rural development: Objectives and constraints. Institute for Agricultural Policy, Bonn University, unpublished document.

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. (August, 1991). Farm adjustment in a period of uncertainty: Some aspects of the British experience. Paper presented at the Contemporary Social and Economic Restructuring of Rural Areas Conference, London, U.K.

Persson, L.O., and Westholm, E. (1991). Changing macro conditions reflected by rural households. Paper presented at Rural Change in Europe, 5th Review Meeting, Sila Greca, Calabria, Italy.

Summers, G.F., Bryden, J., Deavers, K., Newby, H., and Sechler, S. (Eds.) (1988). *Agriculture and Beyond, Rural Economic Development*, Madison, Wisconsin: University of Wisconsin.

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

URBAN-RURAL DICHOTOMY: AN OVERVIEW OF CURRENT CRITERIA AND FUTURE RESEARCH

N. Torrieri and J. Sobel[1]

ABSTRACT

The U.S. Census Bureau's Geography Division (GEO) is reexamining its criteria and procedures for identifying and delineating urban places and urbanized areas (UAs) for the 2000 census. The GEO is considering a number of different approaches, including the elimination of the reliance on place boundaries, to delimit UAs. In addition, the GEO will examine alternative delineation techniques, such as moving averages. The GEO also will examine the feasibility of developing new geographic areas to further classify and differentiate urban population clusters in the U.S. The goal of these activities is to improve the Census Bureau's ability to distinguish urban population from rural population.

KEY WORDS: Urban; Rural; Urbanized area; Urban place; Population density.

1. INTRODUCTION

The Census Bureau defines the urban population of the United States for each decennial census. This urban population includes all persons living in places[2] with minimum populations of 2,500 and those living in geographic statistical areas known as "urbanized areas" or UAs. A UA comprises a place or places and the adjacent densely-settled surrounding territory that together have a population of 50,000. The Census Bureau defines as rural all population that is not urban.

The Census Bureau's urban and rural definitions provide a framework for scholarly research on population distributions, settlement, and change. They also serve a variety of programmatic functions. Many private sector organizations base their marketing and site location decisions on the economic and social characteristics of particular UAs, or the proximity of a site to a UA. In addition, the UA forms the geographic basis for implementing dozens of government programs, from the establishment of automobile emission standards to the determination of Federal reimbursements to hospitals.

Millions of dollars in highway and mass transit funds are allocated to areas that qualify as UAs. Local officials are aware that the UA designation can mean the difference between a budget surplus and a budget deficit, between providing essential services or eliminating them. They sometimes attempt to influence UA designations in their favor by contacting Census Bureau officials directly or through their congressional delegations. Given these factors, the Census Bureau takes extreme care during its UA delineation operations to ensure that the UA criteria are applied consistently and fairly nationwide.

---

[1] N. Torieri & J. Sobel, Geography Division, U.S. Bureau of the Census, Washington, DC 20223-0001 U.S.A.

[2] The term "place" in the UA criteria includes both incorporated places, such as cities and villages, and census designated places (CDPs). A CDP is an unincorporated population cluster for which the Census Bureau delineates boundaries in cooperation with state and local agencies. For Puerto Rico, the Commonwealth government and the Census Bureau cooperatively delineate and recognize zonas urbanas and communidades as place equivalents, which are used when applying the UA criteria.

After each decennial census, the GEO reexamines the criteria and procedures it employs to identify and delineate the Nation's urban population. For the 2000 decennial census, this process now has begun.

2. AN HISTORICAL OVERVIEW OF THE CENSUS BUREAU'S URBAN/RURAL DEFINITIONS

The Census Bureau's urban and rural definitions date to the first census of 1790, which distinguished as a separate data tabulation, but not as "urban" population, those persons living in incorporated places of 2,500 or more. The first Census Bureau publication to designate places of 2,500 or more as urban was a 1906 supplement to the 1900 census[3]. The publication did not present any reason for selecting this threshold except to state that it formed a more realistic dividing line between urban and rural.

The 1900 Census of Manufactures defined industrial districts, the forerunner of today's metropolitan areas, around the four largest cities of the United States: New York, Chicago, Philadelphia, and St. Louis. These districts included population in the fringe areas of these cities, and consisted of densely-populated minor civil divisions (MCDs) within a distance of 10 miles from a central city. MCDs encompassed significant areas of rural territory and population, however, and some MCDs, underwent significant boundary revisions over time. Industrial districts were renamed metropolitan districts for the census of 1910; subsequently, metropolitan districts were defined for the 1920, 1930, and 1940 censuses.

Early nineteenth-century U.S. population settlement patterns, the basis for the original census definition of urban, generally had easily distinguishable and centralized population centers. By the late nineteenth century, the development of transportation technology, coupled with the availability of low-cost or low-rent land and housing, led to the movement of population outside the corporate limits of cities and villages. Simultaneously, just as large population clusters and commercial development started appearing outside of existing cities, many of these same cities were themselves changing, by gaining large tracts of sparsely-settled land through annexations and city/county consolidations. Thus, the political city no longer was suitable as the sole measurement unit for an urban region.

The Census Bureau attempted to improve its definition of urban to reflect these developments for the 1930 and 1940 census by implementing several revisions to the definitions of urban and rural. The most important of these established a population density threshold of 1,000 persons per square mile (ppsm) for governmental subdivisions defined as urban. In 1948, the Census Bureau hosted a "Conference on the Urban Fringe" to consider more far-reaching revisions to the definitions urban and rural populations. Two new statistical measures were adopted as a result of this conference. First, the Bureau of the Budget (predecessor to the Office of Management and Budget), in cooperation with other Federal agencies including the Census Bureau, established the standard metropolitan area (SMA) to define the metropolitan area around large cities. Second, the Census Bureau developed the UA definition to define the urban area and population around large cities.

The SMA provided a means of delimiting a functional zone of economic and social integration around a central place or places. The UA, in contrast, represented a measure of the extent of an urban agglomeration, including the built-up portion of a core place and the densely-settled surrounding area. The Census Bureau implemented its new definition of urban for the 1950 census. Beginning with that census, the Census Bureau recognized two components of the population as urban: those persons in urban places (incorporated or unincorporated population agglomerations of at least 2,500 persons that could be distinguished from scattered surrounding development and were outside of UAs) and those persons living in UAs.

3. CURRENT UA CRITERIA AND THE UA DELINEATION OPERATION

The chief characteristics of the 1990 UA criteria and the UA delineation operation are described below. Figure 1 illustrates the application of the UA criteria to the delineation of a UA for a hypothetical place and surrounding densely-settled fringe area.

---

[3] Twelfth Census of the United States, 1900, Supplementary Analysis, p. 20.

Place and Surrounding Area of 50,000 People

A UA comprises a place and the adjacent densely-settled surrounding territory that together have a minimum population of 50,000.

Minimum Population Density

There is no minimum population density criterion for the place that forms the core of the UA, although the core generally has a population density of at least 1,000 ppsm. The densely-settled surrounding areas adjacent to the place generally consist of territory made up of one or more contiguous census blocks having a population density of at least 1,000 ppsm. The Census Bureau uses population density as a measure of urban agglomeration because this measure is easily calculated and represents a demographically sound basis for determining the extent and distribution of population concentrations. Also, population and areal measurements for geographic entities, the basis for determining population density, are available within approximately six months after the day of the census.

Inclusion of Whole Places

The Census Bureau currently defines UAs in terms of whole places; a place in the surrounding densely-settled fringe area of the central place(s) is included if it has a minimum population of 2,500 and if at least fifty percent of the population of the place lives in qualifying blocks of 1,000 ppsm.

To achieve a better separation of urban and rural population for such "overbounded" places, the Census Bureau adopted the extended city concept for the 1970 census. The extended city definition allows for the recognition within places of component rural areas provided such areas have population densities less than 100 ppsm.

Inclusion of Low Density and Discontiguous Qualifying Area

UAs may contain low-density holes or gaps, such as parks or railyards, within the central place or within the contiguous densely-settled areas surrounding the place. The UA criteria also allow for the inclusion of discontiguous qualifying exclaves separated from the main body of the UA by low-density areas, and connected by road to the main body of the UA.

Exclusion of Certain Land Use Areas from Density Measurements

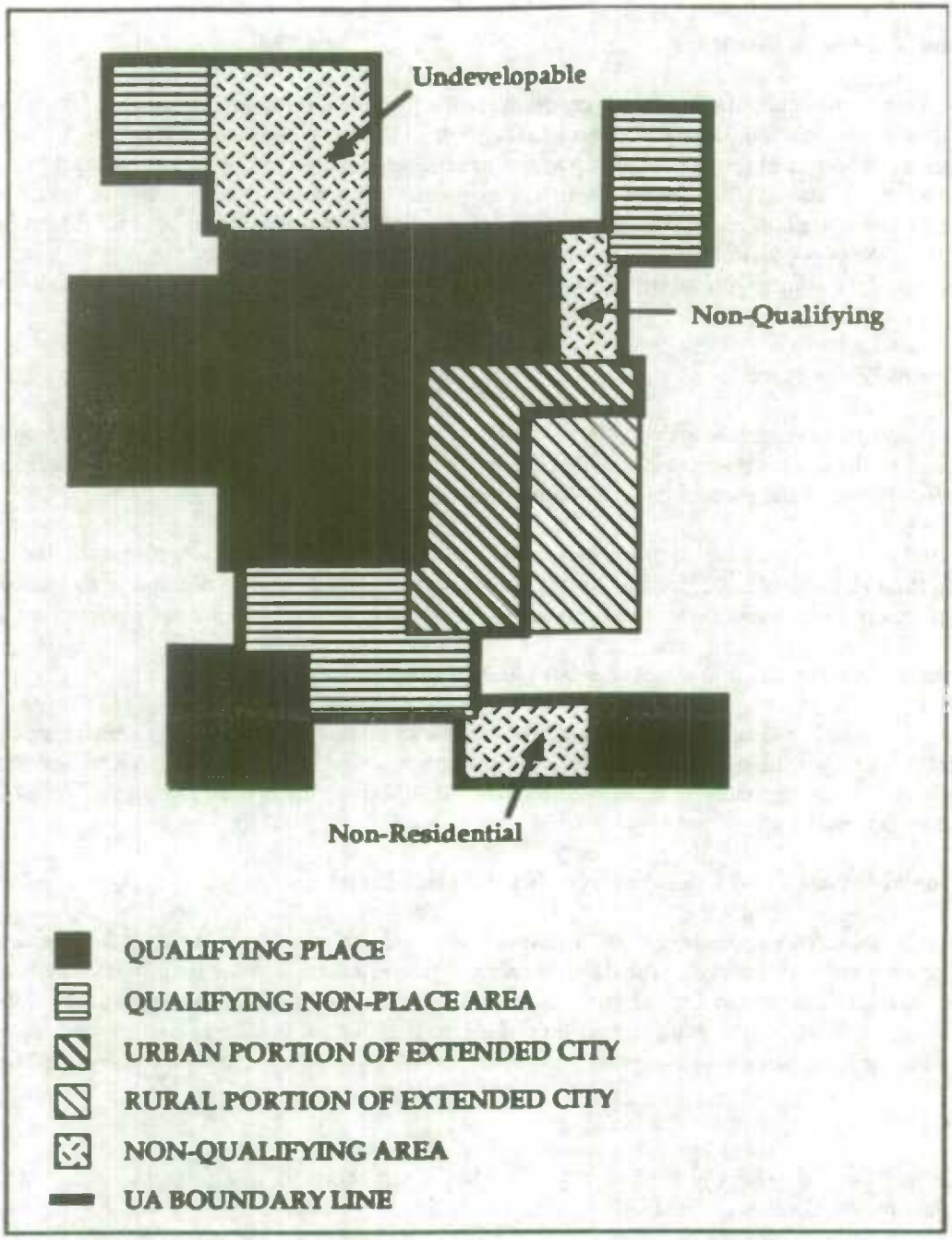
Industrial, commercial, recreational, and transportation land use areas (such as large industrial parks, national parks, or airports) and undevelopable land use areas (such as marshlands or mud flats) may be excluded from density measurements if the Census Bureau has information documenting their existence. When such areas occur along the edge of the qualifying portion of a UA, they do not qualify as part of the UA unless they can be used to reach noncontiguous exclaves of qualifying territory having a population density of at least 1,000 ppsm.

UA Mergers

The Census Bureau generally merges adjacent UAs when they are mostly in the same MSA or Primary Metropolitan Statistical Area (PMSA).

The delineation of UAs for the 1990 census was largely automated. Geographers in the Census Bureau's twelve regional census centers (RCCs) delineated the potential 1990 UAs using software developed by the GEO. Geographers in the GEO then reviewed each delineation and made revisions as necessary. Then, they identified all potential UAs that had a population of at least 50,000 persons; these are the final UAs that the Census Bureau recognizes in its 1990 tabulations.

Figure 1. A Hypothetical Urbanized Area



4. FUTURE DIRECTIONS

We currently are evaluating the definitions, criteria, and methodologies that the Census Bureau uses to identify urban population. Four major research categories are of particular interest: concepts and definitions, delineation techniques, variable criteria, and new geographic areas.

4.1 Concepts and Definitions

Urban Places

Outside of UAs, the Census Bureau classifies entire places (both incorporated places and census designated places) as urban if they have a population of at least 2,500. Within UAs, we also designate entire places as urban, including those blocks that do not meet the 1,000 ppsm population density requirement. The Census Bureau does not recognize any blocks outside of urban places or UAs as urban, regardless of their population or population density. As a result of our UA criteria, many sparsely populated blocks both inside and outside of UAs are delineated as urban blocks. Conversely, many blocks outside of UAs that are densely populated are delineated as rural solely because they are not within a place. We may decide to modify our concept of urban place to account for these situations.

Land Use Classification

The Census Bureau recognizes two classifications of sparsely-populated land that can be included in UAs: undevelopable land and non-residential urban land. By classifying a census block in one of these categories, we then can include additional adjacent densely-populated blocks in a UA that would otherwise be separated, and therefore excluded, from the main body of the UA (Figure 1). Several areas that qualified as UAs for the 1990 census did so only because we were able to include additional blocks in the UA in this fashion. Conversely, some areas failed to qualify as UAs because we could not justify using the land use categories.

Implementing these classifications presents several problems. First, we have not been able to devise objective, unambiguous definitions that can be applied consistently nationwide. Second, the Census Bureau does not have comprehensive data sources describing the characteristics of all land parcels in the United States. For the 2000 census, we want to establish more objective methods for categorizing such areas, determine whether it is practicable to obtain and utilize the source material to support these methods, or else decide whether we can devise alternative delineation techniques that eliminate the need for such classifications altogether.

4.2 Delineation Techniques

Moving Average

Our traditional approach to delineating a UA is to examine a block (or sometimes groups of similar, contiguous blocks), simultaneously taking into account place boundaries and land use. If these blocks qualify to be in the UA, we examine the next separate block or group of blocks. We are considering the use of a "moving average technique," which involves the determination of a boundary between urban and rural populations through a sequence of overlapping block configuration operations. In this process, we configure groups of adjacent blocks and determine their cumulative population density, then accept or reject the blocks for inclusion within the UA on the basis of their population density. We then create another group of contiguous blocks that includes some of the blocks that we previously added to the UA as well as "new" blocks that we have not yet examined (Figure 2). As each subsequent group of blocks qualifies to be in the UA, we continue to include some of these blocks in the next group we examine.

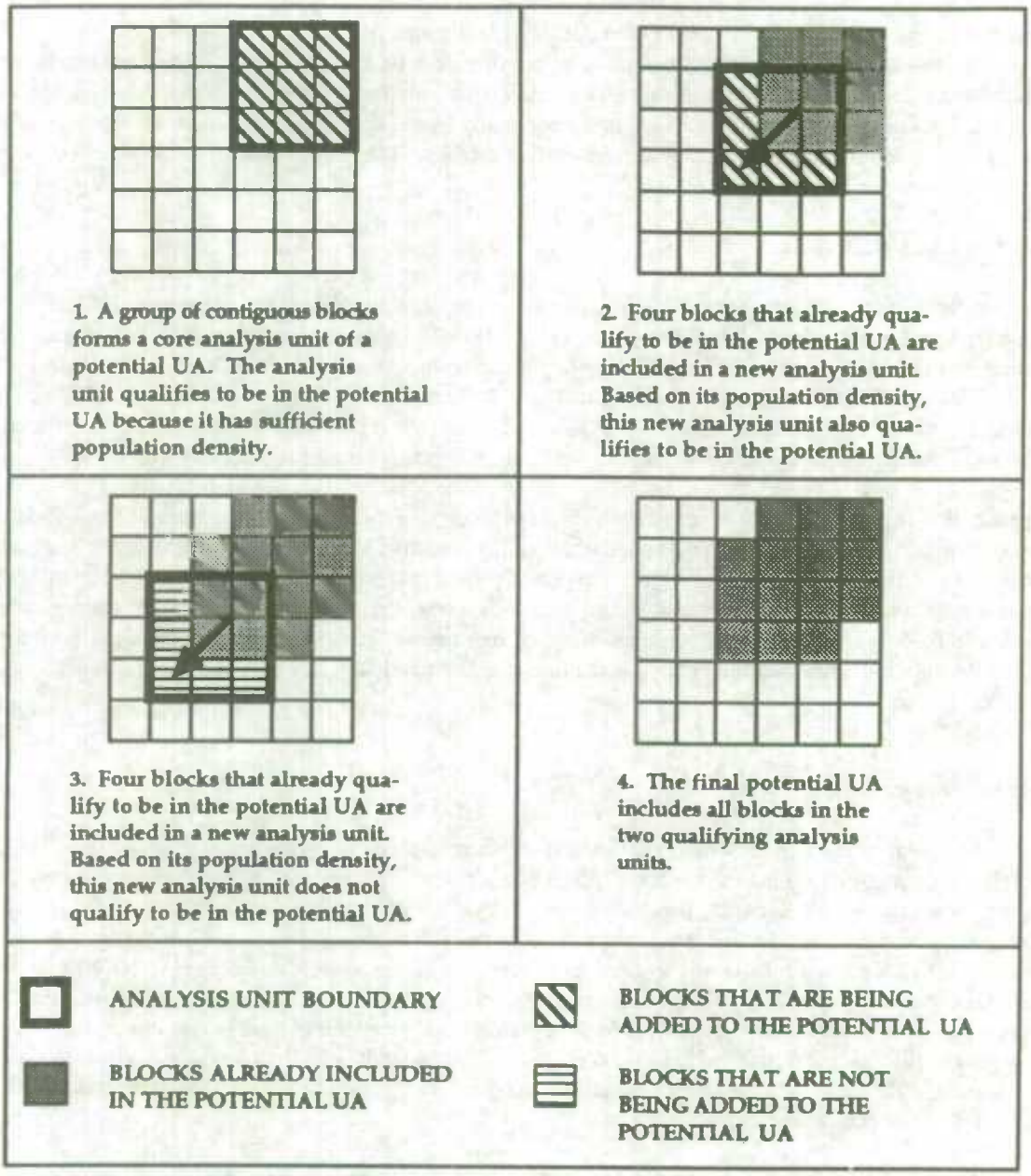
This method frees us from incorporating subjective land use classifications into the delineation process. In addition, by designating a UA boundary on the basis of population density measurements of overlapping groups of blocks, minor and non-representative block-to-block variations in the population density within an urban agglomeration are eliminated because they are averaged with the population density of adjacent blocks. This allows us to better analyze population density patterns of UAs and to identify the significant population density thresholds that separate urban and rural populations.

Population Density Zones

We also will consider examining the feasibility of delineating zones that represent different ranges of population density within areas that already qualify as UAs. Further, we will examine the effect of adopting new population

density criteria to help us delineate these zones more accurately. This latter approach recognizes that our dependence on a single population density criterion to delineate UAs may no longer be acceptable.

Figure 2. The Moving Average Technique



4.3 Variable UA Criteria

Many of the factors that significantly affect our UA delineations vary greatly nationwide. For example, in older, northeastern cities, street networks often are more compact than elsewhere in the nation, particularly the west, where blocks are larger due to less dense street network patterns and an absence of suitable hydrographic features to use as block boundaries. In many states, incorporation laws allow cities to annex territory almost at

will, while in others it is almost impossible for cities to annex. These geographical and morphological variations may support the adoption of spatially variable UA criteria.

For the first time since the Census Bureau began delineating UAs, areas that qualified as UAs in the previous census failed to qualify as UAs in the 1990 census[4]. The Census Bureau did not make provisions to continue recognizing them if they failed to meet the UA criteria. We must decide if it is appropriate for both conceptual and programmatic purposes to "grandfather" UAs, or whether we can identify a limited "grace" period (for example, one census) in which an existing UA's population can fall under 50,000 without the area losing its UA status. We also will consider establishing a lower qualifying population criterion for previously recognized UAs.

4.4 New Geographic Areas

Given the many programmatic uses of urban designations and UA delineations, as well as the great importance many analysts place on data comparability, we may consider not changing our current criteria and techniques drastically for the next census after all. As an alternative, we may choose to produce supplementary delineations between censuses, and let data users choose the products they find most satisfactory for their particular needs.

5. CONCLUSION

The issues presented here will not be resolved easily, and any results and conclusions must reflect the needs of those people and organizations that are most affected by our definitions of urban and rural. We will not alter our definitions of urban and rural without broad-based public support. At the same time, we recognize the need for a dynamic and far-sighted approach to improving our current definitions of urban and rural, an approach that will provide data users with a more useful classification of urban and rural populations in the future.

---

[4] Danville, Illinois and Enid, Oklahoma, two previously qualifying UAs, no longer qualify as UAs as a result of the 1990 census.

SESSION 10

Data Analysis from a Geographic Perspective

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

STATISTICAL ANALYSIS OF SPATIAL URBAN CENSUS DATA IN THE PRESENCE OF MISSING VALUES

D.A. Griffith[1]

ABSTRACT

Missing values estimating methodology is applied to the 1986 geographic distribution of median family income by census tracts for the Ottawa-Hull metropolitan area. This methodology is based upon maximum likelihood techniques. A simple comparison is made between autoregressive response model results obtained with this Canadian census data application, conventional regression model results for this data application, and conditional autoregressive model results obtained with the 1980 United States census data for Houston. Improvement of precision of the missing values estimates in this application is explored.

KEY WORDS: Autoregressive; E-M algorithm; Missing values; Precision; Spatial autocorrelation.

1. INTRODUCTION

Missing values can create troublesome difficulties for the analytical and visual analysis of spatial data. They yield geographic distributions containing holes, and hinder attempts to generalize map surfaces. The purpose of this paper is to illustrate a methodology for handling missing data in Canadian urban census publications. One version of this methodology already has been applied to a United States urban census data set (see Griffith, Bennett, and Haining 1989).

Often sample census data are tabulated by specified areal units (*e.g.* census tracts), and then the aggregated figures made available to the public. Because of confidentiality rules and concerns, selected resulting figures often are suppressed, since this *ex post facto* tabulation does not guarantee some minimum individual areal unit sample size that prevents the possibility of associating published figures with any identifiable household, hence ensuring information about each household is masked. Therefore, there is no *a priori* reason to suspect that the missing values that occur arise from some systematic factor; accordingly, the perspective utilized in this paper treats them as random variables.

1.1 Background

Griffith, Bennett, and Haining (1989) studied the 1980 geographic distribution of median family income by census tracts across the Houston metropolitan area. This urban place had been partitioned into 363 census tracts, with income values having been suppressed for three of them. As is the case for Ottawa-Hull, the censoring process used to delete Houston data was derived from the number of households in a census tract, and hence was considered to be independent of median family income figures.

The Houston income variable displayed relatively strong positive spatial autocorrelation, producing the conditional autoregressive model parameter $\hat{\rho} = 0.1747$ ($\hat{\rho}_{max} = 0.1755$). This geographic distribution exhibited a non-constant mean surface; income in census tract i was found to be a function of the Cartesian coordinates (u_i, v_i) of that tract's physical centroid. Spatial statistical missing value estimates were improvements upon

[1] D.A. Griffith, Department of Geography, and Interdisciplinary Statistics Program, Syracuse University, Syracuse, New York, U.S.A., 13244-1160.

conventional missing value estimates, by 10%, 30%, and 4%. Unfortunately the associated confidence intervals were very wide (poor precision), and in one case the lower bound was negative, and hence for practical reasons had to be truncated at zero. One conclusion reached was that additional explanatory variables needed to be included in the model specification.

1.2 Ottawa-Hull Metropolitan Area

The data to be analyzed in this paper are from the 1986 Canadian Census of Population (a digital copy was supplied by Statistics Canada), and are for the 192 census tracts into which Ottawa-Hull has been partitioned. The total population of this metropolitan area is 819,263, and its total area is 5138.33 square kilometers; hence the average density is 159.44 people per square kilometers. Census tract centroids are given in terms of Universal Transverse Mercator (UTM) coordinates; because they are on an interval measurement scale, these coordinates have been rescaled for convenience. The average median family income by census tract, for those where it is reported, across the metropolitan area is $40,522, with a range of $16,908 to $59,902; the actual median family income for the urban area is $41,775.

One complicating idiosyncracy of Ottawa-Hull is that the metropolitan area is overlayed by two prominent cultural regions. Roughly speaking, the boundary separating these two regions coincides with the Ottawa River, which also is a political boundary. Therefore, one cultural area is located in the Province of Ontario while the other is located in the Province of Quebec. This cultural differentiation will be incorporated into the analysis by including an indicator variable; this variable takes on a value of 1 when a census tract is located in Ontario, and a value of -1 when a census tract is in Quebec (this parameterization allows for a direct difference of means test to be conducted).

2. EXPLANATORY DETERMINANTS OF THE GEOGRAPHIC DISTRIBUTION OF INCOME IN AN URBAN AREA

Urban economics suggests that three salient factors impact upon the geographic distribution of income in cities (see Richardson 1977), namely population density, gradients, and spatial externalities. First, income is inversely related to population density. More affluent households tend to have an income elasticity of demand for space that causes them to have a strong preference for additional space. Their utility maximization produces the aforementioned inverse relationship, as accessibility to centrality is traded off against transport costs. Ideally, a marginal change in income would offset a change in transport costs that would be incurred by choosing a given location. Thus, given the preference of high-income groups for low densities, and their ability to pay, income tends to increase as density falls.

Second, income gradients exist throughout an urban area. The type of income surface may well resemble a circus tent or three-dimensional spider's web, with many peaks and troughs. The highest peak tends to occur at the city's center. Local peaks of higher income occur in sections of a city that are preferred for residential purposes. The troughs are attributable to the presence of competing land uses, or the existence of blighted neighborhoods. This characterization is consistent with Burgess's and Hoyt's conceptualizations of the spatial organization of urban areas.

Third, residential location also is a function of neighborhood amenities, a preference for pleasant living environments, and socio-economic features of neighborhoods. These elements lead to spatial externalities, which arise from specific sites conferring advantages in addition to their accessibility to centrally located places in a city. Thus, similar income groups tend to cluster together in neighborhoods, with the higher-income households being willing to pay for these positive externalities. Such spillover effects embed spatial autocorrelation in the geographic distribution of income. This characterization is consistent with Harris and Ullman's multiple-nuclei conceptualization of the spatial organization of urban areas.

3. SPECIFYING AN AUTOREGRESSIVE MODEL FOR OTTAWA-HULL

Two model properties are of concern here. First is the Jacobian of the transformation from an autocorrelated domain to an unautocorrelated domain, especially when $n = 192$ and the surface is covered by an irregular partitioning. Second is the selection of the spatial autoregressive model specification.

The Jacobian term for Ottawa-Hull involves a 192-by-192 matrix, whose determinant has to be re-evaluated numerous times. Griffith (1992) outlines a useful numerical approximation to this determinant. If the autoregressive model is written in terms of the standardized version of the binary connectivity matrix C ($c_{ij} = 1$ if census tracts i and j are juxtaposed, and $c_{ij} = 0$ otherwise; each row is scaled so that it sums to unity), say W ($w_{ij} = c_{ij} / \sum_{j=1}^{n} c_{ij}$), then the eigenvalues of this matrix are such that

$$\lambda_{max} = 1 \text{ and } \lambda_{min} = 0.62371 \rightarrow -1.601706 < \hat{\rho} < 1.$$

Drawing a systematic sample of size 22 from across this feasible parameter space yields the following Jacobian approximation ($SSE = 0.000099$):

$$\hat{J}_w^{1/n} = 0.237169 \bullet ln(1.873080) + 0.144759 \bullet ln(1.159028) - 0.237169 \bullet ln(1.873080 + \rho) - 0.144759 \bullet ln(1.159028 - \rho). \quad (3.1)$$

This approximation was employed in conducting the research summarized in this paper.

The autoregressive response model has been selected from the family of potential spatial autoregressive models, principally because it minimizes the number of spatial lag variables that must be dealt with. Consider a set of p variables, whose data are entered into matrix X. In addition, a vector of ones, denoted 1, must be concatenated with these p variables in order to include an intercept term. Let Y be the response variable vector, and let ξ be the error vector. The traditional linear statistical model that can be constructed from these terms is expressed as $Y = X\beta + \xi$, where β is a $(p+1)$-by-1 vector of regression coefficients. Accordingly, the spatial autoregressive response model may be expressed as

$$(I - \rho W)Y = X\beta + \xi \quad \text{or} \quad Y = \rho WY + X\beta + \xi,$$

where I is the identity matrix. This is the model specification that will be employed in this paper. It has been described by, among others, Upton and Fingleton (1985). Its parameters can be estimated using standard statistical software packages by rewriting it as

$$Y\exp(\hat{J}_w^{1/n}) = \rho WY\exp(\hat{J}_w^{1/n}) + X\beta\exp(\hat{J}_w^{1/n}) + \xi\exp(\hat{J}_w^{1/n}). \quad (3.2)$$

Griffith (1988b) outlines a strategy for obtaining parameter estimates using equation (3.2).

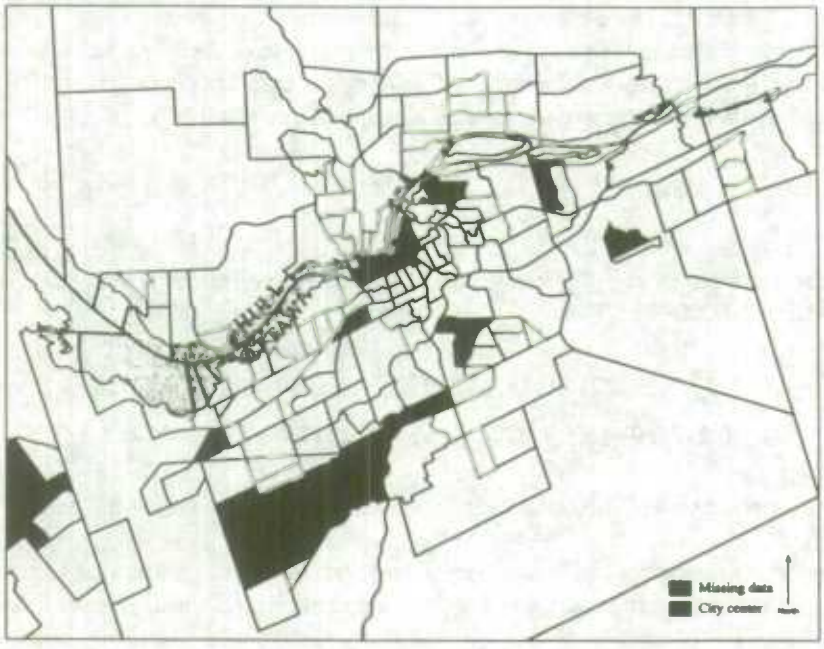
4. ESTIMATING OTTAWA-HULL MISSING CENSUS TRACT MEDIAN INCOME FIGURES

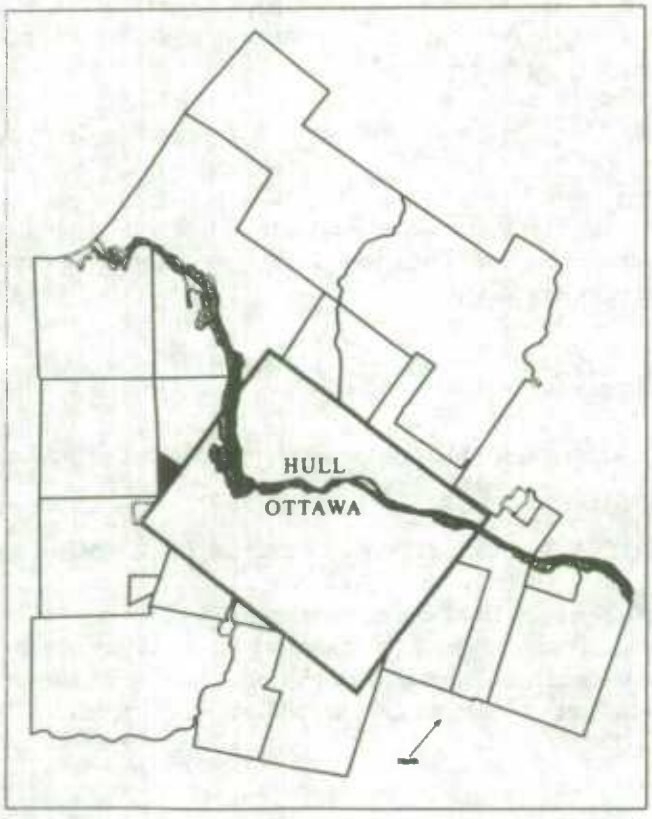
In the case of Ottawa-Hull, there are three census tracts[2] with respective populations of 13, 45, and 64. Rarely would a national unconstrained random sample ever produce the necessary minimum sample size in any of these three areal units that would satisfy confidentiality restrictions. Statistics Canada presumably believes that this contention holds for populations as large as 250. In addition, median family income figures for eight other

[2] #47, #140.01, and #160.03.

census tracts[3] have been suppressed; their corresponding populations range from 1,766 to 6,083, and hence presumably they are victims of the non-spatially stratified sampling design employed by Statistics Canada. The geographic context of these missing values is illustrated in Figure 1. Except for a single triplet cluster, these missing values are dispersed across the city.

Figure 1: Map displaying census tracts with missing 1986 median family income values.





[3] #2.01, #6, #34, #110, #120.01, #125.01, #130.01, and #137.03.

4.1 A Conventional E-M Algorithm Type of Solution

A conventional E-M (the iterative estimation of missing values followed by maximization of the likelihood function; see Little and Rubin 1987) solution to the missing median family income values for Ottawa-Hull involves the following steps:

Step 1: compute $b_{\tau=0} = (X_o^t X_o)^{-1} X_o^t Y$,

Step 2: compute $\hat{Y}_{m,\tau} = X_m b_\tau$,

Step 3: compute $b_{\tau+1} = (X^t X)^{-1} X^t \hat{Y}$, where $\hat{Y}^t = \langle Y_o^t : \hat{Y}_{m,\tau} \rangle$, and

Step 4: iterate between Steps 2 and 3 until $|b_{\tau+1} - b_\tau| \approx 0$.

Execution of these steps was achieved with SAS using PROC NLIN (see Section 5.1).

Table 1: Estimation Results in the Presence of Missing Response Values

Conventional linear regression model with deletion of missing values

Order of	Regression Model		Regression Residuals			Extra sum of	R^2
Model	df	Sum of squares	df	Sum of squares	df	squares	
Zero	2	4033904135.2	178	15167150599	2	4033904135*	0.21
Linear	4	4322552193.5	176	14878502541	2	288648058	0.23
Quadratic	7	4612198066.2	173	14588856668	3	289645873	0.24
Cubic	11	4620744265.8	169	14580310469	4	8546199	0.24

* denotes a significant contribution to the regression sum of squares when compared with the mean square error of the full model.

Conventional linear regression model estimates with missing values

Procedure	β_0	β_1	β_2	MSE
Single step	42419.22165	4577.94548	-1.37701	85208711
Iterative	42419.19599	4577.93092	-1.37700	80249474

Spatial statistical autoregression model estimates with missing values

Procedure	ρ	β_0	β_1	β_2	MSE
E-M type	0.44900	23451.43946	3051.74277	-0.94187	71722838
Iterative	0.47327	22456.38101	2958.47903	-0.92201	71425230

Completion of Step 1 required specification of the regression equation. Griffith, Bennett, and Haining (1989) found that the median family income surface for Houston could be characterized in part by a linear gradient (or trend surface) model. Urban economics theory also casts urban income surfaces as a function of population density, as described in Section 2. And, for the Ottawa-Hull region, cultural differentiation can be taken into account by including an indicator variable as discussed in Section 1.2. Initially all three of these factors were analyzed. Results of this exercise are reported in Table 1. According to the extra sum of squares principle, none of the UTM coordinate terms were found to be significant. In fact, the linear north-south coordinate axis was displaced as being important when population density entered the regression model. Consequently, the conditional average census tract median family income was found to be $42,419, with Ottawa census tracts on average being $4,578 above this mean and Hull census tracts on average being $4,578 below this mean. Further, as population density per square kilometer increases, census tract median family income tends to decrease. These two factors of cultural differentiation and population density account for approximately one-fifth of the geographic variation in median family income across the Ottawa-Hull landscape. Model diagnostics yielded a nonsignificant Shapiro-Wilk statistic of 0.9910, supporting the notion that the parent population of regression

residuals is normally distributed. The graphical evaluation of heteroscedasticity indicates that there may be a slight departure from constant error variance, but nothing serious (the indication is that perhaps several marginal outliers are present). And, the significant Moran Coefficient for the regression residuals of 0.1876 implies that a modest amount of positive spatial autocorrelation is present.

Completing Steps 2, 3, and 4 rendered results that are negligibly different from those obtained in Step 1 (see Table 1). The mean square error (MSE) needs to be multiplied by 189/178 in order to be equivalent to the MSE from Step 1. The re-estimated coefficients obtained in these steps are the ones used to estimate the missing median family income values.

Table 2: Missing Median Family Income Estimates for the Eleven Ottawa-Hull Census Tracts: Missing Values Treated as Parameters

	Conventional Regression Model			Spatial Autoregression Model		
Census Tract Number	Lower 95% CI Bound	Missing Value Estimate	Upper 95% CI Bound	Lower 95% CI Bound	Missing Value Estimate	Upper 95% CI Bound
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	

Estimates of missing median family incomes are reported in Tables 2 and 3. Results appearing in Table 2 are obtained by treating these missing values as parameters. These estimates are accompanied by their 95% confidence intervals (see Section 6). In almost all cases the lower bound is considerably larger than the minimum observed median family income, the estimated value always is larger than the average or the conditional average median family income, and the upper bound always is higher than the maximum observed median family income (by as much as 10%). In contrast, the conditional expectations reported in Table 3 have lower bounds considerably closer to the minimum observed value (with two being less than this value), mean values that are exactly the same as the parameter estimates, and upper bounds that are even more extreme than their estimated parameter counterparts (all being greater than the maximum observed value).

Table 3: Missing Median Family Income Estimates for the Eleven Ottawa-Hull Census Tracts: Missing Values Treated as Observation Expectations

	Conventional Regression Model			Spatial Autoregression Model		
Census Tract Number	Lower 95% CI Bound	Missing Value Estimate	Upper 95% CI Bound	Lower 95% CI Bound	Missing Value Estimate	Upper 95% CI Bound
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 A Spatial Statistical E-M Algorithm Type of Solution: The Algebra

Martin (1984) presents the general equation for estimating missing values with spatial autoregressive models. His solution notes that the Jacobian term needs to be weighted such that, for some general inverse-covariance structure defined by matrix V, it becomes

$$J = -1n(\det|V|/\det|V_{mm}|)/n_o,$$

where $\det|V_{mm}|$ is the inverse-covariance matrix for the configuration of missing values, and n_o is the number of non-missing values. Equation (3.1) approximates $-1n(\det|I - \rho W|)/192$. Eight of the missing values for Ottawa-Hull are dispersed, while three form a linear arrangement. Thus, $\det|I - \rho W_{mm}| = \det|I| \bullet (1-\rho/24)(1+\rho/24)$, since the eigenvalues of matrix W_{mm} are 1/24, nine zeroes, and -1/24. Martin's result means that equation (3.1) needs to be modified as follows:

$$J_w^{1/4} = (192/181)[0.237169 \bullet 1n(1.873080) + 0.144759 \bullet 1n(1.159028) - 0.237169 \bullet 1n(1.873080 + \rho) - 0.144759 \bullet 1n(1.159028 - \rho)] + [1n(1-\rho/24) + 1n(1+\rho/24)]/181. \quad (4.1)$$

As an aside, $\det(|V_{mm}|)$ also can be approximated when sizeable clusters of missing values are present; it reduces to 1 when the missing values are completely dispersed.

Griffith, Bennett, and Haining (1989) have outlined an estimation procedure based upon the conditional autoregressive model, where

$$V = I - \rho C.$$

Griffith (1988a) has outlined an estimation procedure based upon the simultaneous autoregressive model, where

$$V = (I - \rho W)'(I - \rho W).$$

At present both of these models are quite difficult to implement with standard statistical software packages, such as SAS. In contrast, the autoregressive response model selected for this research project has as its missing values estimation equation

$$\hat{Y}_m = [(I - \rho W_{mm})'(I - \rho W_{mm}) + \rho^2 W_{om}' W_{om}]^{-1}\{[(I - \rho W_{mm})'X_m - \rho W_{om}' X_o]\beta + \rho[W_{om}'(I - \rho W_{oo}) + (I - \rho W_{mm})'W_{mo}]Y_o\}. \quad (4.2)$$

This equation reduces to $\hat{Y}_m = X_m\beta$ for spatially unautocorrelated data (*i.e.* $\rho = 0$), the conventional result described in Section 4.1.

Equations (4.1) and (4.2) have been used to obtain the spatial statistical estimates reported in this paper.

4.3 A Spatial Statistical E-M Algorithm Type of Solution: The Estimates

A spatial autoregressive response E-M type of solution to the missing median family income values for Ottawa-Hull involves the following steps:

Step 1: compute $b_{\tau=0} = (X_o'X_o)^{-1}X_o'Y$,

Step 2: compute $\hat{Y}_{m,o} = X_m b_{\tau=0}$,

Step 3: estimate $\hat{\rho}_\tau$ and b_τ for $\hat{Y} = \rho W\hat{Y} + X\beta + \xi$, where $\hat{Y}' = \langle Y_o' : \hat{Y}_{m,\tau}' \rangle$,

Step 4: compute $\hat{Y}_{m,\tau+1}$ using equation (4.2),

Step 5: iterate between Steps 3 and 4 until $|b_{\tau+1} - b_\tau| \approx 0$.

Execution of Steps 1, 2, 3 and 5 also was achieved with SAS using PROC NLIN (see Section 5.2); Step 4 was executed at each iteration with MINITAB code.

Completing Steps 2, 3, 4, and 5 rendered results that are only slightly different from those obtained in Step 1 (see Table 1). The mean square error (MSE) needs to be adjusted by both the Jacobian term and the n/n_o factor, if it is to be comparable with the MSE reported in Step 1. The level of spatial autocorrelation recovered indeed is moderate and positive. The conditional mean median family income has decreased to $22,456, and the difference between median family income in Ottawa and Hull has diminished by about $3,239.

Estimates of missing median family incomes are reported in Tables 2 and 3. Results appearing in Table 2 were obtained by treating these missing values as parameters. These estimates differ, but not substantially, from the predicted values reported in Table 3. Substantially tighter confidence intervals are found for the conditional expectations reported in Table 3, when compared with their aspatial counterparts. Here the lower bounds clearly are greater than the minimum observed value, now nonlinearity causes mean values to deviate from their parameter estimate counterparts, and upper bounds are closer to the maximum observed value, although they remain more extreme in many cases.

5. THE SAS CODE

SAS code has been developed for estimating missing values. This code is built around PROC NLIN, which produces the iterative solutions outlined above. This code is presented in Appendices I and II.

5.1 The Conventional Statistical Estimation Code

The SAS code for the conventional linear regression solution described in Section 4.1 appears in Appendix I. Lines 1-2 define the paths to the input files. Lines 3-12 access the attribute data housed in file "OTT-HULL DATA." Lines 6 and 7 rescale the UTM coordinates. Line 8 creates the cultural indicator variable. Line 11 computes the population density variable. Lines 13-16 compute the linear regression coefficients for the known values (Step 1 of Section 4.1). The predicted values, YHAT, then are used as the first estimates of the missing

values. Lines 17-21 replace the missing values with these first estimates. Lines 23-33 define individual indicator variables for each missing value; this step is in keeping with the treating of missing values as parameters. Lines 36-38 initialize the model parameters. Line 39 is the simple linear regression model. Lines 40-42 cause the missing values to be replaced by their estimates, as parameters BM_j, at each iteration in PROC NLIN. Lines 43-89 are the first and cross-partial derivatives used by PROC NLIN to optimize the objective function constructed from the model statement appearing in Line 39. Line 90 generates the expectations of the missing values as though they were observations; 95% confidence intervals are computed here, too.

5.2 The Spatial Statistical Estimation Code

The SAS code for the spatial autoregression solution described in Section 4.2 appears in Appendix II. Again, Lines 1-2 define the paths to the input files. Lines 3-11 access the attribute data housed in file "OTT-HULL DATA." Now Line 6 creates the cultural indicator variable, and Line 9 computes the population density variable. Line 10 drops the UTM coordinates, since they are not used in this analysis. Lines 12-23 access the connectivity matrix C, housed in file "OTT-HULL CONN,"for the Ottawa-Hull metropolitan area, computes $\sum_{j=1}^{n} C_{ij}$, and computes the first part of the spatial lag variable CY. Lines 24-27 complete the computation of the spatial lag variable, and Lines 28-30 convert these calculations to a column vector. Line 47 converts CY to WY. Lines 35-45 define individual indicator variables for each missing value; once more, this step is in keeping with the treating of missing values as parameters. Lines 53 initializes the model parameters. Lines 59-69 are the output from the MINITAB code that operationalizes equation (4.2) (see Appendix III). Lines 70-71 define the Jacobian of the transformation from an autocorrelated space to an unautocorrelated space; it is equation (4.1). Lines 72-77 cause the missing values to be replaced by their estimates, as parameters BM_j, at each iteration in PROC NLIN. Line 78 specifies the spatial autoregressive response model. Lines 79-84 are the first derivatives used by PROC NLIN to optimize the objective function constructed from the model statement. Line 86 generates the expectations of the missing values as though they were observations; 95% confidence intervals are computed here, too. Lines 87-91 divide out the Jacobian term from the predicted values.

5.3 The MINITAB Code

The iterative estimates of Y_m were obtained using MINITAB because of its matrix operation capabilities; SAS IML also could be used for this purpose. This code is presented in Appendix III. Line 2 reads in the SAS estimates of $\hat{\rho}_{\tau-1}$ and $b_{\tau-1}$. Line 8 reads in the estimation components based upon a partitioning of matrix W. Line 9 constructs matrix I_m. Line 10 computes $W'_{mm} + W_{mm}$. Line 11 computes $W'_{mm}W_{mm} + W'_{om}W_{om}$. Line 12 computes $\hat{\rho}_{\tau-1}(W'_{mm} + W_{mm})$. Line 14 computes $\hat{\rho}^2_{\tau-1}(W'_{mm}W_{mm} + W'_{om}W_{om})$. Lines 15-17 compute $[I_m - \hat{\rho}_{\tau-1}(W'_{mm} + W_{mm}) + \hat{\rho}^2_{\tau-1}(W'_{mm}W_{mm} + W'_{om}W_{om})]^{-1}$. Line 18 computes $W'_{mm}X_m$. Line 19 computes $W'_{om}X_o$. Line 24 computes $W'_{mm}X_m b_{\tau-1}$. Line 25 computes $W'_{om}X_o b_{\tau-1}$. Line 26 computes $X_m b_{\tau-1}$. Line 27 computes $X_m b_{\tau-1} - \hat{\rho}_{\tau-1}(W'_{mm}X_m b_{\tau-1} + W'_{om}X_o b_{\tau-1}) + \hat{\rho}_{\tau-1}(W'_{om}Y_o + W_{mo}Y_o) - \hat{\rho}^2_{\tau-1}(W'_{om}W_{oo}Y_o + W'_{mm}W_{mo}Y_o)$. Line 28 computes the results for equation (4.2). Finally, Line 32 writes a file that is input into the next SAS iteration.

6. PRECISION OF THE MISSING VALUES ESTIMATES

Tables 2 and 3 include information on the precision of the missing value estimates, in terms of interval estimates for them. This information is of two types. First, the missing values Y_m can be viewed as parameters, and then estimated. Second, the missing values can be viewed as conditional expectations for observations, with either confidence intervals or new observation prediction intervals constructed for them.

If a conventional analysis is undertaken, where the geo-referenced observations are assumed to be independent, and the missing values are set equal to the sample mean of the known data, say $\bar{x}_o$, then the entries needed to compute the asymptotic standard errors are as follows:

$$-E[\partial^2 ln(L)/\partial\mu^2] = n/\sigma^2,\ -E[\partial^2 ln(L)/\partial(y_{m_j})^2] = 1/\sigma^2,$$

$$-E[\partial^2 ln(L)/(\partial\mu\partial y_{m_j})] = -1/\sigma^2, \text{ and } -E[\partial^2 ln(L)/(\partial y_{m_j}\partial y_{m_k})] = 0.$$

Since $-E[\partial^2 ln(L)/(\partial\sigma^2\partial y_{m_j})] = 0$ and $-E[\partial^2 ln(L)/(\partial\sigma^2\partial\mu)] = 0$, the result for $-E[\partial^2 ln(L)/\partial(\sigma^2)^2]$ is superfluous, which simplifies the necessary matrix inversion problem. As expected, these values yield σ^2/n_o as the asymptotic variance for $\bar{x}_o$. The asymptotic variance for the estimate of y_{m_j}, where it is treated as a parameter, becomes $n\sigma^2/[(n-1) - n_m/(n-1)]$. In contrast, treating the estimate of y_{m_j} as the prediction of a new observation yields a variance of $\sigma^2(1 + 1/n_o)$. The necessary t -statistic here is $t_{.975,180} = 1.9732$. Consequently, the estimation intervals in this case are approximately the same, being roughly 40521.8 ± 20536.1.

If a conventional E-M type of analysis is undertaken, where the geo-referenced observations still are assumed to be independent, and the missing values are set equal to $X_m b$ where the regression coefficients are based upon the iterative solution involving both the known and the missing data, then the entries needed to compute the asymptotic standard errors are as follows:

$$-E[\partial^2 ln(L)/(\partial\beta\partial\beta)] = X'X\sigma^2,\ -E[\partial^2 ln(L)/(\partial\beta\partial Y_m)] = -X/\sigma^2, \text{ and } -E[\partial^2 ln(L)/(\partial Y_m\partial Y_m)] = I_m/\sigma^2.$$

Again the result for $-E[\partial^2 ln(L)/\partial(\sigma^2)^2]$ is unimportant, simplifying the necessary matrix inversion problem. The asymptotic variance for the estimate of Y_m, where it is treated as a parameter, is given by the second through $(m+1)$-th diagonal entries of $\sigma^2[I_m - X_m(X'X)^{-1}X_m']^{-1}$. In contrast, treating the estimate of Y_m as the prediction of a new observation yields a variance of $\sigma^2[I_m + X_m(X'X)^{-1}X_m']$; the bracketed expression contains the first two entries of an expansion of the previous matrix inversion. It is not surprising that these two estimates also may be nearly the same, as is found in the results reported in Tables 2 and 3. For this case the necessary t-statistic here is $t_{.975,178} = 1.9734$.

If a spatial E-M type of analysis is undertaken, where the geo-referenced observations now are assumed to be dependent, the missing values may be set equal to those estimates given by equation (4.2). Upton and Fingleton (1985, p. 353) already have provided the following entries needed to compute the asymptotic standard errors for this spatial autoregressive model specification, when a geo-referenced data set is complete:

$$-E[\partial^2 ln(L)/\partial(\sigma^2)^2],\ -E[\partial^2 ln(L)/\partial\rho^2],\ -E[\partial^2 ln(L)/(\partial\beta\partial\beta)],$$
$$-E[\partial^2 ln(L)/(\partial\sigma^2\partial\beta)],\ -E[\partial^2 ln(L)/(\partial\sigma^2\partial\rho)], \text{ and } -E[\partial^2 ln(L)/(\partial\rho\partial\beta)].$$

These particular entries may need to be modified slightly when a data set is incomplete. The additional terms needed are as follows:

$$-E[\partial^2 ln(L)/(\partial\beta\partial Y_m)] = [\rho W_{om}'X_o - (I_m - \rho W_{mm})'X_m]/\sigma^2,$$
$$-E[\partial^2 ln(L)/(\partial Y_m\partial Y_m)] = [\rho^2 W_{om}'W_{om} + (I_m - \rho W_{mm})'(I_m - \rho W_{mm})]/\sigma^2,$$
$$-E[\partial^2 ln(L)/(\partial\sigma^2\partial Y_m)] = \{\beta'[X_m'(I_m - \rho W_{mm}) - \rho X_o'W_{om}] + \rho E(Y_o')[(I_o - \rho W_{oo})'W_{om} + W_{mo}'(I_m - \rho W_{mm})]$$
$$- [\rho^2 W_{om}'W_{om} + (I_m - \rho W_{mm})'(I_m - \rho W_{mm})]E(Y_m)\}/\sigma^4, \text{ and}$$
$$-E[\partial^2 ln(L)/(\partial\rho\partial Y_m)] = \{\beta'(X_m'W_{mm} + X_o'W_{om}) + E(Y_o')[2\rho(W_{oo}'W_{om} + W_{mo}'W_{mm}) - (W_{mo}' + W_{om})]$$
$$+ [2\rho(W_{mm}'W_{mm} + W_{om}'W_{om}) - (W_{mm} + W_{mm}')]E(Y_m)\}/\sigma^2.$$

where

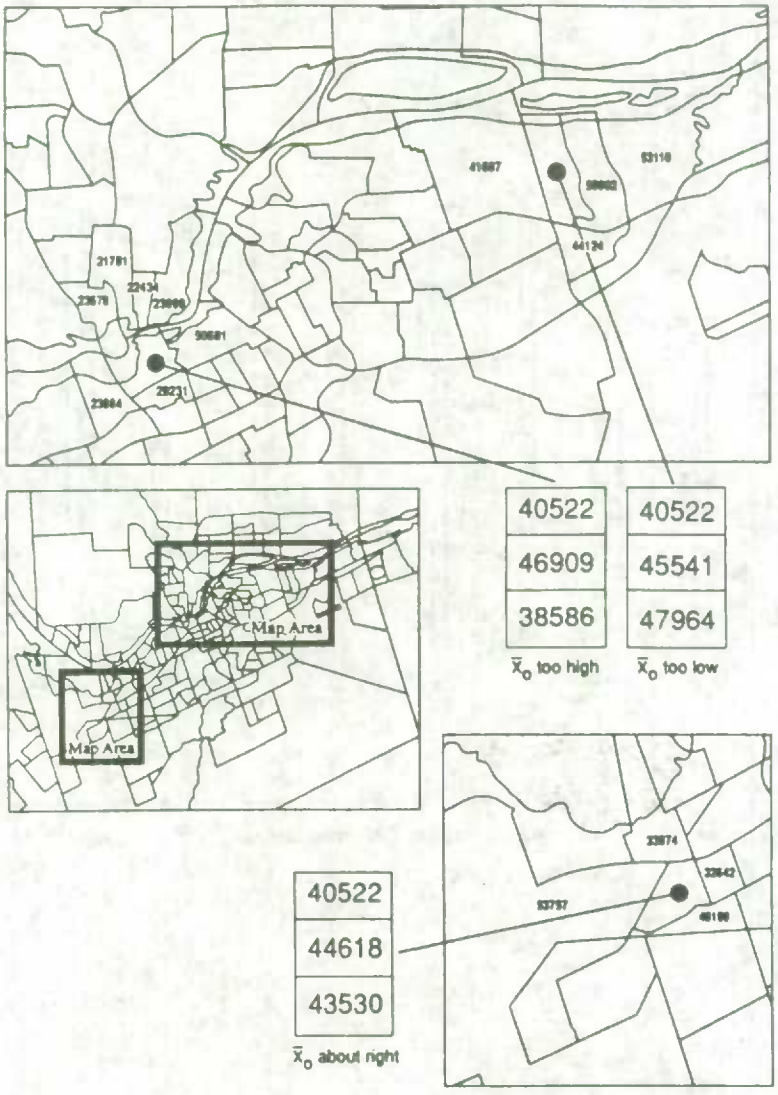
$$E(Y_o) = [(I_o - \rho W_{oo}) - \rho^2 W_{om}(I_m - \rho W_{mm})^{-1} W_{mo}]^{-1}[X_o + \rho W_{om}(I_m - \rho W_{mm})^{-1} X_m]\beta, \text{ and}$$

$$E(Y_m) = \{\rho(I_m - \rho W_{mm})^{-1} W_{mo}[I_o - \rho W_{oo}) - \rho^2 W_{om}(I_m - \rho W_{mm})^{-1} W_{mo}]^{-1} X_o + [(I_m - \rho W_{mm}) - \rho^2 W_{mo}(I_o - \rho W_{oo})^{-1} X_m\}\beta.$$

For this case the necessary t-statistic here is $t_{.975,177} = 1.9735$.

The numerical complexity of this third case, which is illustrated by these analytical results, is attributable to both the presence of non-zero spatial autocorrelation, and non-zero values for $-E[\partial^2 1n(L)/(\partial\sigma^2\partial\rho)]$ and $-E[\partial^2 1n(L)/(\partial\sigma^2\partial Y_m)]$--the matrix inversion is more complex. Therefore, for purposes of making comparisons in this paper, only intervals for the predicted values are presented. A graphic portrayal of these comparisons appears in Figure 2. Of note is that when $\hat{y}_m = \bar{x}_o$, the prediction intervals are constant across the set of missing values; in some cases they are too wide, and in other cases they are too narrow. Variability, and hence uncertainty, are introduced by letting $\hat{Y}_m = X_m b$. Now the intervals are wider for situations further removed from the attribute space mean response. In addition, the estimates themselves are greater than $\bar{x}_o$ in all eleven missing value cases. Finally, equation (4.2) considerably tightens the intervals, yielding estimates of the missing values that are more in line with local geographic contexts.

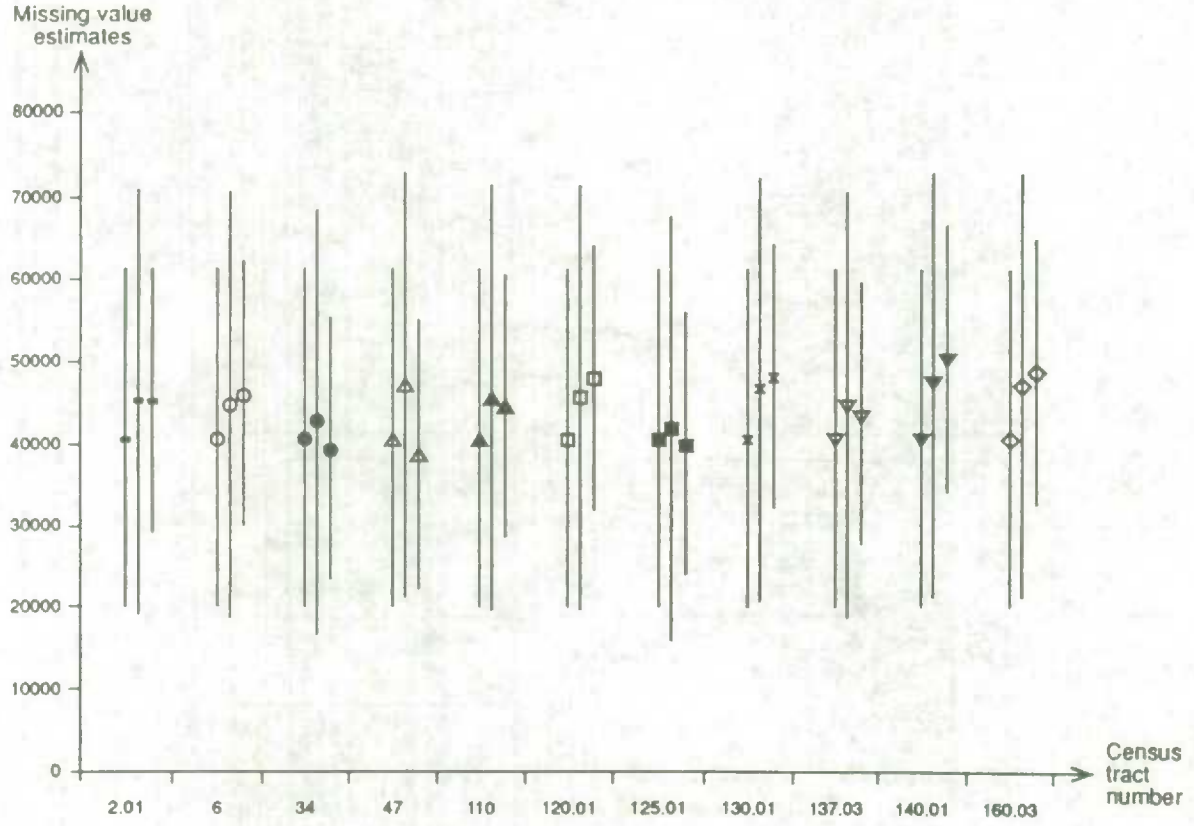
Figure 2: Prediction interval comparisons for the three estimation procedures.



7. CONCLUSIONS

A number of interesting outcomes have been obtained and conclusions have been reached here. First, inclusion of explanatory variables markedly improves the precision of missing value estimates. The earlier study of Houston found confidence intervals that were extremely wide, and in one instance unacceptably wide. That previous research employed a gradient model, which apparently functioned as a surrogate for explanatory factors. In the case of Ottawa-Hull, once population density was included in the model specification, not only did regression diagnostics become reasonably well-behaved, but coordinate terms also became irrelevant. As an aside, the population density measure probably could be enhanced by computing the ratio of people to residential land area, rather than total census tract area. Second, the spatial autoregressive model specification further improves the precision of missing value estimates. This contention is highlighted by Figure 3. The first set of selected estimates illustrate a case where the mean is much lower than the average of surrounding census tract income values, the mean approximately equals the average of surrounding census tract values, and the mean is greater than the average of surrounding census tract values. The second set of estimates are from the traditional E-M type of analysis. Estimates still are not consistent with local geographic contexts; in the first case where the mean is too high, the conditional mean is even more extreme, while in the third case the conditional mean is an improvement upon the mean. The third set of estimates, based upon equation (4.2), more closely reflect their local geographic context; in the first and third cases this is the best estimate of the three available, whereas in the intermediate case this estimate falls between the other two. This type of finding held for the Houston case study, too, where a much stronger positive spatial autocorrelation was uncovered.

Figure 3: Geographic contexts for selected missing values.



Development of the estimation equation (4.2) here is advantageous, especially since it helps supplement the existing array of estimator possibilities for missing geo-referenced data. The approach outlined by Griffith, Bennett, and Haining (1989) is equivalent to kriging, from the theory of regionalized variables (*e.g.*, the exponential model; see Griffith 1991). Kriging relies upon at least four different models (linear, exponential, Gaussian, spherical), finding some more suitable than others for characterizing a particular data set. There is no reason why spatial statistics should not be the same! Hence, effective specific instances of Martin's (1984) general solution need to be developed and their properties explored. Equation (4.2) also is valuable because it allows this technology to be implemented with standard commercial software packages, such as SAS (see Appendices I and II), without being excessively consumptive of computer resources. In contrast, and unfortunately, results for Houston required a considerable amount of FORTRAN programming coupled with IMSL subroutines.

Finally, producing reliable estimates of missing values for urban census data circumvents the problem of confidentiality without compromising it. Statistics Canada can continue to censor sensitive data while making available to researchers essentially complete data sets; this technique is somewhat analogous to the current practice of random rounding use to protect the confidentiality of individual questionnaire responses. The general utility of this approach would benefit tremendously, though, if Statistics Canada conducted internal, undisclosed evaluations of the reliability of missing value estimates. Such an assessment across a national sample, rather than certain urban area tabulations, would be informative without possibly violating the confidentiality of the data. Results obtained for Houston suggest that similar experiments should be done at the United States Bureau of the Census.

REFERENCES

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, in press.

Griffith, D., Bennett, R., and 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 and Planning A*, 21, 1511-1523.

Little, R., and 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., and Fingleton, B. (1985). Spatial Data Analysis by Example, New York: Wiley.

APPENDIX I. SAS CODE FOR THE LINEAR REGRESSION ALGORITHM FOR ESTIMATING MISSING VALUES

```
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
```

APPENDIX II: SAS CODE FOR THE SPATIAL AUTOREGRESSION ALGORITHM FOR ESTIMATING MISSING VALUES

```
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
* START OF THE SPATIAL STATISTICS ANALYSIS *                            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
```

APPENDIX III. MINITAB CODE FOR EQUATION (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
```

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

USING SMALL AREA AND ADMINISTRATIVE DATA TO EXAMINE AGGREGATION EFFECTS IN DEMOGRAPHIC ANALYSIS

C.G. Amrhein[1]

ABSTRACT

The Tax Filer data set from the Small Area and Administrative Data Division of Statistics Canada provides a rich source of data at a sufficiently fine spatial resolution to permit the extensive analysis required to search for various aggregation effects. A previous study examined the nature of an aggregation effect using migration data derived from Canadian income tax data. This study extends Amrhein and Flowerdew's (1990) study by disaggregating the 260 census division flow table by the age and sex of the migrants. The focus is on methodological issues related to aggregation effects, not on providing a model of Canadian migration. Three aggregation algorithms are examined using 130 and 65 regions, and the results are compared to the original 260 region data and provincial data.

KEY WORDS: Aggregation effects; Aggregation algorithms; Migration.

1. INTRODUCTION

The nature and existence of an aggregation effect has been the focus of a small but constant stream of research (see Openshaw 1984 and Amrhein and Flowerdew 1990 for reviews). In a recent study (Amrhein and Flowerdew 1990) using migration data derived from a sample of income tax returns, little evidence of a pronounced aggregation effect was observed. It was suggested that a number of reasons might account for this lack of effect. For instance, since a nonlinear, Poisson model was used, the lack of effect in Amrhein and Flowerdew might be due to the difference between the nonlinear model and the linear models used in earlier studies. Alternatively, the lack of effect might be due to the linear nature of the Canadian urban system, in which each province is dominated by one (or a cluster of) urban area(s) so that as long as the provincial identities are maintained, the migration system is stable. Finally, it is possible that the aggregate nature of the migration data, and the aggregation of a large number of reporting units (260 census divisions), to a smaller number (130, 65, or 10) was convoluting a number of migration processes related to spatial scale (intraurban, interurban, or interregional) and demographic structure (work related, retirement related, and so on). This study addresses one aspect of this last possibility, the aggregation across age and sex.

The importance of age and sex effects as causal variables in migration long has been recognized (see, for example, Shaw 1975; Greenwood 1981; Mueller 1982; Greenwood 1985; and Clark 1986 for reviews) and continues to attract attention (see, for example, Ledent and Liaw 1989; Clark and White 1990; and Haurin and Haurin 1990). This study recognizes the importance of the assumed migration process represented in these works by disaggregating by age and sex the national migration table derived from a sample of the Canadian 1986 income tax returns. A detailed discussion of this data set appears in Flowerdew and Amrhein (1989).

The focus of the following study is to search for an aggregation effect attributable to three factors. The first two factors, the partitioning of space represented by the number of areal units and the grouping of regions represented by the aggregation rules, are examined in Amrhein and Flowerdew (1990). In this study the effect

---

[1] C.G. Amrhein, Department of Geography, University of Toronto, Toronto, Ontario, Canada M5S 1A1. National Center for Geographic Information and Analysis, State University of New York at Buffalo, Amherst, N.Y. 14261, U.S.A.

of partitioning the population by age and sex is combined with the previous two effects. The model used is chosen for its simplicity and parsimony, and is not presented as a complete model of Canadian migration. Far more complex models exist (Shaw 1985 or Flowerdew and Amrhein 1989), and should be used if the purpose is to explain migration. The goal here is limited to the methodological question of aggregation effects.

2. RESEARCH DESIGN

The data are disaggregated by age and sex as well as by census division. As a result, in addition to the questions of scale (the number of reporting units in the analysis, that is 260, 130, 65, or 10), and the rules by which aggregations are created, the additional dimensions of age and sex affect the values of the parameters and statistics used to examine the pattern of migration. Since a detailed comparison of the different aggregation rules is presented in Amrhein and Flowerdew (1990) only a brief outline of the various algorithms is presented here. The four algorithms are:

Poisson Regression Algorithm:

This algorithm is a Poisson regression model adjusted to account for age and sex differences. It is of the form:

$$\ln y_{ij}^{kl} = \beta_o + \beta_1 \ln P_i + \beta_2 \ln P_j + \beta_3 \ln d_{ij}, \tag{1}$$

where P_i is the population of the origin, P_j is the population of the destination, d_{ij} is the distance between census division i and j, y_{ij}^{kl} is the number of people in the k^{th} age category of sex l, and ln is the natural logarithm. Excluding the territories there are 260 regions corresponding to the Canadian Census Divisions. The migration matrix is derived from a sample of income tax returns (The Tax Filer Data Set maintained by the Small Area and Administrative Data Division of Statistics Canada) for 1986. Cells in the 1986 flow matrix were combined as necessary to fit the 1981 census divisions.

Random Neighbour Algorithm (RanNei):

This algorithm is of the form in (1) but operates on fewer, artificially created regions. The synthetic regions are created by first randomly choosing an initial seed region. Next a specified number of contiguous neighbours are chosen to create a new region from original census divisions. The total number of synthetic regions to be created is set by the analyst (in this study, separate analyses are conducted using 130, 65, and finally 10 regions created from the original 260). Problems with islands arising from the definition of contiguity are solved heuristically.

Modified Openshaw Algorithm (ModOpen):

As in RanNei, this algorithm is also of the form in (1) and operates on fewer, artificially created regions. Here the synthetic regions are created by first randomly choosing an initial seed region with a random number of contiguous neighbours chosen to create a new region from original census divisions. The number of contiguous regions used to build a synthetic region varies from one to some specified upper limit. This algorithm is a modification of the algorithm presented in Openshaw (1977).

Maximum Interaction Algorithm (MaxInt):

This algorithm also uses the same procedure as RanNei except that the contiguous region to be merged into the evolving synthetic region is chosen by some specified rule. In the first instance, the contiguous region having the maximum interaction (in terms of migrants) with the seed region is selected. Other rules could certainly be constructed, but are left to future work.

3. RESULTS

The RanNei, ModOpen, and MaxInt algorithms are each analyzed using 130, 65, and finally 10 regions. These experiments are labelled as follows:

Model I: The original Poisson Regression model with 260 census divisions.
Model II: The "provincial model" with 10 aggregations identical to the existing provinces.
Model III: RanNei with 130 regions.
Model IV: MaxInt with 130 regions.
Model V: ModOpen with 130 regions.
Model VI: RanNei with 65 regions.
Model VII: MaxInt with 65 regions.
Model VIII: ModOpen with 65 regions.

A variety of results from this analysis can be reported. The following six tables show the results for Models I and II, and the average results from 100 simulations for each of Models III though VIII. In addition, each of the eight models is analyzed for five age groups (0-15, 16-24, 25-44, 45-64, 65 and older) and for males and females (or a total of (2*5)[2 + (6 x 100)] = 6020 sets of results). Age/sex groups are labelled as follows:

M1 and F1 are males and females, respectively, aged 0-15.
M2 and F2 are males and females, respectively, aged 16-24.
M3 and F3 are males and females, respectively, aged 25-44.
M4 and F4 are males and females, respectively, aged 45-64.
M5 and F5 are males and females, respectively, aged 65 and older.

Results are reported for two summary statistics and the four parameters in equation (1). The results using the aggregate data reported in Amrhein and Flowerdew (1990) for the two summary statistics, and the distance coefficient are repeated here for convenience. The analysis of the results can proceed on at least two levels. The first level relies on a preliminary comparison of the results. The extent to which the results from the various models are different is assessed in a rough context defined by analytical expectations and a straightforward assessment of the meaning that might be attributed to a difference of a given magnitude. Alternatively, given the multinominal nature of the study [four algorithms, number of regions (260, 130, 65, 10), two sexes, and five age categories], an analysis of variance of the results from Models III through VIII can be performed. The tables of results are presented first, followed by the preliminary assessment and finally the ANOVA results (using SAS version 5.0).

Table 1: Proportion of Deviance Explained

	Age/Sex Cohort										
Model	F1	M1	F2	M2	F3	M3	F4	M4	F5	M5	Aggregate
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

Table 2: R square

Age/Sex Cohort

Model	F1	M1	F2	M2	F3	M3	F4	M4	F5	M5	Aggregate
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

Table 3: Constant

Age/Sex Cohort

Model	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

Table 4: Origin Population Coefficient

Age/Sex Cohort

Model	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

Table 5: Destination Population Coefficient

Age/Sex Cohort

Model	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

Table 6: Distance Coefficient

	Age/Sex Cohort										
Model	F1	M1	F2	M2	F3	M3	F4	M4	F5	M5	Aggregate
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. Preliminary Assessment

The main conclusion from a preliminary assessment of the results in the six tables is that only minimal differences among the eight models is evident (compare the rows in Tables 1 through 6). This finding suggests that there is no obvious aggregation effect found in the study of age and sex specific Canadian migration. This may be due to the properties of the Poisson regression model, the algorithms, or the structure of the Canadian urban system. This latter conclusion is consistent with the findings of others (see, for example, Simmons 1980). However, the disaggregation of the population by age and sex does appear to give rise to greater variability than is evident in the aggregate data. For example, in Table 1 the overall performance of the Poisson model, represented by the proportion of deviance explained, does vary more for the various age/sex categories than it does for the aggregate population. Specifically, in Table 1, compare the rows corresponding to Models VI and VII with Model II - the first two columns and the last column. Using the more generally recognized R-square statistic (coefficient of determination) increased variability is evident that might be considered an aggregation effect captured by the statistic. However, the different results created by the R-square and the deviance statistics must be explained, and might be attributable to the difference between a linear and non-linear statistic.

The four parameters (Tables 3 to 6) of the Poisson regression equations reveal patterns consistent with those found in the earlier studies. For example in Table 6, compare the distance coefficients from the various age/sex disaggregated models with the corresponding aggregate results in the last column. In general, the results in each table suggest that in each of these cases, the number of regions appears to generate greater variability than does the aggregation algorithm. This trend is seen in the comparison in each table of the rows corresponding to Models VI, VII, and VIII (65 regions), with Models III, IV, and V (130 regions), with Models I (260 regions) and II (10 regions). This comparison supports the notion of an aggregation effect (Openshaw 1984).

As just seen, a comparison of rows examines the combined effects of the algorithm (RanNei, MaxInt, and ModOpen) used to generate the aggregates, and the number of aggregates in each experiment. In contrast, a comparison of columns in Tables 1 through 6 examines the effect of age and sex on the ability of the model to capture an aggregation effect. For example, in Table 1 the effect of sex and age is seen in the variability among the values in each row. Again the variability, compared to the aggregate results in Model 1, appears greater for the 65 region models (VI, VII, VIII) and the provincial model (Model II), than for the 130 region models. Within each row, the 0-15 age group reflects its dependence on the 25-44 age group as expected. It is also the middle age groups that most closely resemble the aggregate population values. This result is also as expected given the proportion of the absolute flows in the migration matrix represented by this age group. However, since there are so many effects (categories) a more structured analysis is needed, in this case ANOVA in SAS will be used.

3.2 Analysis of variance

An analysis of variance procedure is employed to separate the effects of algorithm, model size, age, and sex of the sub-population analyzed in each of the models. There are 100 observations for each age and each sex for each of Models III through VIII. The original 260 region (Model I) and the provincial models (Model II) are

not included in this analysis since there is only one observation in each case. The results presented below are thus examining the differences among the randomly generated aggregates of the 260 census divisions, not comparing the aggregated partitionings to the original data. Nevertheless, the appearance of an effect due to model size would certainly support the existence of an aggregation effect.

The ANOVA model is structured with three algorithms x two sizes x two sexes x 5 age groups, or 3 x 2 x 2 x 5 = 60 cells. Each cell contains 100 observations so there are 6000 observations in total. Selected results appear below. Among other assumptions required for ANOVA is that the summary statistics and parameter values conform to a normal distribution in the population. The parameters from the Poison regression model are approximately chi-square distributed with degrees of freedom equal to the number of observations minus the number of parameters in the model. Since the chi-square distribution converges on a normal distribution as the number of observations increases, it is assumed that with 6000 observations the assumption of normality holds in the absence of any evidence to the contrary.

When analyzing the ANOVA results, it is important to remember that an observed difference that is not significant with a small number of observations will likely become significant when the number of observations becomes large enough. With 6000 observations, relatively small differences will be significant. In fact, as will be seen below, the ANOVA procedure indicates that almost all of the variables are significant. Finally, extensive analysis of the data set, not reported here, indicates that there are many significant interaction effects among the variables. In addition to the concerns discussed above, the presence of interaction effects should be kept in mind when examining these results.

Table 7 presents the results from the ANOVA model that seeks to explain the proportion of deviance explained by equation (1). The dependent variable is the proportion of deviance explained. The independent variables are the algorithm used to create the aggregation (MOD= RanNei, ModOpen, MaxInt), the sex of the population cohort (SEX=M or F), the age of the cohort (AGE where 1 = 0 to 15 years, 2 = 16 to 24 years, 3 = 25 to 44 years, 4 = 45 to 64 years, and 5 = 65 years and older), and the number of regions in the experiment (R130 = 130 regions and R65 = 65 regions).

Table 7
ANOVA results
Dependent Variable: DEV

Source	DF	Sum of Squares	Mean Squares	F value	PR > F
Model	8	2.1983	0.3648	1345.85	0.00
Error	5991	1.6239	0.0003	Root MSE	
Corrected Total	5999	4.5422		(0.0165)	

Source	DF	ANOVA SS	F Value	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

Given the sample size of 6000, it is not surprising that the model is significant, and that each variable in the model contributes significantly to the explanation of deviance. The ANOVA results, however, tell us only that at least one of the sixty cell means is different from the others. It is useful to determine which of the means are different using a multiple comparisons test for each of the variables. Table 8 presents the Scheffe test for the model tested in Table 7.

Table 8: Scheffe Multiple Range Test

Scheffe Grouping	Mean	N	Model
A	0.7526	2000	ModOpen
B	0.7368	2000	MaxInt
C	0.7348	2000	RanNei
A	0.7614	1200	AGE 4
A	0.7610	1200	3
B	0.7363	1200	2
B	0.7361	1200	5
C	0.7122	1200	1
A	0.7477	3000	R130
B	0.7351	3000	R65
A	0.7481	3000	F
B	0.7348	3000	M

Of technical interest is the finding that each of the three aggregation algorithms is significant and different from the other two, with the Modified Openshaw algorithm (ModOpen) providing the best explanatory power. Ignoring the observation that .75 may not be all that different from .73, the significance of the difference in the performance of the three aggregation rules could be labelled an "aggregation effect." At the very least, it suggests that spatial models are sensitive to the manner in which space is partitioned. This is certainly not a new finding, but it is useful to reiterate occasionally. In the same spirit, the difference attributable to the number of reporting regions in the model also provides evidence of an aggregation effect. Each sex and age group also has a significant and different effect on the explanatory power of equation (1). It is not very surprising that middle and older adults (25 to 64) are most successful in explaining the migration flows since the data represent income tax returns and 25 to 65 year old are the primary wage earning group. In addition, migrants of working age dominate the data file as seen below.

Age group	Males	Females	Total
Child (0 to 15)	164817	156537	322354
Young adult (16 to 24)	120661	128392	249053
Middle adult (25 to 44)	259149	243728	502877
Older adult (45 to 64)	63937	64069	128006
Retired adult (65 +)	23558	33722	57280
Total			1259570

Tables 9 and 10 show the results for the dependent variable RSQ (the coefficient of determination). They are comparable to the results in Tables 7 and 8.

Table: 9
ANOVA results
Dependent Variable: RSQ

Source	DF	Sum of Squares	Mean Squares	F value	PR > F
Model	8	3.7487	0.4686	116.89	0.00
Error	5991	24.0165	0.0041	Root MSE	
Corrected Total	5999	27.7652		0.0633	

Source	DF	ANOVA SS	F Value	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

Table 10: Scheffe Multiple Range Test

Scheffe Grouping	Mean	N	Model
A	0.6603	2000	MaxInt
A	0.6572	2000	RanNei
B	0.6446	2000	ModOpen
A	0.6752	1200	AGE 2
A B	0.6686	1200	4
B	0.6664	1200	3
C	0.6441	1200	5
D	0.6156	1200	1
A	0.6549	3000	R130
A	0.6531	3000	R65
A	0.6640	3000	F
B	0.6440	3000	M

The RSQ values used in the analysis have been calculated from the results of the Poisson regression, they do not represent the results of a separate linear regression experiment. Of primary interest is the lack of significance of the variable SIZE. Other studies (see Openshaw 1984) suggest that a range of values for a statistic can be obtained by modifying areal units. It is therefore not surprising that the RSQ variable fails to be significant given the range of models calibrated here. The variability of the RSQ statistic is large within each set of 100 calibrations, and these ranges overlap across the 60 different sets. This result is apparent in the summary of Scheffe tests. RanNei and MaxInt algorithms perform similarly in this case. Again, since the algorithms differ only in the fashion in which the adjacent neighbour is chosen, this result is not too surprising. Finally, the age groupings are not all significantly different from each other in explanatory power. This result suggests, together with the other tests in Table 10, that the RSQ statistic is less able to distinguish among the range of results than is the DEV statistic. With respect to the age groupings, it is unexpected that young adults have the largest average RSQ value. The next two groups are the same as the first two in Table 8, and together are not significantly different. Young and old adults are also not significantly different. As before, children and retired adults are fourth and fifth in rank.

Finally, Tables 11 and 12 represent the ANOVA and Scheffe results for the distance parameter in equation (1). Results for the other three parameters are similar and are not reported here but are available from the author.

Table 11
ANOVA results
Dependent Variable: COEF4
(the distance coefficient)

Source	DF	Sum of Squares	Mean Squares	F value	PR > F
Model	8	32.2980	4.0373	2655.64	0.00
Error	5991	9.1079	0.0015	Root MSE	
Corrected Total	5999	41.4059		0.0390	

Source	DF	ANOVA SS	F Value	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

Table 12: Scheffe Multiple Range Test

Scheffe Grouping	Mean	N	Model
A	-0.9025	2000	ModOpen
B	-0.9252	2000	MaxInt
C	-0.9317	2000	RanNei
A	-0.8586	1200	AGE 2
B	-0.8665	1200	3
C	-0.8736	1200	1
D	-0.9786	1200	4
E	-1.0217	1200	5
A	-0.8955	3000	R65
B	-0.9441	3000	R130
A	-0.9087	3000	M
B	-0.9309	3000	F

As in the results presented in Tables 7 and 8, each of the 60 cells in Tables 11 and 12 in the ANOVA design are significant and different from other categories of the same variable. The values are negative with the largest absolute values producing the smallest mean values for each Model. The distance parameter (COEF4) provides some interesting interpretations for the age groups. Recall that children are tied to their parents' mobility patterns. The group most sensitive to distance are the retired adults, the least sensitive are the 16-24 year olds, then the middle adults followed by the infants. Older adults are only slightly less sensitive than retired adults. In addition, the 130 region model, with twice as many origins and destinations, includes many more "local" moves than the 65 region model and, as a result, shows migrants more sensitive to distance. Finally, females are more sensitive than males.

Of all the results presented here, the results in Tables 11 and 12 most clearly represent the effect of aggregating space on both the generated statistics (perhaps representing our ability to understand the output of models) and on the description of the behaviour of subgroups of a population. Whether or not the absolute differences are large enough to satisfy some intuitive notion of an aggregation effect, Table 11 and 12 clearly suggest that there is a difference. That is, the spatial resolution affects the modelling results, and these results are affected differently across age and sex specific cohorts. While the findings may be Canadian specific, the sample size is sufficiently large that the reported differences are unlikely to be transient.

3.3 Age/Sex Specific Groupings

Tables 7 through 12 separate the effects of age and sex. Combining these two categories will allow an assessment of whether males and females in the same age group behave differently. Results for the dependent variables DEV, RSQ, and COEF4 are presented in Tables 13, 14, and 15, respectively. The aggregation effect implied is one of demographic rather than spatial aggregation. In particular, compare the Scheffe groupings of the sub-groups when age and sex are separate variables (Tables 8,10,12) to the groupings below. The age/sex groupings are as described before.

Tables 13 suggests that middle and older adults of both sexes are most effective in accounting for the explained deviance in equation (1), with infant and retired males least useful. Age groups 3 and 4 are grouped tightly as in Table 8. With respect to R-square, Table 14 provides similar results with the bottom three groups the same. Of interest in Table 14 is the greater ability of retired females and young adults to account for the observed values of the coefficient of determination. However, there is a greater degree of overlap in the groupings found in this table than in Table 10. This overlap is not surprising given the interaction effect between age and sex, and the significance of these seen before. It does however, suggest that care be taken in examining these results. For example, in Table 13, the first five of the ten groups share at least one category (F3). The conclusion is that the effect of aggregating a population based on age and sex effects is important when examining modelling results.

Table 13: Scheffe Multiple Range Test for Variable DEV where age/sex groups are treated as one variable

Scheffe Grouping	Mean	N	Model
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

Table 14: Scheffe Multiple Range Test for Variable RSQ where age/sex groups are treated as one variable

Scheffe Grouping	Mean	N	Model
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

Table 15 shows the results of the joint age/sex analysis and the distance coefficient. Again, all the variables from the Poisson regression in the ANOVA are significant. These results are very similar to those in Table 12. Older and retired adults in both sexes are more sensitive to distance than are young and middle adults. Young and middle aged males are least sensitive to distance. Interestingly, infant and middle aged females are equally but less sensitive to distance than young females. The close connection between infant and middle-aged females is clear: infants move more with mothers than with fathers. This result is not apparent in Table 12 as a result of combining males and females in age group 3. The other grouping of M4 and F4 is not seen in Table 12. An interesting question is why, when disaggregated, retired males and females (M5 and F5) behave differently? This may be due to different rates of survival among spouses, that is, very old women behave differently, and there are fewer very old men.

Table 15: Scheffe Multiple Range Test for Variable COEF4 where age/sex groups are treated as one variable

Scheffe Grouping	Mean	N	Model
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

Whether or not there is an aggregation effect in Canadian migration is still not entirely clear. A preliminary view suggests that there are clear differences in the results grouped by subset of the population and number of regions, but that these differences are not large absolutely. This assessment must include implicitly the shortcomings of the data and the problems in examining a process-based phenomena such as migration. At the same time, the data provide a large enough base that observed differences are not likely to disappear with additional samples. As such, there is evidence that the results of the Poisson regression model are sensitive to spatial as well as demographic aggregation. Furthermore, these results are sensitive to the fashion in which the combinations of original census divisions are generated. These results imply that the generalizability of the results from spatial interaction models may not be nearly as unconstrained as some may assume. The task is to choose the level of spatial resolution appropriate to the process being studied. The differences in results due to different scales may not be large absolutely, but they are likely to be statistically significant. The appropriate level of spatial resolution is desirable, but this is not necessarily the same as the highest possible level of resolution in every application.

ACKNOWLEDGEMENTS

C. Amrhein acknowledges funding from the National Center for Geographic Information and Analysis (#150-6710A). Data for this study were purchased with a grant from the Institute for Market and Social Analysis, Toronto, Ontario. Computing time was provided by a research grant from the Research Council of the University of Toronto and was subsidized by a grant from the Province of Ontario to the Ontario Centre for Large Scale Computation. Algorithm II was programmed by Felipe Calderon of the Department of Geography, University of Toronto.

REFERENCES

Amrhein, C., and Flowerdew, R. (1990). The effect of data aggregation on a Poisson regression model of Canadian migration. Department of Geography, University of Toronto, available on request.

Clark, W.A.V. (1986). *Human Migration*. Beverly Hills, CA: Sage Publications.

Clark, W.A.V., and White, K. (1990). Modelling elderly mobility. *Environment and Planning A*, 22, 909-924.

Flowerdew, R., and Amrheinc 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., and Haurin, R. (1990). Youth migration in the United States: an analysis of a deindustrializing region. Paper presented to the annual meetings of the Population Association of America. Toronto, May.

Ledent, J., and 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, 1423-28.

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.

SPECIAL INVITED LECTURE

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

THE POTENTIAL FOR SPATIAL MODELS IN THE ESTIMATION ON NONSAMPLING ERRORS

P.P. Biemer[1]

Traditional methods of estimating the components of total mean squared error, such as replicated measurements, record checks, and interpenetrated assignments, are costly and are fraught with inherent data collection problems. Further, the model assumptions forming the bases for the estimates rarely hold in practice so that, in the end, the estimates can only be regarded as rough indicators of the target error components. To avoid the costs and complexities of special field experiments to estimate components of total MSE directly, some researchers have attempted to *model* the uncontrolled experimental variation as a substitute for randomization and/or repeated measurements (referred to as non-experimental methods). Regression estimation and instrument variable estimation are two techniques which have been applied for this purpose. Spatial models present new opportunities to the modeler for providing better estimators of nonsampling error components than can be achieved with other models. Following a review of the literature on non-experimental estimation methods, the spatial modeling approach is discussed in the context of a total census of the population.

The basic idea of the spatial modeling approach is as follows. Let y_{ij} denote the census response from the j-th unit in the i-th enumerator's assignment. Let τ_{ij} denote the true value (or alternatively, the true score) associated with the unit. Assume

$$y_{ij} = \tau_{ij} + b_i + e_{ij},$$

where $b_i \sim (0, \sigma_b^2)$ and $e_{ij} \sim (0, \sigma_e^2)$. Our objective is to estimate σ_b^2 and σ_e^2 without "interpenetrating" enumerator assignments; *i.e.*, by using the usual census assignment of units to enumerators. To do this, the model component $\tau = [\tau_{ij}]$ is reparameterized as $X\beta$ where X is a matrix of known constants and β is a vector of parameters to be estimated. Further, the model error $y - X\beta = \mu$ is modeled via a spatial model which takes advantage of the spatial dependence among neighboring enumerator assignments in a census. Issues of estimability and estimation are also discussed in the paper. Finally, the results of a small simulation study conducted by the U.S. Bureau of the Census are described.

REFERENCES

Biemer, P.P. (1986). A spatial modeling approach to the evaluation of Census nonsampling error, *Proceedings of the Census Annual Research Conference*, Washington, D.C., 7-20.

---

[1] P.P. Biemer, Research Triangle Institute, P.O. Box 12194, Research Triangle Park, NC 27709, U.S.A.

CLOSING REMARKS

Proceedings of Statistics Canada Symposium 91
Spatial Issues in Statistics
November 1991

CLOSING REMARKS

G.J. Brackstone[1]

That brings us to the end of our Symposium 91. I would like to say a few words in conclusion. For those of you who don't know, we have been running a Methodology Symposium in Statistics Canada for many years now and when the topic for this year's Symposium was first suggested, some of us had concerns about whether it was a suitable topic for a Methodology Symposium. We had been previously dealing with more traditional aspects of survey methodology and this was a little bit off the beaten track for us.

I am pleased to say and to acknowledge that our concerns were unfounded. Judging by the size of the registration for this conference, and the participation throughout it most certainly has been a successful symposium from our point of view and I hope that you share that view.

In fact, the registration for the conference was over 400 people. Of course, that includes a lot of people from Statistics Canada, in fact, the majority. But, there were 135 people registered from outside Statistics Canada and, in particular, a good number of people from the United States and some from overseas as well and, in that respect alone, it has been a success.

I think, also, the range of topics that have been covered is quite impressive. We started on Tuesday morning with the excellent presentation from Michael Goodchild which, set the tone for the symposium and laid out some of the problems and some of the challenges. In a sense, we went from that presentation which was from a geographer and introducing some statistical problems, to the final presentation by Paul Biemer, which came from a statistician, a survey methodologist, who is looking at the uses of geographical approaches in a survey methodology problem. In between those end points we've covered most of the functions of a statistical office involved in a survey: collection, frame construction, processing, non sampling error estimation; and we have covered a range of application areas: medical statistics, agriculture, the environment, to name some of them.

I won't even attempt to summarize what we've learnt from the conference. That is for all of you to summarize and take back. But I think a few observations about geography in a statistical agency are appropriate. The role of geography is well established in the context of a population census and is clearly recognized as an essential element in taking a census. In terms of other survey activities, its importance has been less well recognized traditionally. In Statistics Canada, we've been making some efforts, and the geography division has been making efforts to demonstrate the importance and value of geographical tools and geographical concepts in surveys other than the census.

The area which I think is still open for a lot more development, of course, is the analysis area and the geographical or spatial analysis of data collected by statistical agencies. Many of the papers during the last three days have concentrated on these aspects, and have identifed some of the problems that exist with spatial analysis. I hope, at least, that they have raised some interests in people both inside and outside Statistics Canada and perhaps fanned some sparks of interest that could be pursued after this Symposium.

We have heard quite a bit about quality of data. That was actually the theme of last year's symposium. The quality of geographic data is something that we have to bear in mind when we are using it. We heard a lot about the problems, starting with ecological fallacies and various manifestations of that which have been mentioned in the last few days. We've heard quite a lot, too, about programmatic use of geographical data - the use of data to determine benefits to be received by particular areas.

---

[1] G.J. Brackstone, Assistant Chief Statistician, Informatics and Methodology Field, 26'J' R.H. Coats Building, Tunney's Pasture, Statistics Canada, Ottawa, Ontario K1A 0T6.

We have certainly had our experience of that problem in Statistics Canada. I think it was very well referred to in the case of tax treatment in northern areas, and I am glad to hear that a new way of handling it has been found because the old way created problems for us. In other examples, certain areas did not want to be classified as urban so that their residents could be entitled to certain benefits. Sometimes we have other parts of the country that want to be counted as urban because, in their case, the benefits work in the other direction. And we also have cases where representatives of local areas want us not to merge them with other areas, or to merge them with other areas, so that their residents could benefit from particular programs. These are examples of the problems we run into with programmatic use of geographic data.

I would like to say, of course, some words of appreciation to those who were responsible for organizing this meeting. First of all, Mary March who has been the overall co-ordinator for this effort. She has been supported by a organizing team consisting of L. Chatterton C. Weiss, J. Yan and P. Tallon.

Of course there are many other people behind the scenes who have been helping them with the development of the programs, advertising, printing, registration, advertising, catering and reception. I would like to thank the people responsible for organizing this room, Gilbert Gauthier who has been looking after the audio-visual effects to be sure that the right slide showed up with the right speaker. I would also like to thank our interpreters who have been here for three solid days translating everything that has been said into the other official language. Our other sponsors were mentioned at the beginning of this conference: Statistics Canada, and the Laboratory for Research in Statistics and Probability at Carleton University and the University of Ottawa, and the Canadian Association of Geographers. Finally, I must thank all those presenters and chairpersons whose contribution served to make this symposium a success.

We will be producing proceedings from this Symposium as we have done from preceding symposia. We will be continuing with our Symposium series next year when we will be leaving the spatial dimension. We will be going back to the time dimension because we will focus next year's symposium on the design and analysis of longitudinal surveys. This is an area that we have not been greatly involved in at Statistics Canada up to now. But we are becoming increasingly involved in it and we will be looking to benefit from the experiences of people who have been involved in longitudinal surveying in other organisations.

Symposium 91 is closed. Thank you all for your participation and support.