Uwe Deichmann,
Michael F. Goodchild,
and Luc Anselin
National Center for Geographic
Information and Analysis
Department of Geography,
University of California
Santa Barbara, California

Dealing With Errors in Socio-Economic Databases: Selected Findings of a National Research Initiative

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 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.

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.

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 the 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.

An Example: Integrated Regional Modeling in California

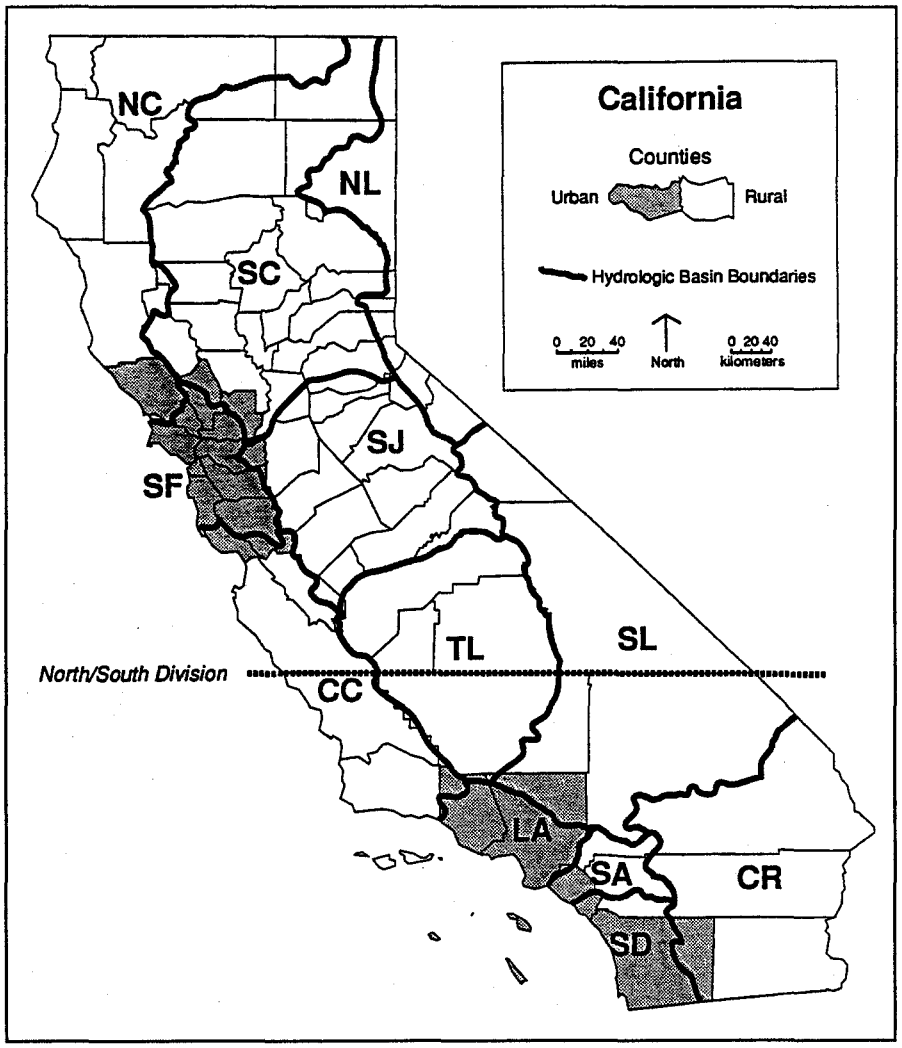
An important part of work in regional science is the development of integrated regional and multiregional models for socio-economic impact analysis (e.g. Isard 1986). An example of such a multiregional modeling effort related to water resources policy in California is outlined in Anselin, Rey and Deichmann (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.

A rather more complicated variation of the problem occurs where the zonal systems are completely or partially incompatible. This

Figure 1
Economic Regions and Hydrological Study Areas



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 5757 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.

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

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

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.

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** edited by M. Goodchild and S. Gopal (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** edited by L. Anselin and M. Madden (London: Belhaven):146-170

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

Arbia, G. 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** edited by M. Goodchild and S. Gopal (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** edited by D.J. Maguire, M.F. Goodchild and D.W. Rhind (New York: Wiley and Sons) 1:191-214

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** edited by M. Goodchild and S. Gopal (London: Taylor and Francis):239-248

Fotheringham, A.S. 1989 'Scale-Independent Spatial Analysis' in **The Accuracy of Spatial Databases** edited by M. Goodchild and S. Gopal (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. (editors) 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** (Zürich) 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** edited by N. Wrigley and R.J. Bennett (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, CA)

Rogerson, P.A. 1990 'Migration Analysis Using Data with Time Intervals of Differing Widths' **Papers of the Regional Science Association** 68:97-106

Tobler, W.R. 1989 'Frame Independent Analysis' in **The Accuracy of Spatial Databases** edited by M. Goodchild and S. Gopal (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

Uwe Deichmann,
Michael F. Goodchild,
et Luc Anselin
National Center for Geographic
Information and Analysis
Department of Geography,
University of California
Santa Barbara, California

Le traitement des erreurs dans les bases de données socio-économiques : résultats choisis d'une initiative nationale de recherche

Introduction

Les systèmes d'information géographiques (SIG) sont de plus en plus appliqués par des géographes, des statisticiens, des planificateurs et des scientifiques de la cognition régionale à titre d'instrument facilitant toute une gamme de formes d'analyses spatiales faisant intervenir des données sociales, économiques et démographiques. Les méthodes nouvelles et l'extension d'applications d'études axées sur des recherches spécifiques à des analyses générales de politiques soulèvent des questions quant à la précision des opérations avec des bases de données spatiales qui n'ont fait l'objet que d'une attention restreinte en analyse spatiale classique. Le problème de la précision des bases de données spatiales est souvent décrit de manière simpliste comme étant un problème du domaine cartographique. Du point de vue de la cartographie, c'est la précision de la position des objets topologiques stockés dans la base de données qui constitue la principale préoccupation, p. ex. dans quelle mesure les lignes de la base de données correspondent-elles aux «vraies» lignes à la surface de la Terre. La précision des bases de données spatiales est cependant également importante du point de vue de la modélisation spatiale ou de la cognition régionale. L'imprécision des bases de données spatiales pose pour plusieurs raisons un problème plus sérieux en traitement numérique des données que ce n'est le cas en cartographie classique. La plus importante de ces raisons tient peut-être au fait que les SIG permettent la création de très grandes bases de données puisant à toute une gamme de sources de données souvent hétérogènes. Une faisabilité accrue sur le plan technique et un partage croissant de l'information spatiale par les institutions rendent nécessaire un examen davantage minutieux des effets de l'intégration de grands ensembles de données spatiales.

Le fait que l'exploitatiuon des SIG soit essentiellement indépendante de l'échelle constitue une deuxième question connexe. L'aptitude à utiliser des données recueillies à diverses échelles entraîne facilement l'introduction d'erreurs reliées à l'échelle. Si un processus spatial est présenté à une échelle, mais est étudié à partir d'ensembles de données basés sur des échelles différentes qui ne conviennent pas, les résultats de l'analyse peuvent être erronés. Enfin, les SIG permettent d'automatiser de manière à les rendre très souples les opérations de l'analyse spatiale. Cela signifie qu'un nombre beaucoup plus grand d'étapes peuvent être franchies dans le cadre d'un projet donné d'analyse spatiale et qu'ainsi il peut devenir impossible de suivre la propagation des erreurs.

Même si cette liste est loin d'être complète (d'autres points sont mentionnés par Goodchild et Gopal, 1989), elle montre que l'utilisation des SIG pour la gestion et l'exploitation de données spatiales aggrave en réalité les problèmes associés aux d'erreurs que doivent régler les analystes de données spatiales. Pour ces raisons, une initiative en recherche a été prise au National Center for Geographic Information and Analysis (NCGIA) dans le domaine de la précision des bases de données spatiales. Elle a pris forme lors d'une réunion de spécialistes tenue en décembre 1988 à Montecito en Californie et ou le programme spécifique de recherche a été élaboré. Bien que cette initiative ait formellement pris fin en novembre 1990 lors d'une série de séances spéciales à la conférence sur les SIG/SIT, un grand nombre des projets individuels de recherche se poursuivent. Dans cet article, nous décrivons brièvement certains des travaux accomplis dans le cadre de cette initiative et nous abordons les problèmes de précision dans la perspective de l'analyse de données socio-économiques. Finalement, nous présentons un exemple tiré d'un effort de modélisation multirégionale. D'après ces expériences, il est soutenu que bien que les SIG peuvent introduire davantage de problèmes, en termes de précision des bases de données spatiales, ils pourraient également constituer un cadre souple permettant à l'analyste de traiter ces problèmes fréquents.

Initiative 1 du NGCIA : précision des bases de données spatiales

Du point de vue de l'analyse spatiale, les recherches qui ont été menées dans le cadre de l'Initiative 1 du NCGIA sur la précision des bases de données spatiales peuvent être regroupées en trois grands domaines :

- Compréhension de la nature des erreurs dans les bases de données spatiales.
- Compréhension des effets des erreurs dans les bases de données spatiales.
- Mise au point de méthodes de réduction ou de gestion des erreurs.

La compréhension de la nature des erreurs exige en tout premier lieu l'élaboration d'une classification commune des erreurs qui tienne compte des implications particulières de diverses structures de données (par exemple tracés en mode-trame versus tracés en mode-vecteur) ainsi que des différents types d'erreurs, comme les erreurs de position versus les erreurs d'attribut (voir Veregin 1989). Une classification commune est une condition préalable à la mise au point de méthodes fiables de modélisation des erreurs dans les divers modèles de données, méthodes par lesquelles le modèle d'erreur ne devrait pas seulement comporter une définition conséquente du type d'erreur, mais devrait également englober des mesures et des statistiques permettant de résumer l'incertitude associée à un ensemble de données (voir Goodchild 1990).

À titre d'exemple, examinons le cas d'une carte pédologique stockée sous forme d'un ensemble de données rastrées. Les limites pédologiques ne sont habituellement pas très bien définies parce que résultant d'une interpolation basée sur des échantillons ponctuels. Cela signifie qu'une incertitude considérable est rattachée à chaque limite. L'une des manières permettant de reconnaître explicitement cette incertitude consiste à attribuer à chaque pixel un vecteur de probabilité d'appartenance à une classe particulière de sol (Goodchild, Sun et Yang 1992). Une application analogue dans le domaine socio-économique serait une recherche sur un marché dans le cadre de laquelle un modèle de type gravitationnel (p. ex. le modèle de Huff) pourrait être utilisé pour prédire des limites bien nettes alors qu'en réalité ces limites sont beaucoup moins nettement définies. Une approche probabiliste comme celle suggérée ajoute une mesure distincte de l'incertitude au produit de SIG tout en le rendant davantage réaliste dans un monde où les relations sont floues. Des travaux connexes ont été menés en rapport avec les limites vectorielles de cartes superficie-classe où un processus de généralisation exigerait essentiellement une vue continue de

l'espace et où la surface représente la probabilité d'appartenance à une classe particulière (Mark et Csillag 1989).

Tel que mentionné dans l'introduction, des successions d'opérations sur SIG mènent à la propagation d'erreurs dans le processus de l'analyse spatiale. L'une des conséquences de ce fait est que des produits libres d'erreurs peuvent être contaminés lorsque combinés à des données renfermant des erreurs. Arbia et Haining (1992) ont mis au point un modèle mathématique théorique du processus de la propagation des erreurs. Une approche plus pratique consiste à mettre au point des indices de la précision des ensembles de données et à les utiliser pour documenter la qualité des cartes dérivées produites en analysant la généalogie des produits dérivés (Lanter et Veregin 1990). Ces travaux de poursuite des erreurs ont des conséquences importantes dans deux domaines : l'analyse des risques associés à la prise de décisions basées sur des produits de SIG d'une précision restreinte et la mise au point de normes de précision pour les organismes (Amrhein et Shut 1990).

Lorsque sont mieux compris la nature et les effets des erreurs dans les bases de données spatiales, l'étape suivante consiste à mettre au point des moyens et des méthodes permettant de réduire ou gérer les erreurs inhérentes à l'exploitation des SIG. Les modèles de représentation de surfaces constituent un exemple de la manière dont différentes structures de données permettent d'améliorer la précision du stockage d'objets spatiaux dans une base de données de SIG. Ces modèles ont été mis au point pour la représentation du terrain, mais ils sont également utilisés pour la modélisation de phénomènes socio-économiques comme les surfaces de densité de la population ou les surfaces de probabilité utilisées en modélisation criminologique. Une bonne compréhension des avantages relatifs de ces structures de données peut permettre de réduire les erreurs introduites par l'utilisation d'une mauvaise spécification du modèle (Kumler et Goodchild 1991).

Les questions reliées à l'échelle spatiale et à l'agrégation spatiale constituent l'un des domaines prioritaires de recherche en analyse spatiale (Amrhein et Flowerdew 1989; Rogerson 1990). Tel que précédemment mentionné, l'une des sources majeures d'erreur en modélisation à l'aide des SIG est le fait que les résultats fournis par les modèles dépendent de l'échelle (Fotheringham 1989). Plusieurs projets de recherche menés dans le cadre de l'Initiative 1 traitaient par conséquent de la mise au point de méthodes d'analyse indépendantes de l'échelle (Tobler 1989) et de l'évaluation des effets de l'agrégation spatiale.

Nous avons indiqué plus haut que l'intégration d'ensembles de données de provenances hétérogènes constitue l'une des raisons pour lesquelles la question de la précision est davantage sensible dans les bases de données spatiales que dans le cas de l'analyse classique. Des ensembles de données nécessaires pour une analyse sont souvent stockés suivant des formats incompatibles. Par exemple, un échantillon ponctuel peut être utilisé avec la superficie couverte par un polygone. Afin de pouvoir utiliser les données dans l'analyse, il faut souvent effectuer une interpolation. Cela signifie que pour l'analyse spatiale au moyen de SIG, de robustes sous-programmes d'interpolation spatiale sont nécessaires pour le traitement de configurations incompatibles de données spatiales (Flowerdew et Green 1989; Deichmann, Goodchild et Anselin 1990).

Dans cette partie nous avons présenté un bref résumé des recherches complétées ou en cours dans le cadre de l'Initiative 1 du NCGIA. Une information plus complète peut être trouvée ailleurs (Goodchild et Gopal 1989; NCGIA 1990; Goodchild 1990). Dans les parties suivantes nous présenterons des réflexions concernant des problèmes spécifiques de précision spatiale rencontrés dans l'analyse spatiale de données socio-économiques.

Implications pour les analyses spatiales faisant intervenir des données socio-économiques

Tout utilisateur de données socio-économiques connaît les problèmes que pose l'obtention du degré d'uniformité nécessaire au niveau des données à utiliser dans l'analyse et la modélisation spatiales. Ces problèmes ont été bien résumés par Wassily Leontief (1986, p. 424) :

> «L'absence de coordination efficace dans le domaine général de la formulation et de la mise en oeuvre des politiques n'a d'égale que l'absence d'un plan général en matière de collecte, de structuration et de présentation des faits et des données dont dépend de manière si critique la prise de décisions tant dans le secteur public que dans le secteur privé».

Dans la pratique, la classification générale suivante des problèmes au niveau des données s'impose : incompatibilités sectorielles/comptables; problèmes temporels; et erreurs spatiales.

Bien que ces trois ensembles de problèmes soient ennuyeux en modélisation régionale, les méthodes mises au point pour les régler sont plus ou moins efficaces. Les problèmes sectoriels et comptables se manifestent dans le processus de la collecte de données en raison d'un manque de coordination entre des organismes (souvent en concurrence les uns avec les autres). Le problème majeur est que les définitions des données socio-économiques changent souvent d'un organisme à un autre. Par exemple, aux États-Unis les définitions de la production industrielle utilisées dans les relevés comptables du Bureau of Economic Analysis diffèrent de celles utilisées par le Bureau of Census même si ces deux organismes font partie du ministère du commerce. Ces problèmes ne peuvent souvent être solutionnés que par des méthodes heuristiques ou intuitives.

Les questions d'ordre temporel ont à l'opposé fait l'objet d'une grande attention, principalement en économétrie et en analyse générale de successions chronologiques. Les mises à jour à intervalles irréguliers des données, les délais entre la collecte et la diffusion des données (p. ex. les relevés comptables nationaux des entrées et des sorties aux É.-U.) et les valeurs manquantes dans les successions chronologiques font partie de ce type de problèmes. Il existe quantité de travaux concernant la prévision des données socio-économiques ainsi que le problème des valeurs manquantes dans les successions chronologiques (voir Vandaele 1983; Granger 1986). De la même façon, le problème des valeurs manquantes dans des ensembles de données spatiales en coupe instantanée a récemment été abordé par Griffith, Bennett et Haining (1989). Cependant, une rationalisation des efforts consentis par les organismes responsables de la collecte des données permettrait d'améliorer considérablement une diffusion opportune et complète d'une grand nombre d'ensembles de données.

Les questions de précision spatiale sont reliées à la nature des erreurs spatiales qui peuvent être attribuables à des erreurs de spécification, à des erreurs de mesure ou à une combinaison des deux (Anselin 1989). Dans un contexte spatial, l'on désigne par erreur de spécification l'utilisation d'un modèle dont l'analyse ne tient pas compte de phénomènes spécifiques à l'emplacement comme la dépendance spatiale ou l'hétérogénéité spatiale (voir l'examen dans Anselin et Griffith 1988). D'une importance spécifique dans ce contexte est la question d'erreurs spatiales d'autocorrélation (à savoir si les erreurs sur une surface spatiale sont uniformément distribuées ou agglomérées). Dans les parties suivantes du présent article nous traiteront principalement les erreurs de mesure. Sous sa forme la plus simple l'erreur de mesure est présente lorsqu'une variable est repérée par rapport au mauvais emplacement géographique. La plupart du temps, cela est attribuable à des erreurs lors du processus de collecte. De toute évidence, des erreurs de position auront une incidence sur les opérations spatiales telles que interpolations, agrégations ou actions tampon. Elles auront de plus une incidence sur les résultats de tests d'effets spatiaux au stade de la modélisation. En plus des erreurs de position, il y a aussi fréquemment des erreurs associées aux attributs stockés

avec les objets spatiaux. Ces erreurs d'attributs peuvent être attribuables à des erreurs de classification ou simplement à des erreurs d'introduction des données.

Les erreurs fréquemment dites conceptuelles (Chrisman 1989) constituent un autre type d'erreur de mesure. Elles ont trait au processus du transfert des entités du monde réel aux objets des bases de sonnées spatiales, ou en d'autres termes, à la manière dont les données sont enregistrées sous formes d'unités spatiales d'observation. Les données socio-économiques sont habituellement recueillies suivant des unités administratives principalement arbitraires qui ne tiennent pas compte des propriétés réelles des distributions des données. Des ensembles inconséquents ou inappropriés d'objets spatiaux utilisés pour la collecte et la structuration des données nuisent fréquemment à une analyse spatiale sensée. Ce problème se manifeste lorsque différents organismes recueillent des données pour différents ensembles de régions, ou lorsque des données sont recueillies à une échelle spatiale, mais nécessaires à une autre échelle, à un niveau moindre d'agrégation. Par exemple, des données peuvent n'être disponibles qu'au niveau national, mais nécessaires au niveau du comté ou du district. En modélisation économique régionale c'est par exemple le cas pour la production régionale brute au niveau du substrat ou pour les comptes d'entrée-sortie publiés.

Ce deuxième type d'erreur de mesure est relié à des questions fondamentales en analyse spatiale qui ont fait l'objet d'une grande attention, mais qui n'ont pas encore de solution satisfaisante. Les données socio-économiques régionales sont habituellement produites en faisant l'agrégation des caractéristiques d'agents économiques individuels (personnes ou entreprises) dans une portion de l'espace (Arbia 1989), ou en compilant des échantillons nationaux par unités de superficie. L'agrégation repose sur des méthodes d'échantillonnage par lesquelles des conclusions pour toute une région sont tirées d'un certain nombre d'observations. Puisque les possibilités d'agrégation d'observations individuelles en groupes d'observations sont considérables, il est possible de soutenir que les conventions spécifiques utilisées pour l'agrégation auront une incidence sur les résultats de l'analyse. Ce problème est appelé *problème de l'unité spatiale modifiable*, et découle, d'après Anselin (1988), du fait que les données sont recueillies pour des unités spatiales ayant des limites arbitraires et irrégulières, ce qui va à l'encontre de la notion générale d'espace homogène impliquant que l'agrégation n'est valide que si les caractéristiques clés sont uniformément réparties dans l'espace. Non seulement le morcellement de l'espace choisi pour l'agrégation, mais aussi le degré d'agrégation influencent les résultats de l'analyse spatiale. Ce problème est appelé *problème de l'erreur écologique* (voir Openshaw et Taylor 1979). D'après Arbia (1989), si les unités sont regroupées en en faisant la somme pour obtenir des unités plus grandes, la moyenne, la covariance et la corrélation varient. Tobler (1989) soutient cependant que les problèmes avec des unités spatiales modifiables et les problèmes dépendant de l'échelle ne devraient pas être attribués à la configuration des données spatiales, mais plutôt à une mauvaise méthode d'analyse.

En pratique, l'analyste n'a que rarement la possibilité d'effectuer un nouveau morcellement de sa région d'étude ou une agrégation à une échelle permettant d'éliminer les phénomènes parasites dans les ensembles de données; il doit donc travailler avec les données disponibles. Il est par conséquent important d'élaborer une notion théorique du problème qui aidera à évaluer l'incertitude rattachée aux résultats à différents degrés d'agrégation. Cela pourrait mener à la mise au point de méthodes d'analyse spatiale qui soient vraiment indépendantes de la configuration spatiale des données ou, comme le dit Tobler, «à une analyse spatiale indépendante du cadre» (Tobler 1989). Afin de se faire une idée des problèmes entravant l'analyse de l'impact socio-économique, il est utile d'étudier un exemple. Dans la partie suivante nous présentons par conséquent un cas d'effort de modélisation multirégionale dans lequel se pose le problème de l'unité spatiale modifiable.

Un exemple : modélisation régionale intégrée en Californie

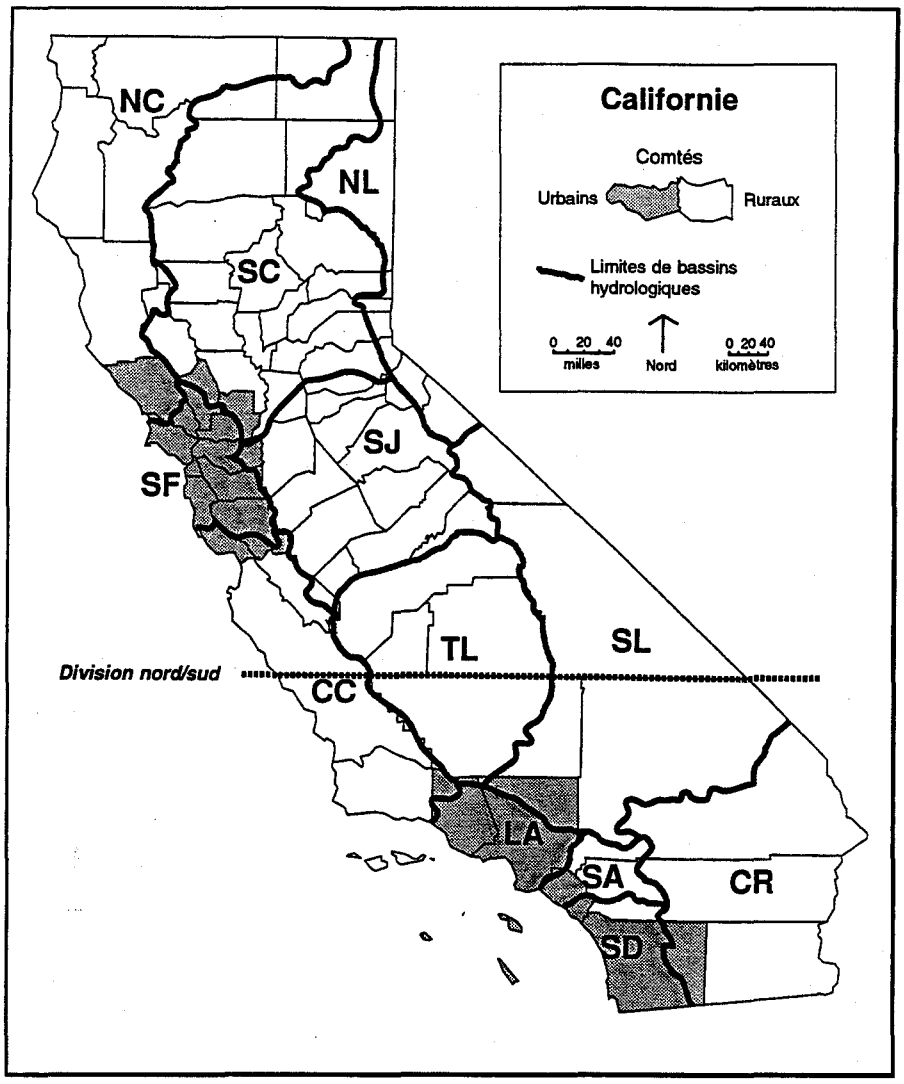
Une part importante des travaux en cognition régionale consiste à mettre au point des modèles régionaux et multirégionaux intégrés pour l'analyse de l'impact socio-économique (p. ex. Isard 1986). Un exemple d'un tel effort de modélisation multirégionale relié aux politiques visant les ressources en eau en Californie est décrit dans Anselin, Rey et Deichmann (1990). Les quatre régions à la base de cette étude sont des agrégats des 58 comtés de Californie représentant une région urbaine et une région rurale, respectivement dans les parties septentrionale et méridionale (figure 1). L'objectif général du modèle est l'évaluation des effets de la dynamique des populations sur les changements de la demande finale des secteurs industriels. Ces changements entraînent à leur tour des changements de l'utilisation de l'eau dans différentes régions et ainsi des modifications de la configuration de l'écoulement de l'eau d'une région à l'autre. Dans le cadre du projet, l'emphase a été placée sur la détermination de l'incidence de différentes stratégies de modélisation sur les résultats du modèle. De manière plus spécifique, l'étude était concentrée sur le choix d'une échelle spatiale faisant intervenir une seule région ou plusieurs régions, sur la sélection d'une approche de mise en relation plutôt que d'une approche d'inclusion des divers modules et sur le problème de l'agrégation spatiale de données zonales imcompatibles (pour les particularités voir Anselin, Rey et Deichmann 1990).

Pendant la collecte des données et la mise au point du modèle, un certain nombre des problèmes mentionnés plus haut se sont posés. Les données par comté pour les sept variables majeures suivantes étaient requises pour l'état de la Californie : emploi; paye, salaires, gains et revenus; production; valeur ajoutée; population; transports et mouvement des marchandises; et approvisionnement en eau ainsi que utilisation et transfert de l'eau. Un examen détaillé de la disponibilité des données et des problèmes qui leur sont associés a déjà été publié (Rey 1988). Nous nous concentrerons ici sur les problèmes spatiaux en cause.

Tel que précédemment mentionné, l'un des problèmes majeurs que pose le travail avec des données à référence spatiale publiées est que différents organismes publient des données concernant des régions ou des zones différentes. Si l'analyse est effectuée à une échelle de très grande agrégation et que les régions s'emboitent les unes dans les autres, le problème peut être réglé par simple agrégation. Si l'analyse est cependant assez détaillée, les données zonales davantage regroupées doivent être décomposées, ce qui exige une méthode d'estimation (p. ex. pour passer de données au niveau national à des données au niveau de l'état). À titre d'exemple mentionnons l'estimation du produit régional brut d'après des variantes de la méthode de Kendrick-Jaycox (p. ex. Weber 1979) dans lesquelles les rapports nationaux de la production sur ses composantes et les données régionales concernant les composantes sont utilisés afin d'estimer la production régionale brute.

Une variante plutôt davantage complexe du problème se manifeste lorsque les systèmes zonaux sont totalement ou partiellement incompatibles. Cela peut se produire lorsque les données proviennent de sources hétérogènes ou lorsque les limites de districts ont changé dans le temps. Puisque les données pour le modèle de l'eau en Californie provenaient d'une multitude de sources, ce problème s'est posé plusieurs fois. Par exemple, l'intégration des données sur l'approvisionnement en eau aux données sur le transfert de l'eau a été compliquée par le fait que pour ces rubriques les données sont publiées pour douze bassins hydrologiques dont les limites sont en grande partie incompatibles avec celles des comtés. D'autre part, les données sur l'utilisation de l'eau par l'industrie sont disponibles par comté. Dans le cadre final d'analyse des activités, il a donc fallu introduire un ensemble de

Figure 1
Régions économiques et régions hydrologiques d'étude



facteurs de pondération traduisant les approvisionnements en eau par bassin hydrologique en demande d'eau par région économique.

Deux méthodes d'interpolation différentes ont été appliquées. La première est une interpolation directe de superficies (Goodchild et Lam 1980), dans laquelle deux ensembles de polygones sont superposés et la superficie des chevauchements pour chaque paire de région source et région cible est calculée. Disposés sous forme de matrice et normalisés, les dénombrements pour les zones sources peuvent ensuite être redistribués sur les zones cibles proportionnellement aux superficies des chevauchements. De toute évidence, l'hypothèse sous-jacente pour cette méthode est que la distribution des variables est homogène à l'intérieur des zones sources et la méthode sera principalement appliquée dans les cas pour lesquels aucune information concernant une variable auxiliaire n'est disponible. Afin d'obtenir une indication de l'erreur potentielle introduite par l'utilisation de pondérations en fonction de la superficie dans une application socio-économique, on a obtenu une deuxième estimation basée sur une carte assez détaillée des centres de gravité de 5757 secteurs de dénombrement avec les populations totales qui leur étaient associées en 1980. Une nouvelle agrégation de ces points en régions économiques et hydrologiques fournit la part de la population de chaque bassin qui habite dans chacune des régions économiques (voir Anselin, Rey et Deichmann 1990 pour les détails). Les deux matrices de pondération sont présentées au tableau 1.

Pour les applications multirégionales multisectorielles ni l'une ni l'autre des deux approches ne devrait être généralement supérieure. Dans les secteurs urbains (comme ceux de la fabrication ou des services), la population fournit une approximation plus judicieuse, alors que des pondérations pour les superficies pourraient sans doute fournir une meilleure approximation pour l'activité agricole. De toute évidence, les configurations d'écoulement réelles prévues avec un cadre d'analyse basé sur la programmation linéaire et l'activité sont très sensibles aux différentes stratégies d'interpolation (voir Anselin, Rey et Deichmann 1990). Dans un contexte de politiques, le choix d'une méthode régionale d'attribution aurait une grande incidence sur les résultats du modèle et ainsi sur des décisions de gestion qui pourraient être d'une grande portée. Étant donné la fréquence des problèmes de dispositions zonales incompatibles, une beaucoup plus grande attention devrait être consacrée à la mise au point de robustes schémas d'interpolation et d'analyses de la sensibilité des résultats de diverses stratégies d'interpolation et de modélisation.

Dans les cas pour lesquels des totaux de contrôle pour une variable auxiliaire ne sont pas disponibles, la dérivation d'estimations pour des zones incompatibles doit reposer sur l'approximation cartographique ou mathématique. Plusieurs autres méthodes ont été proposées en plus de la méthode de la pondération en fonction de la superficie. Dans l'interpolation pycnophylactique (conservant la masse) (Tobler 1979), on suppose une distribution uniforme dans laquelle des emplacements voisins ont des valeurs similaires de la densité. Par cette méthode une approximation de la surface est obtenue en drapant un fin réseau de points sur la région d'étude. La valeur totale pour la variable dans une région est ensuite itérativement distribuée sur les points de manière à ce que le total pour chaque région reste constant et que des points voisins aient des valeurs similaires. Une autre stratégie suggérée par Flowerdew et Green (1989) consiste en une méthode d'estimation statistique par laquelle l'on tient compte de l'information additionnelle disponible concernant la zone.

Les problèmes de systèmes zonaux incompatibles avec le modèle pour la Californie décrit ci-haut ont mené à une généralisation du modèle d'interpolation en fonction de la superficie. Les résultats initiaux ont été présentés par Deichmann, Goodchild et Anselin (1990). Si la distribution dans les zones cibles peut être supposée constante et que le nombre des zones cibles est plus petit que le nombre des zones sources, les densités dans les zones cibles peuvent être estimées sous forme de coefficients d'une courbe de régression passant par l'origine avec les populations des zones sources comme variables dépendantes et les superficies des chevauchements entre les zones sources et les zones cibles comme variables indépendantes. Une ordinaire régression simple des moindres carrés ne guarantit toutefois pas que les coefficients (densités) estimés seront positifs et que la population estimée totale correspond à la population totale connue. Les recherches en cours ont par conséquent été concentrées sur l'utilisation de méthodes de régression avec contraintes (p. ex. Judge et Yancey 1986).

Une extension de cette méthode englobe les cas dans lesquels l'on dispose

Tableau 1
Matrices de pondération pour l'interpolation spatiale

	Facteurs de pondération en fonction de la superficie				Facteurs de pondération en fonction de la population			
	Urbaine sud	Rurale sud	Urbaine nord	Rurale nord	Urbaine sud	Rurale sud	Urbaine nord	Rurale nord
NC	0.0000	0.0000	0.0687	0.9313	0.0000	0.0000	0.4936	0.5064
SF	0.0000	0.0000	0.9998	0.0002	0.0000	0.0000	1.0000	0.0000
CC	0.0222	0.5173	0.0774	0.3831	0.0000	0.4520	0.2424	0.3046
LA	0.9981	0.0019	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
SA	0.2037	0.7963	0.0000	0.0000	0.6003	0.3997	0.0000	0.0000
SD	0.8362	0.1638	0.0000	0.0000	0.9882	0.0118	0.0000	0.0000
SC	0.0000	0.0000	0.0344	0.9656	0.0000	0.0000	0.0489	0.9511
SJ	0.0000	0.0000	0.0225	0.9775	0.0000	0.0000	0.0541	0.9459
TL	0.0018	0.3417	0.0000	0.6565	0.0000	0.3052	0.0000	0.6948
NL	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000
SL	0.0476	0.4875	0.0000	0.4649	0.3175	0.6064	0.0000	0.0761
CR	0.0649	0.9351	0.0000	0.0000	0.0069	0.9931	0.0000	0.0000

d'un troisième ensemble de zones, des zones de contrôle, présentant des densités constantes. Si le nombre des zones de contrôle est inférieur au nombre de zones sources, les densités des zones de contrôle peuvent être estimées de la manière précédemment décrite. La densité des zones cibles peut ensuite être déduite en intégrant la surface de densité de la population représentée par les densités des zones de contrôle sur la superficie de chaque zone cible. Les résultats préliminaires montrent qu'il n'est pas nécessaire que les zones de contrôle soient définies avec très grande précision pour obtenir une amélioration considérable de l'estimation des populations des zones cibles. Dans le cas de l'application californienne, quatre zones de contrôle pour lesquelles on a supposé des densités constantes ont été interactivement introduites dans une couche de SIG «par examen empirique». Ceci pourrait donc être considéré comme l'opinion d'un expert.

Conclusion : l'importance de l'analyse de sensibilité

Parmi les problèmes que posent les efforts de modélisation spatiale, celui de dispositions incompatibles des sources de données est l'un des plus remarquables. Dans les parties qui précèdent, des méthodes de solution de ce problème ont été décrites; certaines de ces méthodes ont été mises au point dans le cadre de l'Initiative 1 du NCGIA sur la précision des bases de données spatiales. Puisque les bases de données spatiales deviennent plus grandes à mesure que progresse la technologie dans le domaine et en raison d'une tendance croissante à la combinaison de bases de données différentes, il devient de plus en plus nécessaire d'intégrer des données de provenances hétérogènes. D'autre part, la technologie des SIG offre la possibilité d'un cadre de mise en oeuvre de méthodes permettant de traiter ces tâches d'intégration. Les questions traitées dans le cadre de l'Inititive 1 du NCGIA sur la précision des bases de données spatiales sont en grande partie recoupées par les sujets de recherche menées dans le cadre de l'Initiative 14 sur l'analyse géographique et les SIG qui doit être lancée au printemps de 1992. L'amélioration des possibilités d'analyse au moyen de SIG exige qu'il existe des outils de manipulation et d'exploration de données permettant à l'utilisateur de concentrer ses efforts sur l'analyse de données confirmatives (voir Anselin et Getis 1992; Goodchild, Haining et Wise 1992).

L'un des atouts majeurs des SIG est le fait que les données y soient stockées de manière à assurer une grande souplesse dans l'adoption d'un cadre spatial d'analyse. Une application particulière découlant de ce caractère fonctionnel est possible lorsque de l'information auxiliaire peut être utilisée pour améliorer des estimations statistiques, de manière analogue à la tendance actuelle à l'utilisation de couches d'un SIG dans une base de données afin de faciliter la classification d'images obtenues par télédétection (Davis et Simonett 1991). L'utilisation d'images classées de l'appareil de cartographie thématique du Landsat pour faciliter une estimation croisée de la population d'une région (Langford, Maguire et Unwin 1990) constitue un pas dans la bonne direction. Ainsi, des zones de contrôle réelles ou subjectives stockées dans un SIG pourraient facilement être utilisées dans la méthode générale décrite par Deichmann, Goodchild et Anselin (1990) et l'interpolation pycnophylactique de Tobler pourrait être modifiée de manière à permettre de tenir compte d'étendues de la région d'étude dont on sait qu'elles ne présentent pas une distribution uniforme. En termes de mise en oeuvre réelle, un SIG pourrait servir de support pour une batterie de méthodes d'interpolation dans laquelle l'analyste choisirait celle qui est la plus compatible avec la nature des données et les hypothèses concernant leur distribution.

Il reste encore beaucoup à faire avant que soit mis au point un système générique de manipulation de données qui permette de travailler avec des données socio-économiques à référence spatiale sans avoir à se préoccuper de l'échelle et de la disposition des unités spatiales. L'on soutient par conséquent que dans l'analyse spatiale une emphase beaucoup plus grande doit être placée sur l'analyse de sensibilité qui devrait être utilisée à tous les stades du processus de la modélisation. Cela implique une analyse menée avec soin des propriétés statistiques des méthodes utilisées, et la mise à l'épreuve d'un certain nombre de configurations spatiales possibles et de méthodes d'analyse, telles que mesures de corrélation croisée ou méthodes d'interpolation. L'objectif devrait être de fournir aux décideurs, qui sont les utilisateurs des résultats de l'analyse spatiale, une forme ou une autre de limites de confiance constituant une indication claire de l'incertitude associée au produit de l'analyse. De toute évidence jusqu'à maintenant l'expérience technique et l'effort de calcul considérables nécessaires en modélisation spatiale ont

nui à la mise au point de telles méthodes. Les progrès accomplis en technologie des SIG et les efforts visant la mise au point de bases de données spatiales et de systèmes d'analyse véritablement intégrés devraient vraisemblablement permettre d'accéder à la souplesse qu'exigent des formes robustes d'analyse spatiale.

Remerciements

Le présent article traite de travaux de recherche actuellement menés au National Center for Geographic Information and Analysis (NCGIA). Nous sommes reconnaissants de l'aide fournie par la National Science Foundation des É.-U. (subvention SES-88-10917).

Bibliographie

Amrhein, C.G. et Flowerdew, R. 1989 'The Effect of Data Aggregation on a Poisson Model of Canadian Migration' in **The Accuracy of Spatial Databases** édité par M. Goodchild et S. Gopal (London: Taylor and Francis):229-238

Amrhein, C.G. et Schut, P. 1990 'Data Quality Standards and Geographic Information Systems' **Actes, SIG pour la prochaine décennie** (Ottawa: Association canadienne des sciences géodesiques et cartographiques):918-930

Anselin, L. 1988 **Spatial Econometrics: Methods and Models** (Dordrecht: Kluwer Academic Publishers)

Anselin, L. 1989 'What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis' Rapport technique 89-4, National Center for Geographic Information and Analysis (University of California, Santa Barbara, CA)

Anselin, L. et Getis, A. 1992 'Spatial Statistical Analysis and Geographic Information Systems' **Annals of Regional Science** 26 (sous presse)

Anselin, L. et Griffith, D.A. 1988 'Do Spatial Effects Really Matter in Regression Analysis' **Papers of the Regional Science Association** 65:11-34

Anselin, L., Rey, S. et Deichmann, U. 1990 'The Implementation of Integrated Models in a Multi-Regional System' dans **New Directions in Regional Analysis: Multi-Regional Approaches** édité par L. Anselin et M. Madden (London: Belhaven):146-170

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

Arbia, G. et Haining, R.P. 1992 'Error Propagation through Map Operations' **Technometrics** 34 (sous presse)

Chrisman, N.R. 1989 'Modeling Error in Overlaid Categorical Maps' dans **The Accuracy of Spatial Databases** édité par M. Goodchild et S. Gopal (London: Taylor and Francis):21-34

Davis, F.W. et Simonett, D.S. 1991 'GIS and Remote Sensing' dans **Geographical Information Systems: Principles and Applications** édité par D.J. Maguire, M.F. Goodchild et D.W. Rhind (New York: Wiley and Sons) 1:191-214

Deichmann, U., Goodchild, M.F. et Anselin, L. 1990 'A General Framework for the Spatial Interpolation of Socio-Economic Data' **Proceedings, Advanced Computing in the Social Sciences Conference** (Williamsburg, VA)

Flowerdew, R. et Green, M. 1989 'Statistical Methods for Inference Between Incompatible Zonal Systems' dans **The Accuracy of Spatial Databases** édité par M. Goodchild et S. Gopal (London: Taylor and Francis):239-248

Fotheringham, A.S. 1989 'Scale-Independent Spatial Analysis' dans **The Accuracy of Spatial Databases** édité par M. Goodchild et S. Gopal (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. et Gopal, S. (éditeurs) 1989 **The Accuracy of Spatial Databases** (London: Taylor and Francis)

Goodchild, M.F. et Lam, N.S. 1980 'Areal Interpolation: A Variant of the Traditional Spatial Problem' **Geoprocessing** 1:297-312

Goodchild, M.F., Sun, G. et Yang, S. 1992 'Development and Test of an Error Model for Categorical Data' **International Journal of Geographical Information Systems** 6 (sous presse)

Goodchild, M.F., Haining, R. et Wise, S. 1992 'Integrating GIS and Spatial Data Analysis' **International Journal of Geographical Information Systems** 6 (sous presse)

Granger, C.W.J. 1986 **Forecasting Economic Time Series** (Boston: Academic Press)

Griffith, D.A., Bennett, R.J. et Haining, R.P. 1989 'Statistical Analysis of Spatial Data in the Presence of Missing Observations: A Methodological Guide and an Application to Urban Census Data **Environment and Planning A** 21:1511-1523

Isard, W. 1986 'Reflections on the Relevance of Integrated Models for Policy Analysis' **Regional Science and Urban Economics** 16:165-180

Judge, G.G. et Yancey, T.A. 1986 **Improved Methods of Inference in Econometrics** (Amsterdam: North Holland)

Kumler, M.P. et Goodchild, M.F. 1991 'A New Technique for Selecting the Vertices of a TIN, and a Comparison of TINs and DEMs Over a Variety of Surfaces' **Technical Papers, 1991 ACSM-ASPRS Annual Convention**, (Baltimore) 2:179

Langford, M., Maguire, D.J. et Unwin, D.J. 1990 'Cross Area Population Estimation Using Remote Sensing and GIS' **Proceedings, 4th International Symposium on Spatial Data Handling** (Zürich) 1:541-550

Lanter, D.P. et Veregin, H. 1990 'A Lineage Meta-Database Program for Propagation of Error in Geographic Information Systems' **Proceedings, GIS/LIS 90** (Anaheim) 1:144-153

Leontief, W. 1986 **Input-Output Economics** 2nd ed. (New York: Oxford University Press)

Mark, D.M. et Csillag, F. 1989 'The Nature of Boundaries on Area-Class Maps' **Cartographica** 21:65-78

NCGIA 1990 'NCGIA 18 Month Report' Rapport technique 90-7, National Center for Geographic Information and Analysis (University of California, Santa Barbara, CA)

Openshaw, S. et Taylor, P. 1979 'A Million or So Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem' dans **Statistical Applications in the Spatial Sciences** édité par N. Wrigley et R.J. Bennett (London: Pion):127-144

Rey, S. 1988 'Data Availability and Data Problems for an Integrated Multiregional Model of California Water Resources' Rapport W-700-88.2, Community and Organization Research Institute (University of California, Santa Barbara, CA)

Rogerson, P.A. 1990 'Migration Analysis Using Data with Time Intervals of Differing Widths' **Papers of the Regional Science Association** 68:97-106

Tobler, W.R. 1989 'Frame Independent Analysis' dans **The Accuracy of Spatial Databases** édité par M. Goodchild et S. Gopal (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' Rapport technique 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