Title
PROBLEMS IN INTEGRATING GEOGRAPHIC DATA

Author
Merrill, D.W.

Publication Date
1987-05-01

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA, BERKELEY

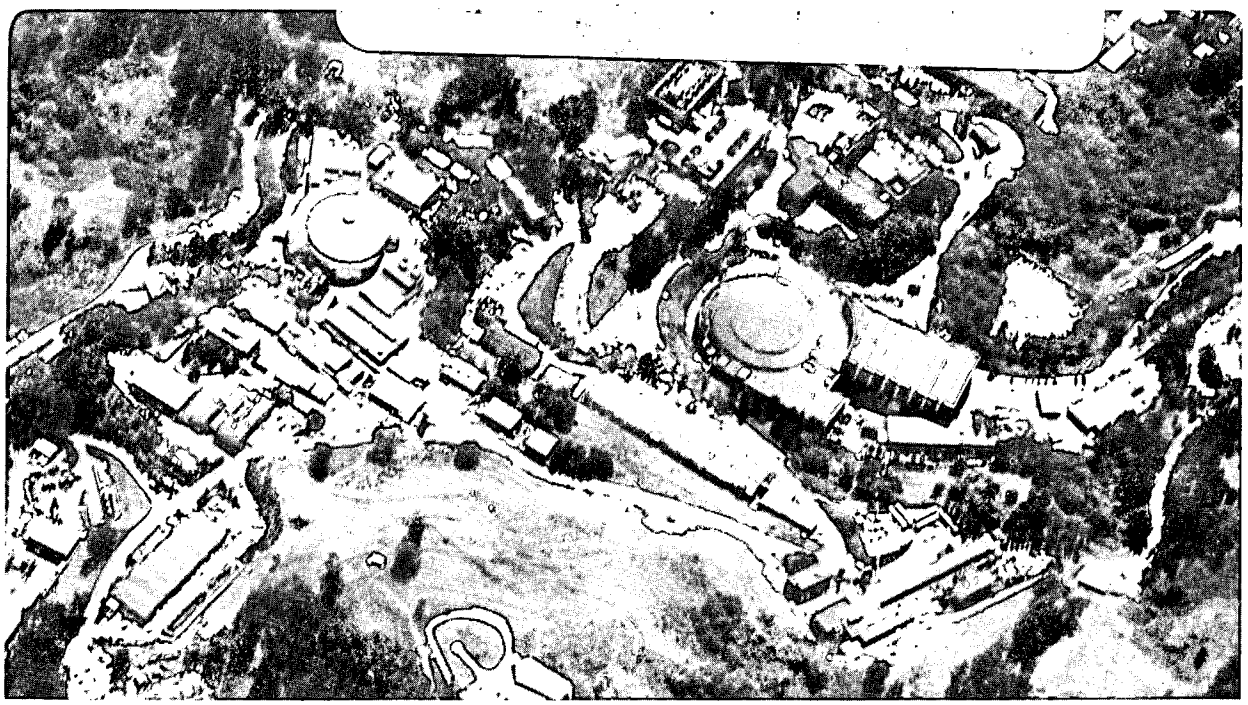
Information and Computing Sciences Division

Presented at "Piecing the Puzzle Together:
A Conference on Integrating Data for Decisionmaking,"
Washington, DC, May 27–29, 1987

PROBLEMS IN INTEGRATING GEOGRAPHIC DATA

D.W. Merrill, Jr.

May 1987



Prepared for the U.S. Department of Energy under Contract DE-AC03-76SF00098

DISCLAIMER

LBL-23601

Problems in Integrating Geographic Data

Deane W. Merrill, Jr.

Computer Science Research Department
Lawrence Berkeley Laboratory
University of California
Berkeley, CA 94720

May, 1987

This research was supported by the Office of Health and Environmental Research, U.S. Department of Energy under contract DE-AC03-76SF00098.

PROBLEMS IN INTEGRATING GEOGRAPHIC DATA

Deane W. Merrill, Jr.
Lawrence Berkeley Laboratory, Berkeley, California 94720

ABSTRACT. Four major types of integrated data systems are: bibliographic, econometric, demographic, and composite. An example of the latter type is Lawrence Berkeley Laboratory's Socio-Economic Environmental Demographic Information System (SEEDIS). Selected SEEDIS data can be downloaded in various self-describing formats, for combination with the user's private data. The problem of combining data at different geographic levels (e.g. 1970 vs. 1980 counties) is discussed. SEEDIS defines over 120 different levels and the relationships between them. Data can be automatically converted from one level to another with the use of geographic correspondence files. Data measured at discrete points, e.g. air quality, can be estimated for population units (e.g. census tracts) by various techniques including interpolation and two-dimensional weighted averaging. Density-equalizing map projections (cartograms) provide a means of analyzing geographic distributions when events are too sparse for reliable calculations of rates.

KEYWORDS

cartogram; data aggregation and disaggregation; data integration; density equalizing projections; geographic data; information systems.

BACKGROUND

LAWRENCE BERKELEY LABORATORY

I would like to describe some experiences of the last fifteen years in attempting to integrate geographically linked data bases at Lawrence Berkeley Laboratory (LBL). LBL is a Department of Energy (DOE) facility operated by the University of California (UC). It is on a hilltop just above the UC Berkeley campus and looks out on San Francisco Bay and the Golden Gate Bridge. Like its sister laboratory Lawrence Livermore National Laboratory (LLNL) 50 miles to the east, LBL is named for Ernest Orlando Lawrence, who invented the cyclotron in Berkeley in the early 1930's. Lawrence was the first of nine LBL Nobel Prize laureates. LBL's major research efforts are in the physical sciences. Unlike LLNL, which is a major weapons laboratory, LBL is closely associated with the UC Berkeley campus and has done no classified work since the 1950's.

SEEDIS

My perspective comes from experience with LBL's Socio-Economic Environmental Demographic Information System (SEEDIS), which has been supported since the early 1970's by DOE, the Department of Labor (DOL), the Army Corps of Engineers, and various other agencies (McCarthy et al. 1982). SEEDIS presently operates in ten DEC VAX computers at DOE and DOL installations, in a nationwide network linked by DECNET. In the future, the Centers for Disease Control (CDC) plan to use SEEDIS as a model for a future national public health information network - a facility to be used by CDC staff and epidemiologists in state and local public health agencies. The front-end user interface will be based in IBM PC-compatible microcomputers and the data archive will reside in CDC's IBM mainframe computer. At a later date the data archive will be copied to optical disks, so users anywhere can access the entire data archive from an inexpensive desktop system.

INTEGRATED INFORMATION SYSTEMS

SEEDIS should be viewed in the broader perspective of integrated information systems, specifically systems of geographically linked data. At the 1981 Integrated Data Users' Workshop I described the most important systems of that type which existed at that time (Merrill 1981). They fell into four separate groups, and the distinction is still valid today.

Bibliographic

The best known group is the bibliographic information systems, of which the largest and most important is DIALOG. These systems contain primarily textual data -- although they do contain some numeric data bases, they were not designed to handle such data, and their computational and analytic capabilities are severely limited. Beyond indexing on a consistent set of keywords, data integration is not a severe problem for the bibliographic systems.

Econometric

The second group is the econometric time series systems, for example those of Data Resources Incorporated (DRI) and Chase Econometrics. These are large systems with thousands of continuously updated time series, and they are accessed daily by thousands of users who need to make up-to-the-minute economic forecasts. The data are well integrated with respect to time, but there is very little geographic detail, and no serious attempt to provide geographic consistency among various data files.

Demographic

The third category is what I call demographic systems. They contain a great amount of geographic detail -- typically down to the ZIP code or block group level. A strong mapping capability is usually included. On the other hand, these systems generally contain only decennial census data, or projections of census data up to the present. So these are not really integrated systems in the sense of this conference, except that 1970 and 1980 census areas are related in order to obtain intercensal estimates. These systems are used primarily by marketing analysts in deciding sites for new retail outlets, shopping centers and the like. Already most of these systems have migrated to single-user microcomputers, with CD-ROM-based mass storage.

Composite

The fourth category -- which I call composite systems -- have truly tackled the spatial integration problem and are the systems of greatest interest for this conference. In 1981, the three most advanced systems of this type were SEEDIS; the Decision Information Display System (DIDS), supported by the Executive Office of the President; and UPGRADE, developed by the Council for Environmental Quality. Unfortunately both DIDS and UPGRADE were discontinued in 1983 due to lack of funding -- in DIDS' case after a very serious but unsuccessful attempt to obtain stable funding from a large consortium of federal agencies.

This leaves SEEDIS as the largest remaining system of its kind. Included are over 100 different data bases acquired from government agencies in connection with various applications projects. In terms of development effort, SEEDIS represents about 100 person-years, approximately one-third in software development, one-third in data compression and installation, and one-third in metadata (structured documentation describing the nature, units, source, reliability of the data, and geographic linkages among the various data files).

DATA IN SEEDIS

In data management circles we hear a lot about "very large" data bases. There are people who describe data bases, usually their own, as "very large" or "enormous", but have no idea HOW large. For SEEDIS the answer is about 22 gigabytes, the equivalent of 220 tapes at 6250 bpi, or about 20-40 optical disks, depending upon the manufacturer. This corresponds to about 29 billion numeric data values, which is a more meaningful measure. The average value of less than one byte per data value was achieved through efficient data compression techniques (run-length encoding and multiple index files). Major SEEDIS data files are listed in Table 1.

Table 1. Major Databases in SEEDIS

File Name		megabytes	mega-values
1980 Census (total)	*	6,400	5,600
Summary Tape File 1	*	430	200
Summary Tape File 2	*	1,200	1,700
Summary Tape File 3	*	1,400	500
Summary Tape File 4	*	3,300	3,200
Equal Employ Opportunity	*	30	50
Master Area Reference File	*	50	5
1970 Census (total)		9,000	1,600
Counts 1,2,4,5,6		7,500	1,300
Public Use Sample		1,000	200
Industry/Occup Employment		500	100
Miscellaneous		6,100	21,700
1973-81 SEER Canc Incidence	*	50	400
1968-78 Tabulated ICD8 Mort	*	80	11,000
1979-84 Tabulated ICD9 Mort	*	70	8,000
1950-84 Unit Rec Mortality		2,500	1,500
1970-80 Current Pop Survey		1,000	300
1964-82 Cnty Bus Patterns		400	80
1976 Survey Income & Educ		500	100
Other		1,500	320
Total		21,500	28,900

* databases stored in compressed format.

Source: adapted from: Databases in SEEDIS, Computer Science Research Department, Lawrence Berkeley Laboratory.

Table 1. Major SEEDIS databases as of May 1987, with the size of each in megabytes and "mega-values" (millions of data values).

THE FUTURE OF INTEGRATED SYSTEMS

Given the changes in the technology and economics of computing, large integrated systems like SEEDIS will probably not be created in the future. More likely, data vendors will integrate small-area data sets for specific applications, drawing from individual data files maintained on mainframes or distributed on optical media. For example, the Topologically Integrated Graphic Encoding and Referencing (TIGER) system developed for the 1990 Census exceeds 200 gigabytes. It will probably be stored in its entirety in few places outside the Census Bureau, let alone integrated with other data.

THE FUTURE OF SEEDIS

I view SEEDIS as a national trust. Its most important role is the preservation and transmission of historic data and metadata, which will be valuable for future epidemiological and sociological research. Only LBL purchased nationwide coverage of the 1980 Census Summary Tape Files (STF) 1 through 4, to say nothing of the 1970 Census and 100-odd other files added since the early 1970's. Through the courtesy of various users, SEEDIS also acquired 20 years of air quality data, complete mortality data back through 1962, and cancer mortality data back through 1950. The latter data are no longer available through public sources. The day is not too far off when copying and storing a mere 22 gigabytes of data will cease to be a data processing nightmare, and the SEEDIS data archive can be moved to modern mass storage media.

PROBLEMS IN INTEGRATING DATA

Next I will discuss a few specific problems frequently encountered by analysts in integrating geographic data: data management problems, inconsistent geographic boundaries, interpolation of point data, and rates of discrete events. In dealing with these problems, the SEEDIS staff has developed some useful concepts and techniques, which will be described.

DATA MANAGEMENT PROBLEMS

Inexperienced data analysts are invariably astounded to discover how much of their time is spent in obtaining, cleaning up, and organizing their data, and how little time is left over for the actual data analysis. An efficient data management system is absolutely essential, one in which the user can quickly and interactively review and correct data. The associated metadata, descriptive information about the data values themselves, must be carried along and appropriately updated at the same time as the data. There are a number of systems that perform the task adequately. The best choice is generally whatever system is already available and comfortable for the user. SEEDIS provides data to the user in a self-describing, eye-readable, flat-file format. Tools are provided for automatic conversion of this format to Statistical Analysis System (SAS), Statistical Program for the Social Sciences (SPSS-X), and Data Interchange Format (DIF). The latter can be converted by vendor-supplied software to Lotus 1-2-3 or dBASE III format.

INCONSISTENT GEOGRAPHIC BOUNDARIES

Another difficulty in integrating geographic data is the problem of inconsistent geographic definitions. Differences in nomenclature or code definitions are easily solved, but changes in geographic boundaries over time pose special problems. In Figure 1 we illustrate a typical situation for some counties and independent cities in Virginia. Between the 1970 and 1980 Census the independent city of Poquoson broke away from York county; during the same period Nansemond county joined with the independent city of Suffolk. Clearly the 1970 and 1980 definitions of York county are inconsistent, as are the 1970 and 1980 definitions of the independent city of Suffolk. In publications and data files such as the City-County Data Book data from both censuses are coded as "Poquoson" or "Suffolk". Some users and data base administrators in the past have neglected to read the explanatory footnotes and, even worse, have attempted to interpolate between the two census data points.

Figure 1. Geographic boundary changes

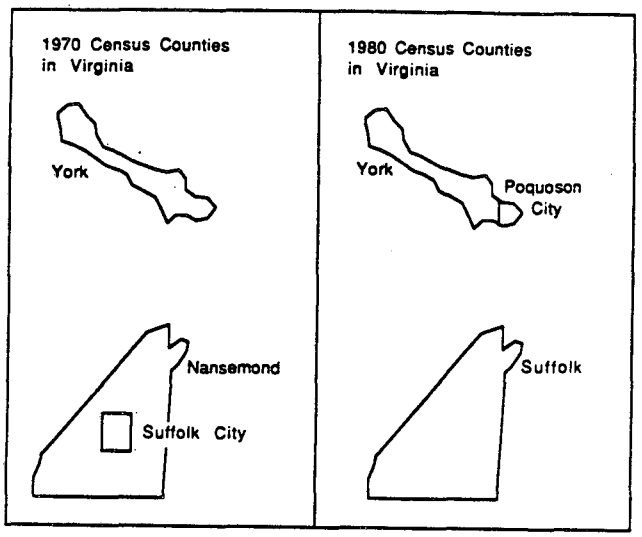


Figure 1. An example of geographic boundary changes. Between the 1970 and 1980 Censuses the independent city of Poquoson broke away from York county; during the same period Nansemond county joined with the independent sity of Suffolk.

SEEDIS protects users against such obvious blunders by clearly differentiating between geographic levels, for example 1970 and 1980 Census county definitions, or Standard Metropolitan Statistical Areas (SMSA's) as defined in different years. In all over 100 different geographic levels are defined in SEEDIS, some of which are listed in Table 2.

The correspondence among various geographic levels is defined in SEEDIS, permitting data at one geographic level to be aggregated to other coarser levels. Similarly, data can be disaggregated to finer levels, with the use of appropriate proportionality assumptions.

Table 2. Geographic Levels in SEEDIS

M	Level	Units	Description
*	NATION80	233	Nations, FIPS 1980
*	FED	10	Federal Regions
*	STATE	55	States + DC + territories
*	AQCR	247	Air Quality Control Regions
*	SMSA81	323	Std Metro Stat Areas, 1981
*	PUS70	408	Public Use Sample county groups
*	COUNTY	3141	Counties, 1970 Census
*	COUNTY80	3137	Counties, 1980 Census
*	NCHS	3082	Counties, Nat Ctr Health Stat
*	NCI	3061	Counties, Nat Cancer Institute
*	MSP	3075	Counties, J Hopk Mort Surv Prog
*	AQMS	6625	Air Quality Monitoring Stations
*	PLACE	12000	Places, 1970 Census
	PLACE80	22450	Places, 1980 Census
*	MCD80	35000	Minor Civil Div, 1980 Census
*	TRACT	35000	Tracts, 1970 Census
*	TRACT80	48000	Tracts, 1980 Census
*	TRACT80PT	99000	TRACT80/MCD80/PLACE80 pieces
*	BGED70	250000	Block Grps + Enum Dists, 1970
*	EDBG80	320000	Block Grps + Enum Dists, 1980

* mapping capability available in SEEDIS

Source: adapted from: Geographic Levels in SEEDIS, Computer Science Research Department, Lawrence Berkeley Laboratory.

Table 2. Major geographic levels in SEEDIS, with the number of geographic units in each.

INTERPOLATION OF POINT DATA

Data measured at point locations, such as air quality, pose special data integration problems. In Figure 2 are shown 1974-1976 levels of total suspended particulate from measurements at individual air quality monitoring stations in California. It may be required to compare such data with population data, which correspond to geographic areas. In some applications, analysts have simply assigned to the county the average air quality of stations within the county, or the value measured at the nearest station. A better method involves interpolation -- constructing a smooth "surface" passing through all the measured values -- but this is inappropriate for data like those in Figure 2, where quite different values have been measured at the same or nearly the same location. A simpler method is a two-dimensional weighted average, where the weight assigned to a given station is some function of the distance (and possibly direction) from the point of interest. A number of reasonable weighting functions have been suggested (Johnson 1983). The smooth weighted average can then be integrated over the area of interest, or simply evaluated at the population centroid. Perhaps a better technique, though not so simple, is "kriging" -- a minumum-variance regression estimate which, unlike a simple weighted average, does not ignore correlation between adjacent points.

Figure 2. Air quality in California, 1974-1976

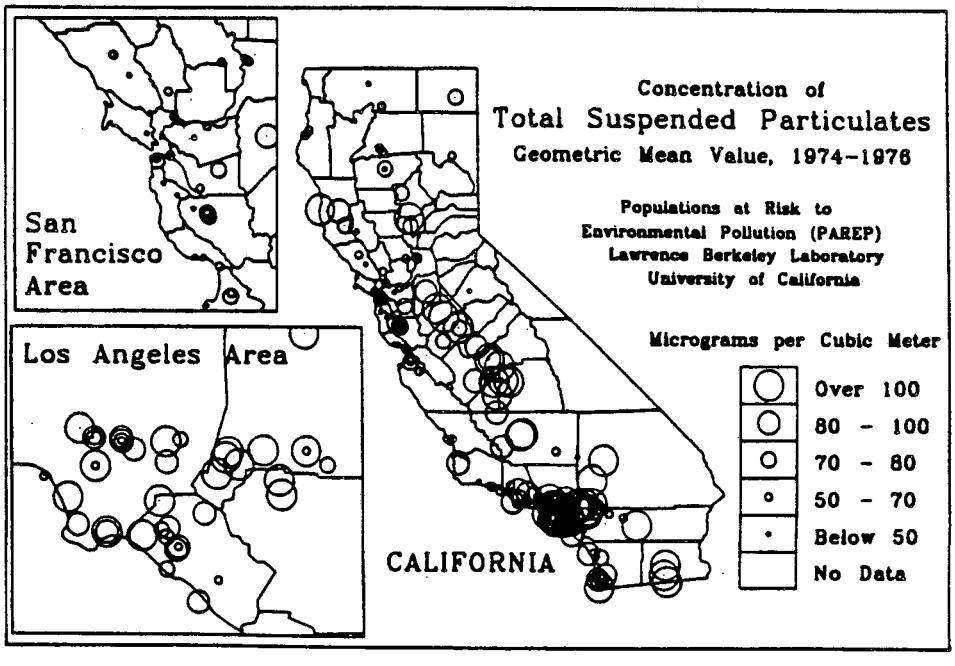


Figure 2. Levels of total suspended particulate in California, from measurements at individual air quality monitoring stations in 1974-1976. The geometric mean values of measurements from individual stations are indicated by circles of varying sizes. Note the presence of concentric circles in the Los Angeles area, corresponding to differing measurements at the same location.

RATES OF DISCRETE EVENTS

The geographic variation of discrete events, such as deaths or cases of disease, poses difficult analytic questions. Local health officials are frequently asked to evaluate the statistical significance of a reported disease cluster. Even ignoring the difficulty of demonstrating causality, the measurement of statistical significance is greatly complicated by the variation of population density. Conventional analytic methods involve calculation of rates for individual subareas, but this is impractical if the number of events is small. Moreover, correlations between rates in adjacent subareas are not easily incorporated into the analysis.

One approach to this problem is the use of density-equalized maps, or cartograms, which have been used to analyze health data since the 1920's. Recently the technique has been computerized, permitting its integration with quantitative spatial analysis techniques (Schulman 1986, Selvin et al. 1987a, Selvin et al. 1987b). The boundaries of individual subareas, for example states, are adjusted so that after the transformation the area of each state is proportional to its population. The locations of individual events, for example deaths, are changed by the same transformation. After the transformation, population density is equal over the entire map, so that boundaries of subreas can be ignored in the subsequent analysis. An example of a computer-generated density-equalizing map transformation is shown in Figure 3.

Figure 3. Density-equalizing map transformation

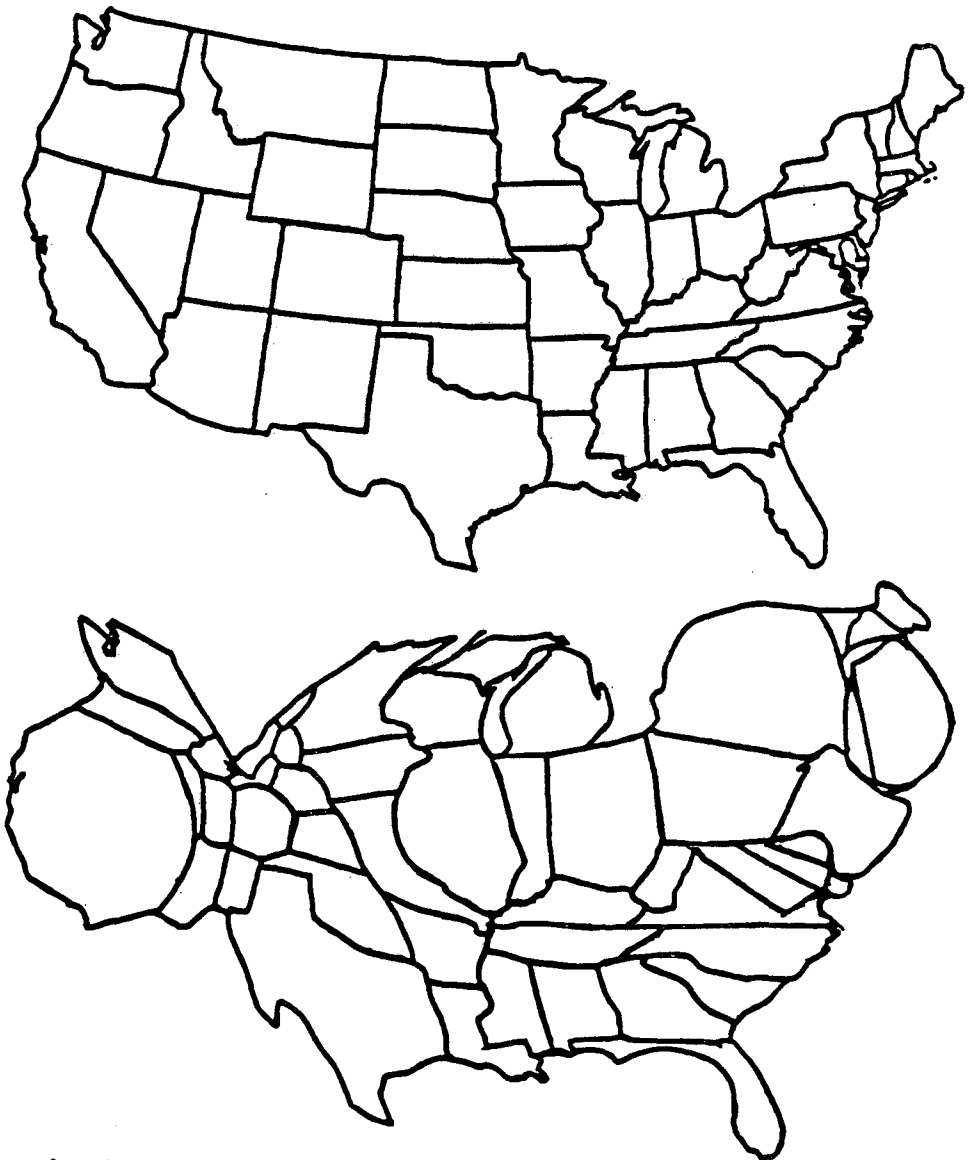


Figure 3. An example of a density-equalizing map transformation. The upper part of the figure is a customary geopolitical map of the United States. In the lower part of the figure the boundaries of the states have been transformed to make their individual areas proportional to population.

ACKNOWLEDGMENTS

DEC, DECNET and VAX are the registered trademarks of Digital Equipment Corporation. IBM PC is the registed trademark of International Business Machines Corporation. DIALOG is the registered trademark of Dialog Information Services, Inc. Lotus and 1-2-3 are the registered trademarks of Lotus Development Corporation. SAS is the registered trademark of SAS Institute, Inc. SPSS-X is the registed trademark of SPSS Inc. dBASE III is the registered trademark of Ashton-Tate.

This research was supported by the Office of Health and Environmental Research, U.S. Department of Energy under contract number DE-AC03-76SF00098.

REFERENCES

Johnson, L. D. 1983. The geographic and statistical analysis of air quality data in the United States. Ph.D. thesis. University of California, Berkeley.

McCarthy, J., W. Benson, A. Yen, D. Merrill, A. Marcus, F. Gey, H. Holmes, and C. Quong. 1982. SEEDIS: A Research and Development Project on Social, Economic, Environmental and Demographic Information Systems. Computer Graphics World 5(6):35-44.

Merrill, D. 1981. Integrated Analysis Systems. pp. 5-18. In Proc., Integrated Data Users Workshop, Reston, Virginia. Oak Ridge National Laboratory report CONF-8110199.

Merrill, D. 1982. Problems in Spatial Data Analysis. pp. 218-223. In Proc., Seventh International Conference of SAS Users Group International, San Francisco, California.

Merrill, D. and J. McCarthy. 1983. CODATA Tools: Portable software for managing self-describing data files. pp. 245-251. in Computer Science and Statistics: Twelfth Annual Symposium on the Interface, Houston, Texas.

Schulman, J. 1986. The statistical analysis of density-equalized map projections. Ph.D. Thesis. University of California, Berkeley.

Selvin, S., G. Shaw, J. Schulman, and D. W. Merrill. 1987. Spatial distribution of disease: three case studies. J. Nat. Cancer Inst. (in press).

Selvin, S., D. Merrill, J. Schulman, S. Sacks, L. Bedell, and L. Wong. 1987. Transformations of maps to investigate clusters of disease. Soc. Sci. Med. (submitted).

For further information, please contact:

Deane W. Merrill, Jr.
Building 50B, Room 3238
Computer Science Research Department
Lawrence Berkeley Laboratory
Berkeley CA 94720
Telephone: 415-486-5063
Electronic mail:
ARPANET: DWMerrill@lbl.arpa
BITNET: MERRILL@LBL
DIALCOM: 64:DOE1209
ASYNC: 329-1020

LAWRENCE BERKELEY LABORATORY
TECHNICAL INFORMATION DEPARTMENT
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720