Title
INTEGRATED DATA BASES IN ENVIRONMENTAL HEALTH APPLICATIONS

Author
Merrill, D.W.

Publication Date
1989-06-01

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

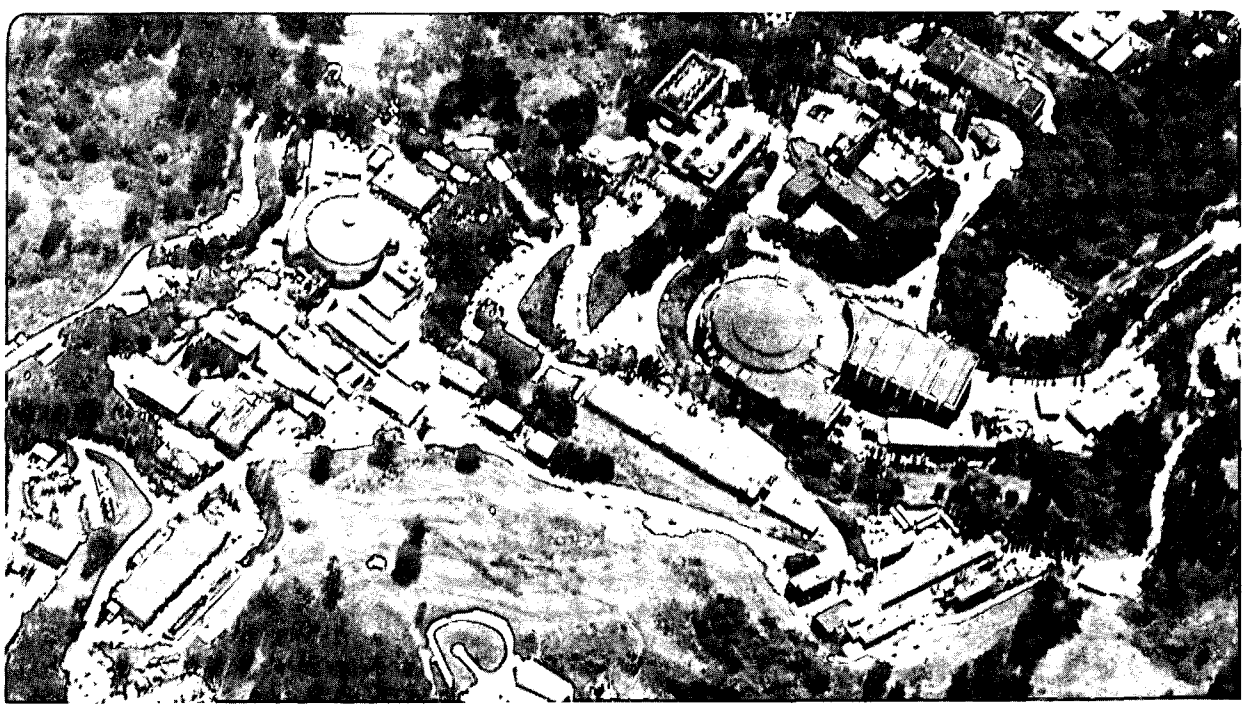
Information and Computing Sciences Division

To be published in the Proceedings of the Environmental Databases Workshop, San Antonio, TX, June 19, 1989

Integrated Data Bases in Environmental Health Applications

D.W. Merrill

June 1989



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

DISCLAIMER

LBL-27436

Integrated Data Bases in Environmental Health Applications

Deane W. Merrill, Ph.D.

Information & Computing Sciences Division
Computing Sciences Research & Development
Lawrence Berkeley Laboratory
1 Cyclotron Road
Berkeley, California 94720

June 1989

Proceedings of the Environmental Databases Workshop, San Antonio, Texas, June 19, 1989, sponsored by the Agency for Toxic Substances and Disease Registry (ATSDR), the Centers for Disease Control (CDC) and the Association of State and Territorial Health Officers (ASTHO).

This work was supported by the Director, Office of Energy Research, Office of Health & Environmental Research, Human Health & Assessments Division, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

INTEGRATED DATA BASES IN ENVIRONMENTAL HEALTH APPLICATIONS

Deane W. Merrill, Ph.D.
Information and Computing Sciences Division
Lawrence Berkeley Laboratory
Berkeley, California

ABSTRACT

Increasingly, public attention is focusing on the health effects of environmental hazards including air pollution, radioactive and toxic wastes, and indoor radon. Typical exposures are at low levels, and have a geographic distribution that can be estimated *a priori.* In such cases, analysis of routinely collected health surveillance data (*e.g.* vital statistics, cancer incidence or birth defects data) can be faster, cheaper and more sensitive than specially designed case-control or cohort studies. The health data are linked by geographic location with independently collected census data and environmental exposure data. The units of analysis are geographic areas rather than individuals; such data are known as ecologic data.

In analyzing such data, rates calculated for individual subareas are not a suitable measure of risk. If the subareas are too small, stable rates cannot be calculated; if too large, geographic detail is lost. In either case one cannot easily describe trends which vary smoothly across adjacent subareas. We describe a preferred alternative, the mathematical technique of density equalizing map projections.

Because ecologic data often cover broad areas and long time periods, the same data find repeated use in different applications. Comprehensive data files are generally too large and complex for commercial data base systems. Conventional analytical techniques are not generally valid for ecologic data. Below the county level, geographic data integration is a difficult task requiring the use of geographic base map files. For all these reasons it is cost effective to permanently integrate major ecologic data files in a publicly shared computing environment.

The role of integrated data bases (and ecologic data) in environmental health applications:

An important area of environmental research involves the human health effects of environmental hazards. In particular, the potentially harmful effects of toxic wastes and nuclear radiation, and of high radon concentrations inside homes, have recently caused widespread public concern. This paper discusses the design and application of integrated data bases which contain not only environmental data, but also health, demographic, and socioeconomic data needed for the study of environmental health effects. Such data bases normally contain summary data, known as *ecologic* data.

What are ecologic data?

Ecology is the science of the relationships between organisms and their environment. Epidemiologists define *ecologic data* as summary data which describe a group of individuals rather than a number of individual cases and controls. The two definitions are related, because environmental conditions pertaining to an individual are not known unless

personal monitoring data have been collected. Generally, environmental data *are* ecologic summary data.

Health and demographic data, on the other hand, are generally collected for individual cases and controls, along with personal risk factor data such as diet and smoking habits. Environmental exposure estimates can be appended to the individual records; for example the estimated air pollution concentration of a census tract can be appended to the records of all individuals living in that region, and the records can be individually analyzed. However, if one is interested in the health effects of the *environment*, one can often reduce costs without losing essential information by analyzing, instead, average tract-level values of environmental, health, demographic and personal risk factor data. Ecologic data records have no personal identifiers. Instead, records are identified by categorical descriptors, for example "white males, age 35-44, living in Los Angeles county and dying of lung cancer in 1975."

Why study ecologic data?

As noted above, except for personal monitoring data, environmental data *are* ecologic data. Personal environmental monitoring data are extremely expensive to obtain. The same is true for health and risk factor data that must be collected by interview for a specific case control or cohort analysis. On the other hand, routinely collected surveillance data (including environmental monitoring data, health surveillance data, and census data) are readily available and inexpensive to the end user. To protect the privacy of individuals, census data and health data are provided *only* as ecologic summary data. Such data are generally comprehensive, covering long time periods and broad geographic areas. Ecologic data can provide a large number of cases even for rare diseases, and they can be used to retroactively analyze phenomena that were unsuspected at the time of their occurrence.

How to study ecologic data?

Special statistical techniques are required for the analysis of ecologic data. The "ecologic fallacy," *i.e.* attempting to apply ordinary statistical techniques such as multiple regression analysis to ecologic data, can lead to erroneous conclusions. As in any study, results must be interpreted with caution. Potential confounders like smoking habits are generally unmeasured in ecologic data, and can be estimated only approximately from measured socioeconomic variables.

Ecologic data require special data management techniques that are not available in commercial data base management systems (DBMS). A DBMS for ecologic data must efficiently handle multidimensional data arrays (*e.g.* deaths by age, sex, race, county, year, and cause of death). It must efficiently store large and very sparse data arrays. It must intelligently manage the *metadata, i.e.* the textual and structural information *about* the data. It must provide transparent access to files which are too large for permanent on-line storage, and files which are too large for one physical storage device.

Because ecologic data are often comprehensive in time and space, they find repeated use in different applications. Geographic integration of dissimilar files is an expensive task that should not be repeated more often than necessary. Hence it is cost effective to integrate data files in a publicly accessible network environment; or to furnish multiple copies of the entire data base on high-volume media (*e.g.* CD-ROM); or both.

Geographic data problems:

Some environmental data, for example air quality data, are measured at specific point locations. Other data, for example water runoff or soil type or land use data, pertain to

irregularly shaped areas determined by geologic features. Ecologic health and demographic data, on the other hand, are generally provided for geopolitical units like counties or census tracts. Sophisticated geographic data integration techniques are required; simple standardization of geographic codes is not sufficient. Even in combining purely geopolitical data, for example 1970 Census data and 1980 Census data, the use of standard geographic codes is inadequate due to changes of geographic boundaries over time.

Tools for geographic integration:

The basic tool for geographic data integration is the *Geographic Base File (GBF),* which describes monitoring locations, boundaries of geopolitical units, *etc.* in terms of latitude-longitude coordinates. A GBF is required for every geopolitical unit having associated data. The GBF must correspond to the boundaries in use at the time the data were collected; a typical ecologic analysis might require data (and corresponding GBF's) from the 1970, 1980, and 1990 Censuses.

Useful for geographic *aggregation* are *geocode correspondence files* which tell, for example, which census tracts belong to a particular county. In the case of boundaries which mutually overlap, for example states and Metropolitan Statistical Areas (MSA's), one needs to consider the smallest undivided regions, for example the Illinois portion of the Chicago MSA.

Geographic *disaggregation* is more difficult and requires assumptions of proportionality. For example, one might wish to estimate tobacco sales in the Chicago SMA, knowing only tobacco sales by state. One might assume tobacco sales to be proportional to population. In this case one needs to know the population separately for the MSA and non-MSA portions of Illinois and Indiana. Or, one might assume sales to be proportional to the over-18 population. Total population and over-18 population, for state parts of MSA's, are two different variables that might be included in a *proxy variable file.*

In addition to the geographic files just described, software tools are required, for example to overlay GBF polygons, to convert data from grid to polygon format, to interpolate point measurements, or to integrate functions over irregular geographic areas. Many of these functions are available in commercial Geographic Information Systems (GIS). However, the geographic data files required for useful analysis are generally lacking or prohibitively expensive.

The representation of time-varying geographic data is largely an unsolved problem. In particular, GBF's, geocode correspondence files, and proxy variable files must provide correct relationships at any point in time without requiring redundant storage. To avoid loss of information, data in general should be archived in terms of the geographic units for which they were collected; necessary transformations should be performed only as needed for integration and analysis.

Density Equalizing Map Projections (DEMP):

The analysis of geographic disease clusters, or of disease distributions in general, demands specialized statistical techniques. In simply analyzing the locations of cases, one improperly ignores variations in population density. Calculating rates for geographic subareas is likewise unsuitable. If the subareas are too large, geographic detail is lost; if too small, meaningful rates cannot be calculated. Most subareas will have no cases and hence a rate equal to zero; a few will have one or two cases and hence a meaningless rate. A measure is needed that reflects the geographic relationships among the observed cases, however few, adjusting for varying population density.

The use of density equalizing map projections (DEMP), or cartograms, to analyze disease distributions was proposed as early as the 1920's. In such an analysis, case locations are plotted on a map which has been continuously transformed so as to make population density everywhere equal. Under the null hypothesis of equal risk, cases should be randomly distributed over the entire transformed map; the null hypothesis is rejected if any geographic pattern is observed that is inconsistent with chance variation. The implementation of the DEMP transformation, and subsequent testing of the null hypothesis, are computationally challenging problems.

The DEMP transformation is not unique. Fortunately, *any* DEMP is suitable for testing the null hypothesis. Conversely stated, if risk is truly equal, *no* DEMP that is truly density-equalizing can introduce a spurious geographic pattern inconsistent with random variation. However, to aid in recognition and interpretation, one wants a DEMP that distorts the original map as little as possible. Further theoretical work is needed in this area.

Lawrence Berkeley Laboratory (LBL):

To conclude, we summarize progress at Lawrence Berkeley Laboratory (LBL) in integrating and analyzing ecologic data. LBL is a Department of Energy (DOE) installation in Berkeley, California, operated by the University of California. Part of the LBL and DOE mission is the study of geographic disease distributions, in particular around energy-related facilities or in areas where high indoor radon concentrations have been reported.

Populations at Risk to Environmental Pollution (PAREP) project:

The Populations at Risk to Environmental Pollution (PAREP) project is an ongoing collaboration between LBL and the University of California, Berkeley, School of Public Health. The PAREP project has been supported since 1978 by the DOE Office of Health and Environmental Research (OHER). Project activities include biostatistics and computer science research, training and support of graduate students, and maintenance of a productive computing environment for epidemiologic research. The PAREP project has pioneered the use of the DEMP technique in the analysis of geographic disease distributions.

Socio-Economic Environmental Demographic Information System (SEEDIS):

A major resource of the PAREP project is the SEEDIS information system, developed with multi-agency support and greatly enhanced by the PAREP project. The result of 100 person-years of development, SEEDIS is the largest existing integrated data base of health, demographic, and socioeconomic data. The data base includes about 100 billion data values, reduced by data compression techniques to about 10 billion characters. For purposes of environmental health studies, SEEDIS complements but does not duplicate major environmental data holdings in the Environmental Protection Agency (EPA), Agency for Toxic Substances and Disease Registry (ATSDR), Centers for Disease Control (CDC), and other government agencies. For handling large multidimensional data bases, SEEDIS has implemented sophisticated data management techniques that surpass those available in commercial systems.

Major databases in SEEDIS:

SEEDIS contains approximately 200 major data files, fully documented and geographically integrated. LBL is the only institution, public or private, holding complete 1970 and 1980 U.S. Census data down to the census tract level. Other important data include 1970 and 1980 census tract GBF's, intercensal population estimates, cancer mortality since 1950, all-

cause mortality since 1962, and cancer incidence data from the Third National Cancer Survey (TNCS) and the Surveillance, Epidemiology, and End Results (SEER) program.

Major SEEDIS geographic levels:

SEEDIS contains GBF's and geocode correspondence files for over 100 geographic levels, including nations, regions, states, counties, minor civil divisions, places, Metropolitan Statistical Areas (SMA's), air quality control regions, air quality monitoring stations, water resource areas, state economic areas, 1970 and 1980 census tracts, and many others. Separate definitions are included to reflect changes in boundaries over time.

World Health Data Base (WHDB) project:

A new initiative is a proposed US-USSR collaboration known as the World Health Data Base project. The purpose of the project is the promotion of joint epidemiological research activities involving the United States and the Soviet Union. Present efforts center around electronic communication and establishment of a bilingual (English-Russian) electronic mail facility. A long-term objective is creation of a bilingual information system containing comprehensive health and environmental data from both countries. Future extensions are envisaged which will incorporate additional languages and data from other countries.

For more information, please contact:

Deane W. Merrill
Building 50B, Room 3238
Information and Computing Sciences Division
Lawrence Berkeley Laboratory
Berkeley, California 94720
Tel: 415-486-5063
FTS: 451-5063
FAX: 415-486-5401
Electronic mail:
Internet: DWMERRILL @ LBL.GOV
Bitnet: MERRILL @ LBL
Compuserve: 71001,62
also: TELEX, DIALCOM, MCIMAIL, and San Francisco-Moscow Teleport

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

LAWRENCE BERKELEY LABORATORY
TECHNICAL INFORMATION DEPARTMENT
1 CYCLOTRON ROAD
BERKELEY, CALIFORNIA 94720