Title
A LBL PERSPECTIVE ON STATISTICAL DATABASE MANAGEMENT

Author
Wong, H.K.T.

Publication Date
1982-12-01

LBL-15393
Excerpt
c.2

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

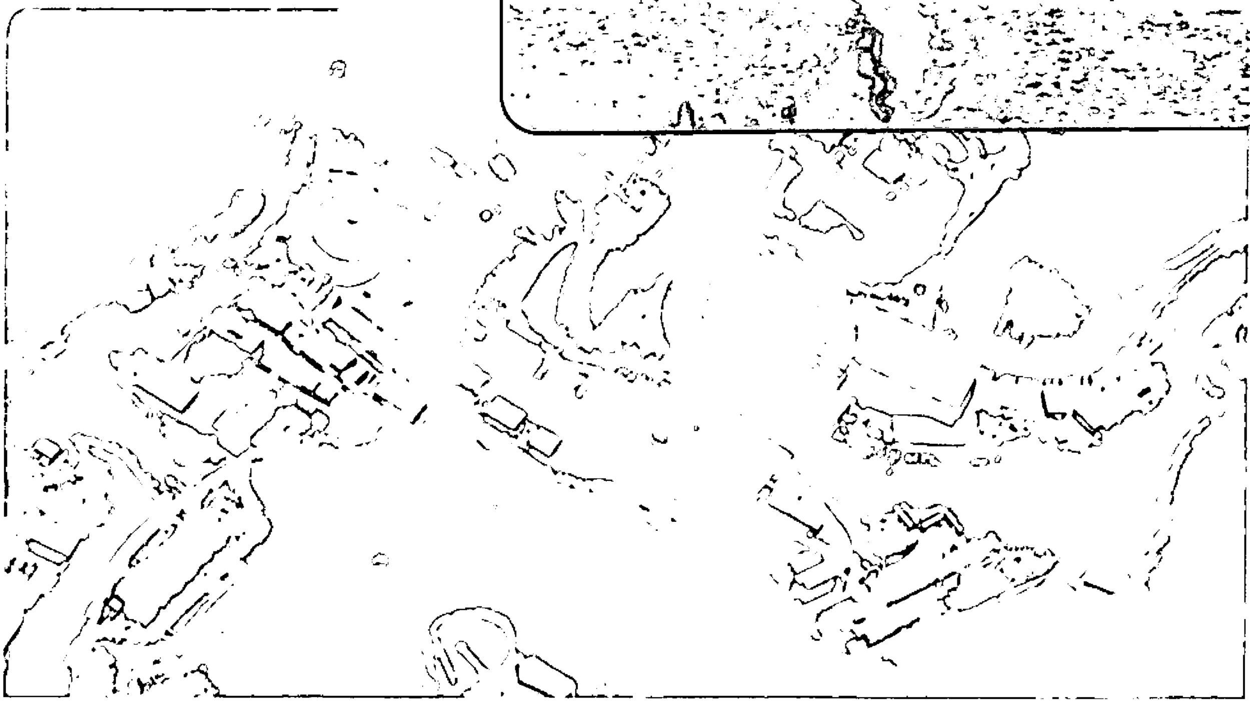
Physics, Computer Science & Mathematics Division

POPULATIONS AT RISK TO ENVIRONMENTAL POLLUTION (PAREP): PROJECT OVERVIEW, 1976-1982

A Chapter in A LBL Perspective on Statistical Database Management

D. Merrill and S. Selvin

December 1982



LBL-15393 excerpt c.2

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

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

LBL-15393
Excerpt

POPULATIONS AT RISK TO ENVIRONMENTAL POLLUTION (PAREP): PROJECT OVERVIEW, 1976-1982

A Chapter in A LBL Perspective on Statistical Database Management

Deane Merrill and Steve Selvin

Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

December 1982

This work was supported by the Director, Office of Energy Research, Office of Basic Research Sciences, Division of Engineering, Mathematical and Geosciences of the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

Populations at Risk to Environmental Pollution (PAREP): Project Overview, 1976-1982

Deane Merrill and Steve Selvin
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

Table of Contents

1. Introduction

Populations at Risk to Environmental Pollution (PAREP) is an ongoing project involving the Computer Science and Mathematics (CSAM) Department at Lawrence Berkeley Laboratory (LBL) and the School of Public Health (SPH) at the University of California, Berkeley (UCB).

The PAREP project was initiated in 1976 under funding from the Environmental Protection Agency (EPA); since 1978 it has been supported by the Office of Health and Environmental Research (OHER) of the Department of Energy (DOE). In 1980 supplementary funding was provided by the Electric Power Research Institute (EPRI). Later, activities gradually expanded to cover additional broad research areas, including the analysis and interpretation of epidemiological phenomena, the critical evaluation of statistical methodologies, and the development of computer science techniques for data management and display.

2. History

2.1. 1976-1978 Getting Started

Until 1978, the PAREP project was called PARAP, or Populations at Risk to Air Pollution. The original mandate of the project, defined in 1976, was the creation of an integrated county-level data base containing mortality data,

Prepared for the U.S. Department of Energy under Contract DE-AC03-763F00098.

population counts, socioeconomic indicators, and estimates of air quality for the United States. Like its predecessor, the POPATRISK database created for the EPA by System Sciences Inc. [G1], PARAP was to be incorporated in the SYSTEM 2000 database management system in the EPA Univac 1100 computer.

The integration of the population and socioeconomic data was a simple task, given the availability of these data in the 1970 Census databases at LBL. In addition, two separate county-level age-adjusted mortality files were obtained: 1950-1969 cancer mortality from Thomas Mason of the National Cancer Institute, and 1968-1972 mortality from Herbert Sauer Associates (courtesy of Sigma Data Computing Corp). Differences in county definitions, especially in Virginia, presented problems which had been anticipated in advance. These were solved by writing special purpose programs.

The major defect of the POPATRISK database, and all available county-level air quality databases, was that air quality for each county was estimated simply as an average of measurements from stations within the county, taking no account of the actual population distribution or the locations of the monitoring stations. Such an approach satisfied EPA's regulatory needs but was totally inadequate for producing estimates of populations at risk to various levels of pollution, which were required for epidemiological research.

The PARAP project proposed to estimate pollution levels for each county at its population centroid, as a weighted geometric mean of values from nearby monitoring stations. For a given county centroid, the weight assigned to each station was f*exp(-.5*(d/dO)**2), where f is the fraction of time the station was active, d is the distance in kilometers between the county centroid and the monitoring station (which may be outside the county), and dO is a constant, empirically chosen to be 20 (kilometers). The form of the weighting function and the choice of dO are the subject of continuing study [P15, P18].

The method involved calculation of population centroids from the 1970 MEDX file, and required information about the geographic coordinates of the air quality monitoring stations. Shortly after the project was begun, gross errors were discovered in the geographic coordinates of the air quality monitoring stations (see Fig. 1). Rather than abandon the novel estimation approach and merely duplicate previous unsatisfactory work, the PARAP staff decided to produce an accurate air quality monitoring station directory. The task required laborious comparison of numerous files from independent sources. Unfortunately, adequate data management tools were not available.

Other difficulties were encountered. Support of the SYSTEM 2000 database management system (DBMS) on the LBL computers was discontinued, necessitating running on unfamiliar distant computers. The second-choice DBMS, the Berkeley Database Management System (BDMS), proved to be too bug-ridden and expensive for use in the PARAP project. In addition, LBL's Control Data Corporation (CDC) computers were becoming hopelessly overloaded.

In the face of these difficulties, it was decided in 1978 to move the PARAP project to a DEC VAX computer which had just been acquired by the Computer Science and Mathematics (CSAM) department. Existing DBMS's on the VAX were inadequate for a major application project, so new software had to be written. Fortunately, just when EPA funding could no longer be continued, DOE was able to begin supporting the project. The project name was changed to PAREP, to accommodate DOE's interest in other forms of pollution, including low-level radiation.

2.2. 1978-1980 System Development

Work was already under way to move portions of SEEDIS, the Socio-Economic Environmental Demographic Information System to the VAX. (For SEEDIS documentation, see [S1] through [S14], especially [S8].) Beginning with the move to the VAX in 1978, SEEDIS and PAREP became almost synonymous. Major SEEDIS enhancements were direct responses to PAREP needs, and PAREP databases were the first databases installed in the new VAX SEEDIS. PAREP applications continue to account for about 15 percent of total SEEDIS usage (see Fig. 2).

Urgently needed were simple data management tools which could conveniently handle simple self-describing files larger than a few megabytes. Deane Merrill wrote a set of routines, the Codata Tools, to operate on files in the Codata format used by SEEDIS [S12]. Following the Software Tools philosophy, the tools are modular, are written in RATFOR (a transportable FORTRAN preprocessor), and follow the UNIX conventions of standard input and output. Accurate and complete documentation is readily available. The Codata tools can be used to extract specified rows and/or columns from a file, to sort a file, to perform relational joins, to perform tabulations by aggregating on common key values, and the like. Though inefficient and limited in their capabilities, they have proved simple and robust enough to become widely used outside the PAREP project. SEEDIS, which is largely written in Digital Command Language (DCL), makes liberal use of the Software Tools and Codata Tools.

Another need was a consistent set of geocode files and geographic correspondence files which could be used to automatically combine data from different geographic levels. The PAREP project put the first few such files in SEEDIS; as of December 1982 the list has expanded to include 82 different geographic levels [S6]. Metadata concepts related to aggregation and disaggregation have been further developed by Chan and Shoshani [G8] and Johnson [G9].

PAREP and SEEDIS require efficient data compression schemes. A preliminary compression algorithm, implemented by Fred Gey, Bob Healey, and Edna Williams for SEEDIS, has spurred further research by Susan Eggers [G10].

Most of 1978-1980 was spent in developing SEEDIS and installing in PAREP data bases to be used in subsequent analyses. County level data on mortality, population, socioeconomic characteristics, and air quality were completed for the entire United States (Figs. 3 and 4). In addition, cancer incidence data from the 1969-1971 Third National Cancer Survey were tabulated and installed at the census tract level, along with corresponding data on population, socioeconomic characteristics, and air quality (Figs. 7 and 8).

Some prototype analyses were performed and published with the use of preliminary data ([P1] through [P8]). The most noteworthy of these was a study of air quality and cancer incidence in the San Francisco Bay Area, report LBL-10847 (see Figs. 7 and 8 and [P6]). Census tracts were the unit of study. Some 700 census tracts were classified into 12 categories of approximately equal risk with respect to age and socio-economic status. The incidence rates for these 12 categories for 26 histology-specific cancer sites were then analyzed non-parametrically. Groups experiencing high levels of pollution (total suspended particulate, sulfur dioxide and nitrogen dioxide) were compared to groups with low levels for differences in cancer incidence. Although this approach lacks some sensitivity, it does not depend on statistical models or ecologic regression techniques. Tract-level incidence data are clearly superior to the county-level mortality data that have been used in most ecologic studies to date.

The results of the San Francisco study were tantalizing but inconclusive: a slight association of gastro-intestinal cancer with total suspended particulate was noted.

2.3. 1980-1982 Research Results

In 1980, after four years of software and database development, the PAREP staff could begin to turn its attention to interesting statistical and epidemiological questions.

The ecologic patterns of disease for ten selected causes of death were described and analyzed in report LBL-10627 [P8]. This analysis involved the use of standardized age-adjusted mortality rates for county level data. A preliminary analysis relating cancer mortality and incidence to air pollution in California [P9] was published in November 1980. The methodology and results served as a prototype for a nationwide replication of the same type of investigation.

The problems, methodology and application of the biostatistical approach were described in report LBL-12217 (see [P12] and Fig. 9). The consequences of ecologic regression analysis were extensively studied, particularly in cases where this statistical approach is applied to the association between mortality and air quality. An innovative approach for dealing with ecologic data, where the sampling units (e.g. counties) differ in size, was described in report LBL-12216 [P10].

Another important epidemiological effort was the analysis of cancer incidence using data from the 1969-1971 Third National Cancer Survey. Major histologic types of cancers of the gum and mouth, esophagus, larynx and lung were discussed in report LBL-13337 [P11]. The relationship between income level and melanoma incidence was analyzed in report LBL-14107 [P14].

A descriptive study of the epidemiology of cell-type specific leukemia mortality (adult and childhood) is now complete and will be published in February 1983 (see Fig. 6 and [P17]). This study employed 1969-1977 leukemia mortality data, which are now integrated into the PAREP data base, to study the recent epidemiology of leukemia in greater detail than was previously possible.

In preparation for analyses to be completed in 1983 and 1984, important new data files were added to SEEDIS. Files added include 1970-1977 population by age, sex and race; 1970-1977 leukemia mortality; 1970 MEDX populations and centroids; geographic centroids of census tracts; 1970 socioeconomic data for Third National Cancer Survey (TNCS) census tracts; 1969-1971 TNCS cancer incidence by site and histology (see Fig. 7); 1973-1977 SEER cancer incidence; 1974-1976 air quality for individual stations, for discrete station locations, and for TNCS census tracts; 1968-1972 and 1973-1976 age adjusted mortality; 1968-1972 age specific mortality; 1980 population by race; and 1980 Census Summary Tape File 1. All SEEDIS data files are fully documented.

PAREP and SEEDIS user documentation was greatly improved and widely circulated in the scientific community. Presentations were made at a number of conferences and workshops [S1, S4, S5, S6, S9, S14, G2, G3, G4, G5, G6].

SEEDIS was installed at several additional sites -- Brookhaven National Laboratory and the Environmental Protection Agency Health Effects Research Laboratory, Chapel Hill, North Carolina. Additionally, collaborative relationships are being developed with state and local health monitoring agencies, such as the Northern California Cancer Program (NCCP). PAREP project personnel are helping new users -- epidemiologists and statisticians -- to become familiar with the tools and data at their disposal.

3. Future Plans

During 1983 and 1984, the PAREP project will integrate new sources of health and environmental data in order to study important epidemiological hypotheses. These efforts will include the development and testing of innovative statistical techniques. Improved documentation and further system development will make SEEDIS and the PAREP data base increasingly accessible to the epidemiological research community. Five broad research areas are described below.

3.1. Air Quality and Cancer Incidence

The previously completed analysis of air pollution and cancer incidence in the San Francisco Bay Area [P6] will be repeated for all nine areas of the 1969-1971 Third National Cancer Survey: Atlanta, Birmingham, Colorado, Dallas-Fort Worth, Detroit, Iowa, Minneapolis-St. Paul, Pittsburgh, and San Francisco-Oakland; and for the eleven areas of the 1973-1977 Surveillance, Epidemiology and End Results (SEER) study: Connecticut, Detroit, Iowa, Atlanta, New Orleans, New Mexico, Utah, Seattle, San Francisco-Oakland, Hawaii, and Puerto Rico. Air quality data for the entire 1969-1977 time period will be acquired and utilized. 1970 and 1980 Census data will be combined to provide accurate intercensal population denominators at the tract level. With 20 times more data than in the previous study, and with the possibility of comparing data from different geographic regions and time periods, the PAREP staff hopes to resolve the question of a possible relationship between air pollution and cancer incidence.

3.2. Air Quality Estimation and Interpolation

The relationship of air quality to disease is an important health issue. In order to study this relationship one requires estimates of air quality for population units such as counties or census tracts. Presently, the PAREP air quality data are derived from 1974-1976 measurements at some 6000 individual monitoring stations. This file, and the county- and tract-level files derived from it, have been widely distributed throughout the scientific community. An air quality modeling capability, recently integrated into SEEDIS [S13], uses the same station-level file to estimate air quality for an arbitrary set of geographic points specified by the user.

Recently, a new file has been created, which contains average measurements for each distinct station location. This file is interpreted more easily than the station-level file, which contains some conflicting measurements at the same geographic coordinates.

The interpolation method used in the PAREP project is a two-dimensional moving average. The weight assigned to each station is w=f*exp(-0.5*(d/d0)**2) where f is the fraction of time the station was active, d is the distance between the station and the point of estimation, and d0 is an empirical scaling parameter of the order of 10 or 20 kilometers. Different scaling parameters, other functional forms, and even different interpolation methods, could possibly be more appropriate and are being tested [P15, P18].

Data exploration and validation will determine what is necessary to estimate human exposure to air pollution, and to see how much accuracy is attainable with various interpolation algorithms. Time trends will be investigated. 1969-1979 air quality data from the EPA SAROAD (Storage and Retrieval of Aerometric Data) system and other sources such as the SURE Sulfate Regional Experiment will be installed in SEEDIS for comparison with the existing data. Spurious data values will be flagged for verification and possible rejection.

An optimum model will be chosen according to various criteria; these criteria include self-consistency in time and space, consistency with independently collected data, and the degree to which data at a given station can be reliably predicted from data at other nearby stations. Finally, the model will be refined, and the variance, covariance and bias of the estimates will be empirically determined.

3.3. Epidemiological Consequences of Employing Census Data

Fundamental to much of epidemiological research is the computation of incidence, morbidity and mortality rates. The numerators for these rates are typically obtained from the National Center for Health Statistics or the National Cancer Institute, whereas the denominators are generally extrapolated from the most recent decennial census. During the intercensal period 1971-1979 most epidemiological studies used unofficial population estimates provided by the Census Bureau. Now that 1980 census data are available, we propose to investigate the epidemiological consequences of possible errors and biases in the earlier intercensal estimates.

The SEEDIS system will probably be the first in the nation to have sex-, race-, and age-specific population counts for both the 1970 and 1980 U.S. Censuses. With these data it will be possible to analyze the accuracy of the intercensal estimates, particularly at the county level.

In general, mortality for almost all causes of death has decreased over the last decade. There is no widely accepted reason for this decrease. One possible explanation is consistent overestimation of intercensal populations. A comparison of the 1980 and 1970 census data will permit an assessment of this potential source of bias. Adjustments of rates, particularly mortality rates, will be based on the actual numbers as given in 1970 and 1980 rather than Census Bureau estimates. The adjusted rates should more accurately reflect the mortality risk. Investigators can then better examine the question of declining mortality rates and the reasons for such a decline if, indeed, it is still observed in the adjusted rates.

Another issue concerns the racial categories used in the census. Such categorization is often subject to changing social attitudes. For example, many members of the Mexican-American population classified themselves as white in the 1970 census but preferred to specify their race as "other" in the 1980 census. This type of change will be documented and analyzed using 1970 and 1980 census data. An estimate of the magnitude of this sort of misclassification bias is absolutely essential before epidemiological methods can be usefully applied to the study of disease. For example, a recent large increase in cancer rates was reported among Mexican-Americans living in the state of New Mexico. Before the significance of this observation can be fully understood, the influences of changing racial classification in the computation of these rates must be understood.

Other issues such as the underestimation of minority populations, extrapolation of counts to years beyond 1980, and possible overenumeration of persons over 65 years of age will be addressed. Attempts will be made to provide corrections to rate calculations to compensate for these biases.

3.4. Epidemiological Investigation of Cancer Mortality

A set of mortality data consisting of 21 million U.S. death records for the period 1969-1978 has been compiled by Alan Gittelsohn's Mortality Surveillance Project (Johns Hopkins University). This information, derived from the Public Use Tapes from the National Center for Health Statistics, has been condensed into compact records containing only relevant epidemiological information. These data are classified by sex, race, age, county of residence, and cause of death code (International Classification of Diseases and Accidents -- Eighth Revision) These data will be installed in the SEEDIS system, which includes the PAREP data base, and will serve as a valuable resource for PAREP-related studies.

The PAREP project has recently completed a comprehensive study of leukemia mortality data using a subset of the John Hopkins data (see Figs. 5 and 6, and [P16].) Several new studies will be conducted, using statistical techniques which were developed for the leukemia project. One such study involves the gastro-intestinal cancers which were observed to be distributed similarly in the nine areas of the Third National Cancer Survey [G11]. This observation suggested that other characteristics of these cancers (such as sex ratios, urban/rural differences, age and racial distributions) might be usefully examined in order to assess the possibility that these cancers have similar etiologies. Using county level data, such as those of the Mortality Surveillance Project data base, necessitates the development of new methodologies. In the leukemia study the PAREP project developed an innovative method for analyzing pairwise associations in county level data. A multivariate extension of this technique is necessary for adequate investigation of complicated phenomena such as the epidemiology of gastro-intestinal cancers.

A second example involves kidney cancer, a cancer that has a fairly high mortality rate but has not been extensively studied, particularly for the entire nation. The PAREP project will provide a descriptive study of this disease, employing specially developed statistical techniques to identify and explore its epidemiology.

Another area that will be studied is the epidemiology of childhood cancer. Most cancer mortality files contain too few cases of rare cancers to calculate reliable mortality rates for these cancers. However, a file of 21 million deaths provides adequate numbers of cases to study the rare occurrence of cancer among children less than 15 years of age. Thus, this project will fill a gap in the epidemiology of cancer and may provide important clues about its etiology. Even in past situations where the investigators had adequate numbers of rare cancers so that mortality rates could be calculated, attempts to examine confounding variables with cross-classification methods produced tables that were rather sparse (many empty cells). This sparseness necessitates the development of statistical methods to deal with such data. Some techniques are suggested in the statistical literature [G12] but these methods are just a beginning and do not specifically address many of the problems found in epidemiological data.

3.5. Retrieval, Analysis and Display of Data

SEEDIS is a testbed for advanced computer science research in network communications, user interfaces, and data management. The combination of epidemiological and computer science expertise in the PAREP project is a unique resource. The ability to answer increasingly complex epidemiological questions depends on the continued improvement of data retrieval, analysis and display techniques.

Of particular importance to the PAREP project is computer science research permitting the integration of large and diverse databases from various government agencies. A set of utilities (CODATA tools) for easily manipulating self-describing data files, and automatic aggregation and disaggregation of data between different geographic levels were developed to address this issue. The 1980 Census, required for population denominators and for indicators of socioeconomic status, requires data management techniques more sophisticated than those presently available in SEEDIS. The same applies to mortality files to be obtained from Johns Hopkins University and the National Center for Health Statistics, the integration of which will require flexible summarization techniques. The unit-record mortality files, too large for on-line disk storage, are archived and must be summarized to age and disease classifications specified by the user. Techniques must be further developed for estimating air quality for arbitrary geographic units, and for more efficiently converting data between different geographic entities.

SEEDIS, including the PAREP data base, demonstrates advanced development of data analysis and display techniques available to the epidemiological research community. In order to make these resources more widely available, LBL staff will implement important enhancements. Installation and test procedures for transferring the system software to other computers will be refined and generalized to cope with site-specific differences in operating systems. Documentation content and distribution facilities will be further improved. SEEDIS is now available through the National Energy Software Center and the National Technical Information Service.

The SEEDIS data base, on the order of 30 gigabytes, vastly exceeds the storage capacity of many computer systems that may choose to install SEEDIS. An efficient caching mechanism for locally storing extracted portions of the archived data base is an important component of making data and analytical resources available to and sharable by researchers at many sites. Equally important, users require mechanisms for conveniently documenting and archiving their own data on local storage devices. At the discretion of the installer and assuming an external network connection, such locally stored data could be available to users at other SEEDIS installations. This development will facilitate automatic sharing of data independently collected and maintained by different government agencies.

4. Philosophy

Epidemiology is the study of the factors involved in the occurrence or nonoccurrence of a disease. For statistical analysis, one would like to have data on each individual in a population, with information on such variables as health characteristics, smoking and eating habits, occupational status, etc. Such data can be obtained only at great expense, by conducting special surveys.

A less rigorous but much cheaper approach for epidemiological research involves the statistical analysis of ecologic data, i.e. summary data for geographic areas such as census tracts or counties. Relevant data that are available, having been consistently collected by federal government agencies over time, include data on mortality, socioeconomic and demographic characteristics, environmental pollution, and climate. A simple-minded approach is to treat counties or census tracts as the unit of analysis, analyzing them statistically as one would analyze individuals in a biological or medical experiment.

The "ecologic fallacy" is the false assumption that aggregating from the individual to the county or tract level does not fundamentally change the conclusions that can be drawn. Computers, and the availability of large machine-readable files, have simplified statistical analysis to the point that some authors have simply copied computer output into papers for publication, with hardly a word of caution. Some researchers have even gone so far as to mistake correlations for causality, claiming that burning so many tons of coal will cause so many excess deaths, etc.

Increasingly, the PAREP staff has grown wary of uncritical analysis of ecological data. With so many pitfalls for the unwary and so many powerful tools at our disposal, it is time to step back and look carefully at statistical methodology.

1982 was a year during which many other epidemiology research programs and integrated data systems (e.g. DIDS, UPGRADE, GEOECOLOGY) succumbed to budget cuts by federal funding agencies. SEEDIS and PAREP are indeed fortunate to have survived. No less important, the PAREP staff are uniquely provileged to find themselves in a productive computer science research environment, where the careful development of flexible information systems and sound statistical methodology take precedence over the hasty publication of new results. It is hoped that the happy combination of epidemiological, statistical, and computer science expertise that is PAREP will remain intact for years to come.

5. Acknowledgments

The success of the PAREP project is a fitting testimonial to the vision and energy of its creator Craig Hollowell, who died suddenly and tragically early in 1982.

Co-principal investigators of the PAREP project are Deane Merrill (Computer Science and Mathematics Department - CSAM) and Steve Selvin (CSAM and School of Public Health, University of California at Berkeley - SPH). Other major contributors since 1976 include Susan Sacks (CSAM and University of California Medical Center, San Francisco - UCSF), Linda Wong (formerly Linda Kwok, CSAM), Laura Johnson (CSAM), Warren Winkelstein, Jr. (SPH), Donald M. Austin (formerly CSAM) and Barbara Levine (formerly CSAM). Shorter-term but important contributions were made by Elizabeth Holly (formerly SPH), Lynn Levin (SPH), Virginia Ernster (UCSF), Bill Hogan (formerly CSAM), Brad Heckman (formerly CSAM), Jos Polman (formerly CSAM), Simcha Knif (formerly CSAM), Jim McMahon (formerly CSAM), Claudette Lederer (formerly CSAM), Jim Schofield (formerly CSAM) and Norm Handy (Jackson State University).

The PAREP project is indebted to Carl Quong, head of the LBL Computer Science and Mathematics Department, for sustained interest and access to the facilities and staff supported by other CSAM projects. The entire CSAM staff, numbering about 50 professionals, is responsible for the success of the PAREP project.

The PAREP staff thanks the persons who made the project possible: Bill Nelson of the Biometry Division of the Environmental Protection Agency (EPA), Walter Weyzen, John Viren, Bob Goldsmith and Joe Blair of the Office of Health and Environmental Research of the Department of Energy, and Ron Wyzga of the Electric Power Research Institute.

Other persons who have provided special assistance include Sandra Stinnett of the University of North Carolina; Wilson Riggan, John van Bruggen, Carol Evans, Carolyn Chamblee, John O'Neill, Joe Ryan, Robin Davis and Loren Hall, all of the Environmental Protection Agency; Tom Mason and Earl Pollack of the National Cancer Institute; Donald F. Austin of the Northern California Resource for Cancer Epidemiology; Carmen Benkovitz of Brookhaven National Laboratory; Dick Olsen of Oak Ridge National Laboratory; Herb Sauer of Herb Sauer Associates; and Carol Graves, Ralph Tartaglione, Jacob Thomas and Larry Milask, all previously of M/A-Com Sigma Data Corp.

6. References

6.1. PAREP

P1. Merrill, D.W. Jr.; Sacks, S.T.; Selvin, S.; Hollowell, C.D.; and Winkelstein, W. Jr.; Populations at Risk to Air Pollution (PARAP): Data Base Description and Prototype Analysis: UCID-8039, August 1978.

P2. Sacks, S.T., Selvin, S.; and Merrill, D.W.; Building a United States Data Base: Populations at Risk to Environmental Pollution; Presented at the Conference on Demographic and Health Information for Aging Research: Resources and Needs; National Institute on Aging,

National Institutes of Health, Bethesda MD, June 25-27, 1979; LBL-9636.

P3. Merrill, D.; Levine, B.; Sacks, S.; and Selvin, S.; PAREP: Populations at Risk to Environmental Pollution; Presented at the 4th International Conference on Computer Assisted Cartography, Reston VA, November 4-8, 1979; LBL-9976.

P4. Selvin, S.; Sacks, S.T.; Winkelstein, W. Jr.; Holly, E., and Merrill, D.W.; Populations at Risk to Environmental Pollution: Prospectus: FY 80 and FY 81; LBID-164, January 1980.

P5. Selvin, S.; Sacks, S.T.; and Merrill, D.W.; Standardization of Age-Adjusted Mortality Rates; LBL-10323; February 1980.

P6. Selvin, S.; Sacks, S.T.; and Merrill, D.W. and Winkelstein, W. Jr.; The Relationship Between Cancer Incidence and Two Pollutants (Total Suspended Particulate and Carbon Monoxide) for the San Francisco Bay Area; LBL-10847 and UC-11; June, 1980.

P7. Austin, D.M.; Kwok, L.; Merrill, D.W.; Sacks, S.T.; Selvin, S.; and Winkelstein, W. Jr.; August 1980; PAREP (Populations at Risk to Environmental Pollution): included in Lawrence Berkeley Laboratory Annual Report for 1979.

P8. Selvin, S., Sacks, S. and Merrill, D.W. Patterns of United States Mortality for 10 Selected Causes of Death -- Lawrence Berkeley Laboratory Report, LBL-10627, November 1980. Submitted (8/81) to American Journal of Public Health.

P9. Winkelstein, W. Jr.; Selvin, S.; Holly, E.A.; Sacks, S.T. and Merrill, D.W.; Cancer Mortality-Incidence and Air Pollution in California, University of California School of Public Health Publication; November, 1980.

P10. Selvin, S., Merrill, D. and Sacks, S. An Alternative to Ecologic Regression Analysis of Mortality Rates -- Lawrence Berkeley Laboratory Report, LBL-12216, March 1981. American Journal of Epidemiology Vol. 115, No. 4, pp. 617-623 (1982).

P11. Ernster, V.L.; Selvin, S.; Sacks, S.T.; Merrill, D.W.; and Holly, E.A.; Sex, Income Level and Major Histologic Types of Cancers of the Gum and Mouth, Esophagus, Larynx and Lung: U.S. Third National Cancer Survey, 1969-1971; LBL-13337; September 1981. Presented at the 109th Annual Meeting of the American Public Health Association, Los Angeles CA, November 1981. Journal of the National Cancer Institute (in press).

P12. Selvin, S., Merrill, D., Kwok, L. and Sacks, S. Ecologic Regression Analysis and the Study of the Influence of Air Quality on Mortality -- Lawrence Berkeley Laboratory Report, LBL-12217, October 1981. Submitted (9/82) to Environmental Health Perspectives.

P13. Computer Science and Mathematics Department, Environmental and Epidemiological Studies -- contribution to 1980 Annual Report of Physics, Computer Science and Mathematics Division -- Lawrence Berkeley Laboratory Report, LBL-12760, December 1981.

P14. Holly, E., Selvin, S., Sacks, S., Merrill, D. and Ernster, V. Melanoma Incidence and Income in the U.S. Third National Cancer Survey -- Lawrence Berkeley Laboratory Report, LBL-14107, February 1982. Submitted to the Journal of the National Cancer Institute.

P15. Johnson, L., Merrill, D. and Selvin, S. Predicting a Continuous Spatial Variable from Discrete Point Measurements -- Lawrence Berkeley Laboratory Report, LBL-14235, March 1982. To be included in L. Johnson dissertation.

P16. Selvin, S., Levin, L., Merrill, D. and Winkelstein, W. Jr. Selected Epidemiological Observations of Cell-specific Leukemia Mortality in the United States, 1969-1977 -- Lawrence Berkeley Laboratory Report, LBL-14234, April 1982. Presented at 15th Annual Meeting of Society for Epidemiological Research (SER), Cincinnati, Ohio, June 17-19, 1982. To be published (2/83) in the American Journal of Epidemiology.

P17. Winkelstein, W. Jr., Merrill, D., Pinto, J. and Syme, S. Time Trends for Ischemic Heart Disease Mortality in the USA: Differences Between States by Age and Sex -- Lawrence Berkeley Laboratory Report, LBL-15235. Submitted to the American Heart Association, 23rd Conference on Cardiovascular Disease Epidemiology, Dallas TX, March 3-5, 1983.

P18. Johnson, L. Statistical and Geographical Analysis of Air Quality in the United States -- Dissertation, University of California at Berkeley (work in progress).

6.2. SEEDIS

S1. Merrill, D. Handling Spatial Data in SEEDIS, Socio-Economic Environmental Demographic Information System; in proceedings of 1980 Integrated County-Level Data User's Workshop, Reston, Virginia, October 1980; Oak Ridge National Laboratory Report, CONF-8010139.

S2. Computer Science and Mathematics Department, Release Notes for SEEDIS Version 1.1 -- Lawrence Berkeley Laboratory Internal Document, LBID-357 Rev, January 1981.

S3. Computer Science and Mathematics Department, Release Notes for SEEDIS Version 1.2 -- Lawrence Berkeley Laboratory Internal Document, LBID-357 Rev, June 1981.

S4. Merrill, D. Automatic Aggregation and Disaggregation of Data -- Lawrence Berkeley Laboratory Report, LBL-14236, December 1981. Presented at the Workshop on Statistical Database Management, Menlo Park, CA, December 1981.

S5. Merrill, D. Air Quality Data in SEEDIS: Socio-Economic Environmental Demographic Information System -- Lawrence Berkeley Laboratory Report, LBL-14237, December 1981. Presented at Stanford University Department of Statistics, SIMS Statistics and Air Pollution Seminar, December 1981.

S6. Merrill, D. Problems in Spatial Data Analysis -- Lawrence Berkeley Laboratory Report, LBL-14047, February 1982. Published in proceedings of Seventh International Conference of SAS Users Group International, San Francisco CA, February 1982, pp. 218-223.

S7. Department of Labor, Employment and Training Administration and Lawrence Berkeley Laboratory -- Reports 1A and 1B: Population Characteristics: 1980 Census of Population -- Lawrence Berkeley Laboratory Report, LBL-14636, April 1982.

S8. McCarthy, J., Merrill, D., Marcus, A., Benson, W., Gey, F., Holmes, H. and Quong, C.; The SEEDIS Project: A Summary Overview of the Social Economic, Environmental, Demographic Information System; Lawrence Berkeley Laboratory Report PUB-424 Rev; May 1982.

S9. Merrill, D. SEEDIS: Socio-Economic Environmental Demographic Information System -- Lawrence Berkeley Laboratory Report, LBL-14233, May 1982. Presented at IASSIST: International Association for Social Science Information Service and Technology Annual Conference and Workshops, San Diego CA, May 1982. To be published in IASSIST newsletter, 1983.

S10. Computer Science and Mathematics Department, Release Notes for SEEDIS Version 1.3 -- Lawrence Berkeley Laboratory Internal Document, LBID-357 Rev, May 1982.

S11. McCarthy, J., Benson, W., Yen, A., Merrill, D., Marcus, A., Gey, F., Holmes, H. and Quong, C. SEEDIS: A Research and Development Project on Social, Economic, Environmental and Demographic Information Systems -- Lawrence Berkeley Laboratory Report, LBL-14417, May 1982. Condensed version published in Computer Graphics World, Vol. 5, No. 6, June 1982, pp. 35-44.

S12. Merrill, D. CODATA User's Manual -- Lawrence Berkeley Laboratory Internal Document, LBID-021 Rev, August 1982.

S13. Computer Science and Mathematics Department, Release Notes for SEEDIS Version 1.4 (preliminary) -- Lawrence Berkeley Laboratory Internal Document, LBID-357 Rev, November 1982.

S14. Merrill, D. Distributed Data Management in a Minicomputer Network: the SEEDIS Experience -- Lawrence Berkeley Laboratory Report, LBL-15075. To be published in Proceedings of the 1982 Integrated Data Users Workshop, Reston VA, October 1982.

6.3. General

G1. Freedman, S.J., Lewis-Heise, E., Wilson, J. and Hardy, A. Population at Risk to Various Air Pollution Exposures: Data Base "POPATRISK" -- Environmental Protection Agency Report, EPA-600/1-78-051, June 1978.

G2. Merrill, D. Geographic Information Systems -- presented at Congressional Seminar of Computer Graphics, Congressional Research Service, U.S. Library of Congress, April 1981.

G3. Merrill, D. Integrated Analysis Systems -- Lawrence Berkeley Laboratory Report, LBL-13418. In proceedings of 1981 Integrated Data Users Workshop, Reston, Virginia, October 1981; Oak Ridge National Laboratory Report, CONF-8110199, pp. 5-18.

G4. Merrill, D. Accessing and Analyzing United States Data Files -- Lawrence Berkeley Laboratory Report, LBL-14232, April 1982. Two lectures for seminar on Health Care and Health Statistics, presented at University of California School of Public Health, Berkeley CA, April 7 and 9, 1982.

G5. Merrill, D. On-Line Socio-Economic and Demographic Data Bases Lawrence Berkeley Laboratory Report, May 1982. Presented at IASSIST: International Association for Social Science Information Service and Technology Annual Conference and Workshops, San Diego CA, May 1982. To be published in IASSIST newsletter, 1983.

G6. Merrill, D. Overview of Integrated Data Systems: Context, Capabilities and Status -- Lawrence Berkeley Laboratory Report, LBL-15074. To be published in Proceedings of the 1982 Integrated Data Users Workshop, Reston VA, October 1982. Revised version to be published (spring 1973) in American Demographics.

G7. McCarthy, J. Metadata Management for Large Statistical Databases -- Lawrence Berkeley Laboratory Report, LBL-14151. Presented at the Eighth International Conference on Very Large Data Bases, Mexico City, 1982.

G8. Chan, P. and Shoshani, A. SUBJECT: A Directory Driven System for Large Statistical Databases -- in Proceedings of the First LBL Workshop on Statistical Database Management, Menlo Park, California, December 1981 -- Lawrence Berkeley Laboratory Report LBL-13851, UC-13, CONF-811208.

G9. Johnson, R. A Data Model for Integrating Statistical Interpretations -- in Proceedings of the First LBL Workshop on Statistical Database Management, Menlo Park, California, December 1981 -- Lawrence Berkeley Laboratory Report LBL-13851, UC-13, CONF-811208.

G10. Eggers, S., Shoshani, A., Efficient Access of Compressed Data March 1, 1980, LBL-10648.

G11. W. Winkelstein, S.T. Sacks, V.L. Ernster and S. Selvin; Correlations of Incidence Rates for Selected Cancer in the Nine Areas of the Third National Cancer Survey, J. Am. Epid., 1977.

G12. Bishop, Y.M.M., Feinberg, S.E., and Holland, P.W., Discrete Multivariate Analysis, MIT press.

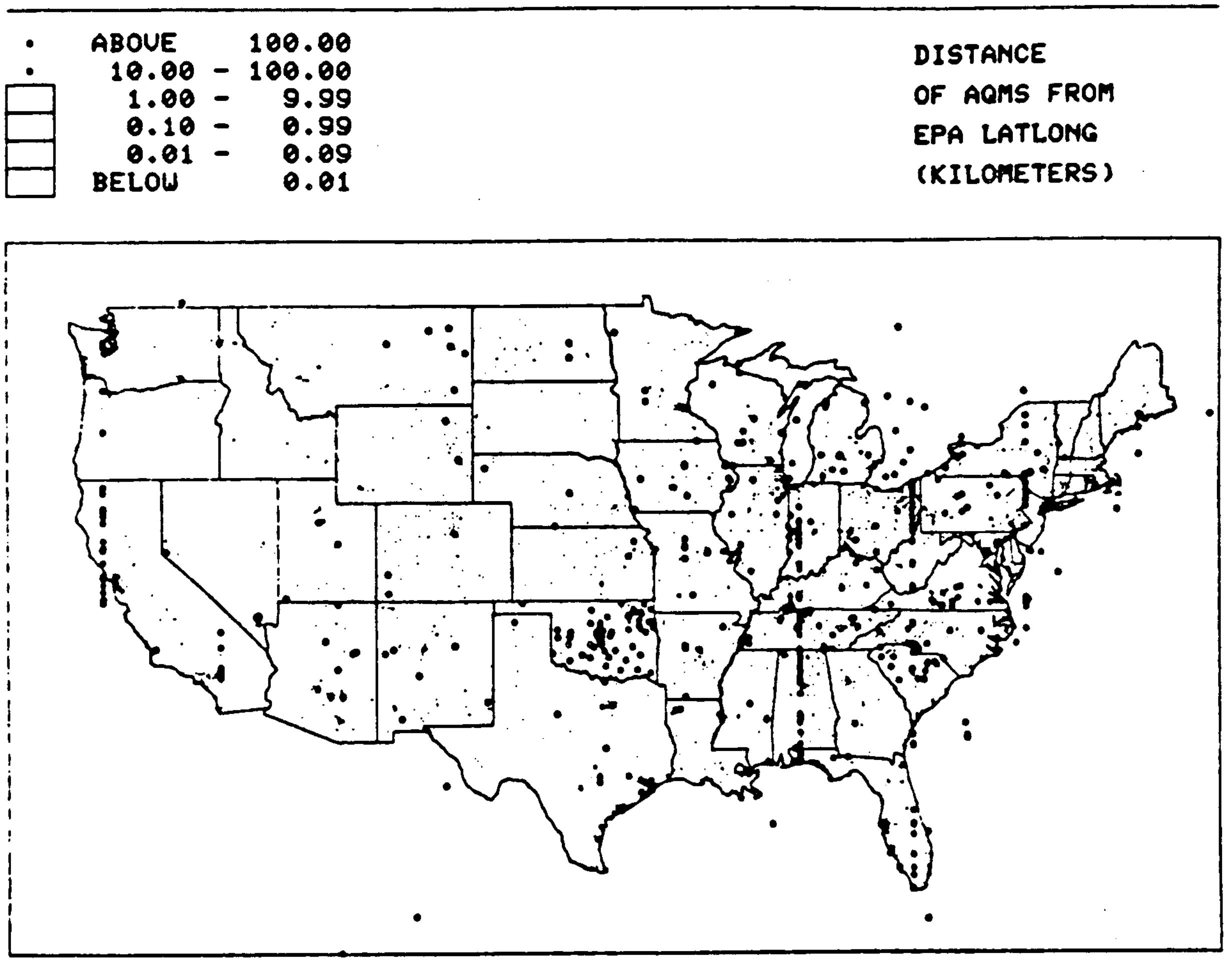


FIGURE 1

1) Locations of air quality monitoring stations, as recorded in the Storage and Retrieval of Aerometric Data (SAROAD) data base of the Environmental Protection Agency (EPA). Only stations active during 1974-1976 are included. Stations indicated by small dots (barely visible) are probably within 10 kilometers of the correct location; stations indicated by circles were corrected at LBL by more than 10 kilometers. In addition to random errors (e.g. stations in Mexico and the Atlantic Ocean), many stations have incorrect longitudes which happen to correspond exactly to Universal Transverse Mercator (UTM) zone boundaries. EPA was advised of the problems as early as 1977 (by LBL, Brookhaven National Laboratory, and others) but has not yet corrected the SAROAD data base, as far as is known.

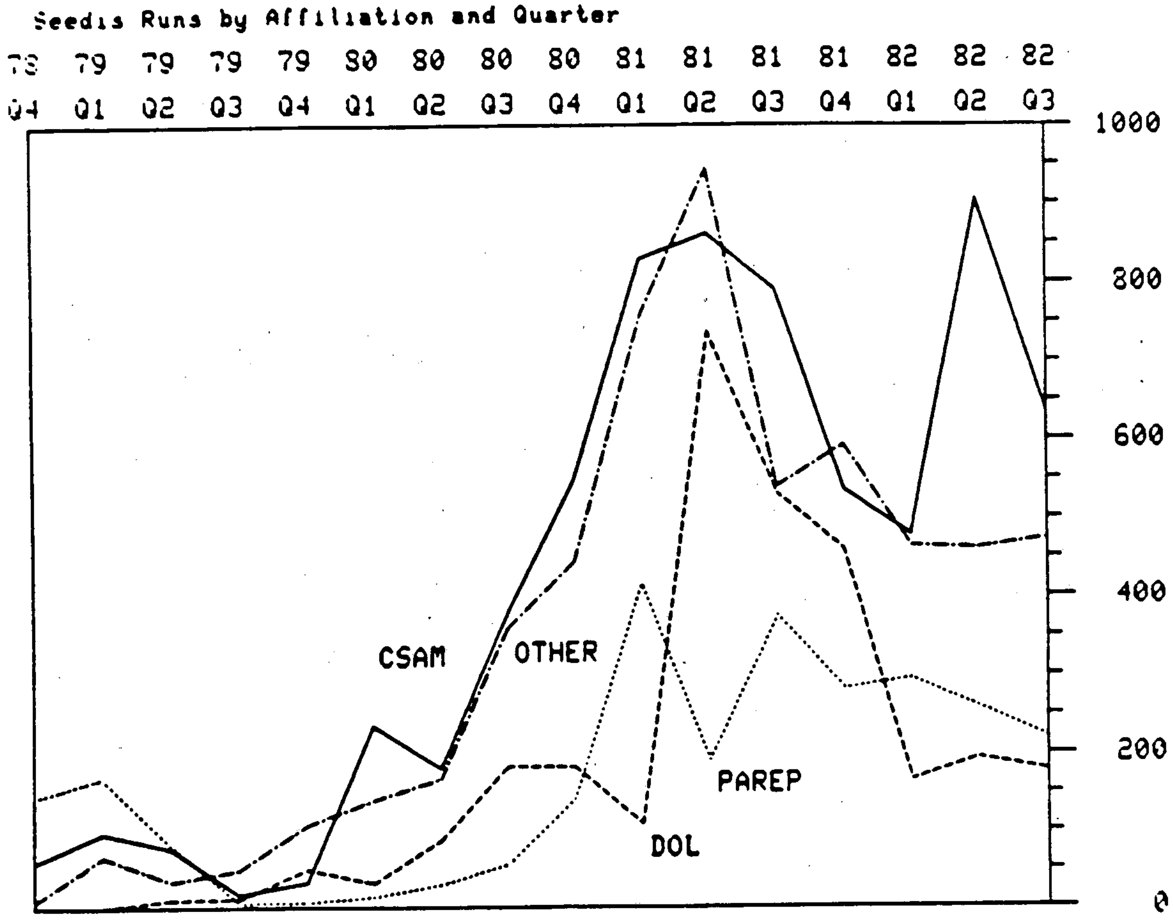


FIGURE 2

2) Number of SEEDIS runs by affiliation and quarter, from the last quarter of 1978 through the third quarter of 1982. "PAREP" includes runs for PAREP applications; "CSAM", which includes development and testing, is the remainder of the LBL Computer Science and Mathematics Department; "DOL" is Department of Labor; "OTHER" includes the rest of LBL, Army Corps of Engineers, Survey Research Center, and miscellaneous. There were 1500 SEEDIS runs during the third quarter of 1982, of which PAREP applications accounted for 15 percent; since 1978, there have been 19,000 SEEDIS runs, of which PAREP applications there also accounted for 15 percent. Runs of late 1978 and early 1979 were on the CDC (Control Data Corporation) computer of the LBL computer center; runs beginning in late 1979 were on DEC VAX computers. Low usage occurred in 1979 during the CDC-DEC transition period. Reduced usage in late 1981, especially for DOL and OTHER, resulted from budget cutbacks.

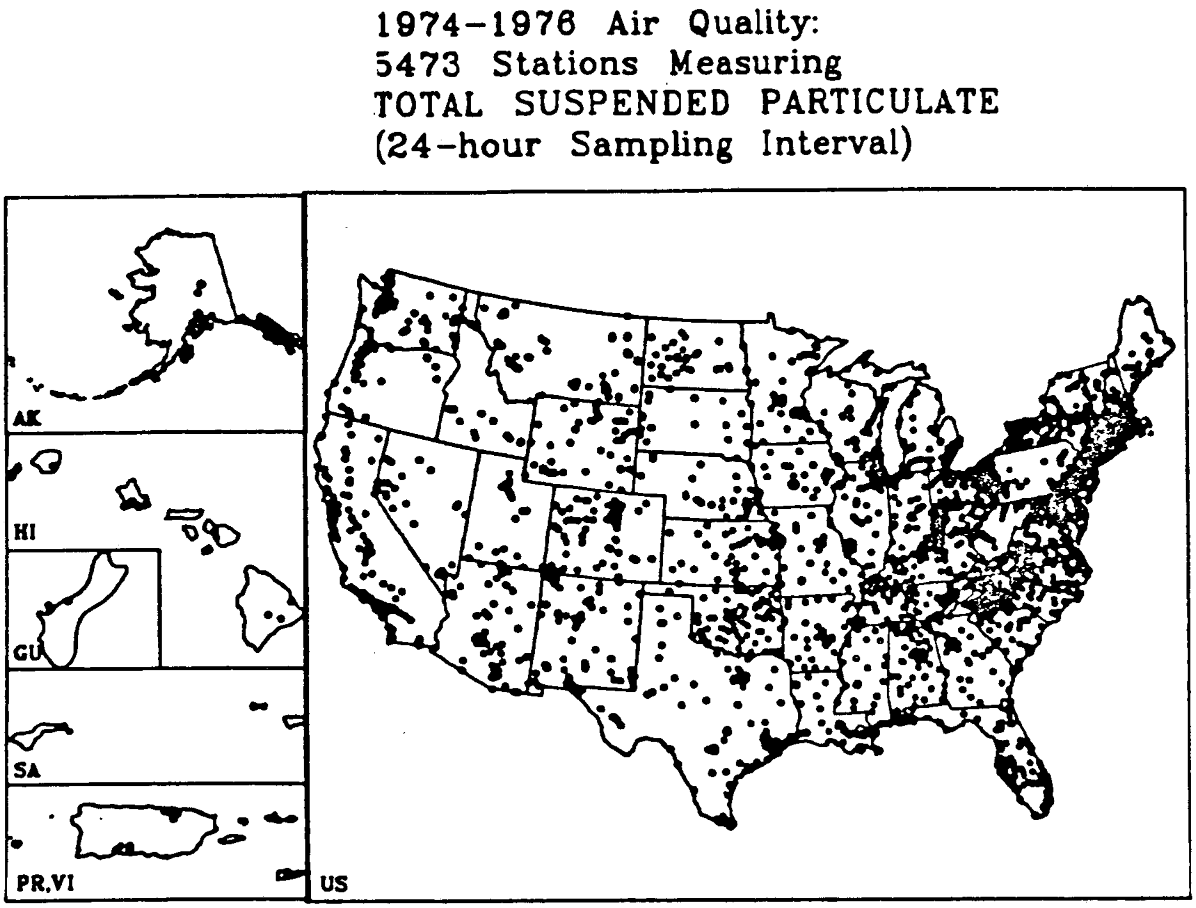


FIGURE 3

3) Locations of air quality monitoring stations, as recorded in SEEDIS. The geographic coordinates in the SAROAD data base were corrected after comparison of various independent sources. Only stations which measured total suspended particulate (TSP) during 1974-1976 are included. Coverage includes the United States plus the territories of Guam, Puerto Rico, and the Virgin Islands; stations in American Samoa were not active during 1974-1976. TSP is measured over a 24-hour sampling interval; at most stations, data are regularly collected every sixth day. TSP monitoring is most active in urban areas; the population concentrations along major highways in California and Oregon are plainly visible.

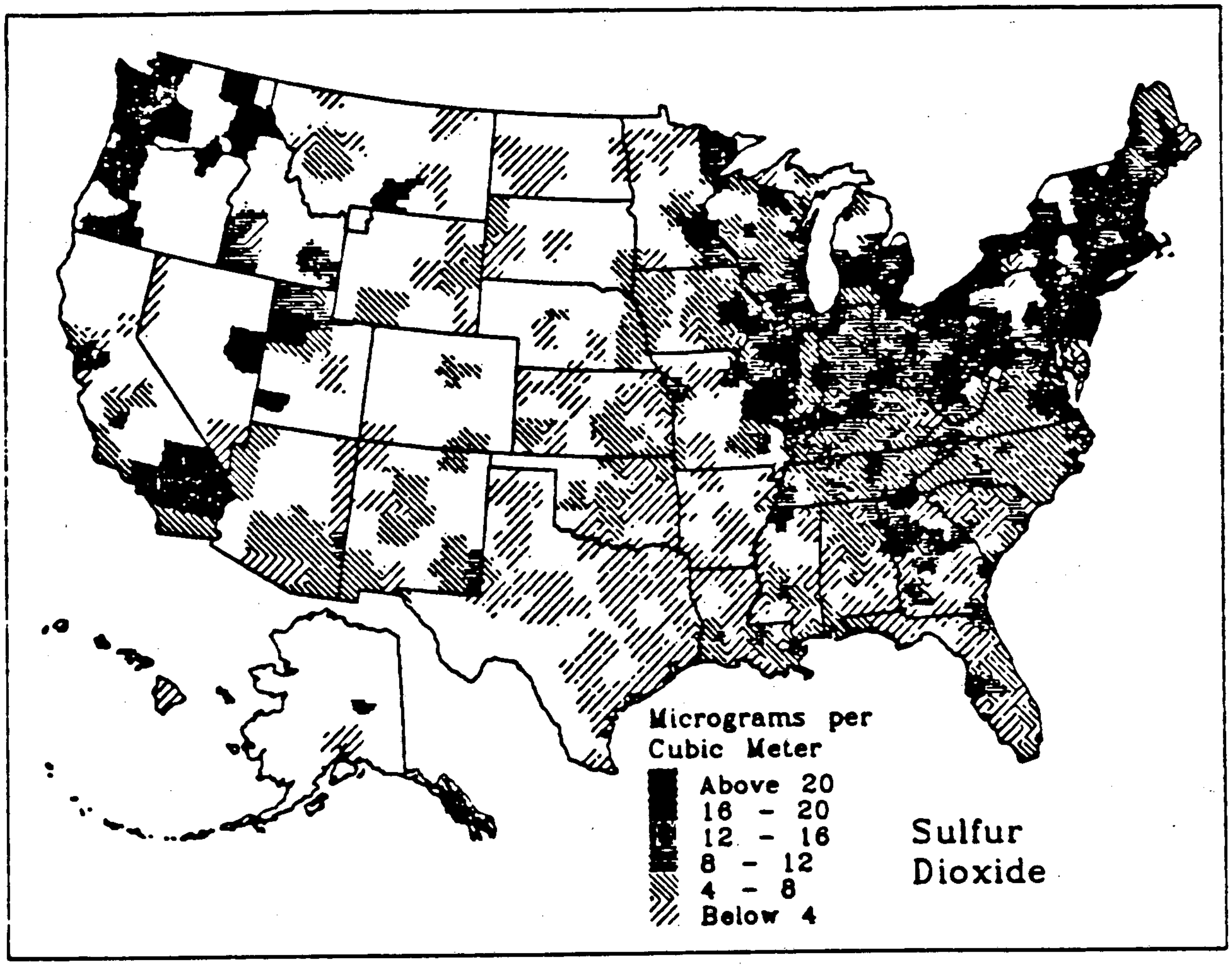


FIGURE 4

4) Estimated sulfur dioxide concentration in the United States in 1974-1976. For each county, concentration was estimated at its population centroid as a weighted geometric mean of values from nearby monitoring stations. For a given county centroid, the weight assigned to each station was f*exp(-.5*(d/d0)**2), where f is the fraction of time the station was active, d is the distance in kilometers between the county centroid and the monitoring station (which may be outside the county), and d0 is 20 (kilometers). The form of the weighting function and the choice of d0 are the subject of ongoing study. Data from stations with 1-hour sampling intervals and 24-hour sampling intervals were combined; each 24-hour observation received 24 times the weight of a single 1-hour observation. Unshaded areas in the map correspond to missing data, i.e. counties for which there were no active monitoring stations within 60 kilometers of the population centroid. Sulfur dioxide concentration is highest in urban areas, especially the industrialized northeast.

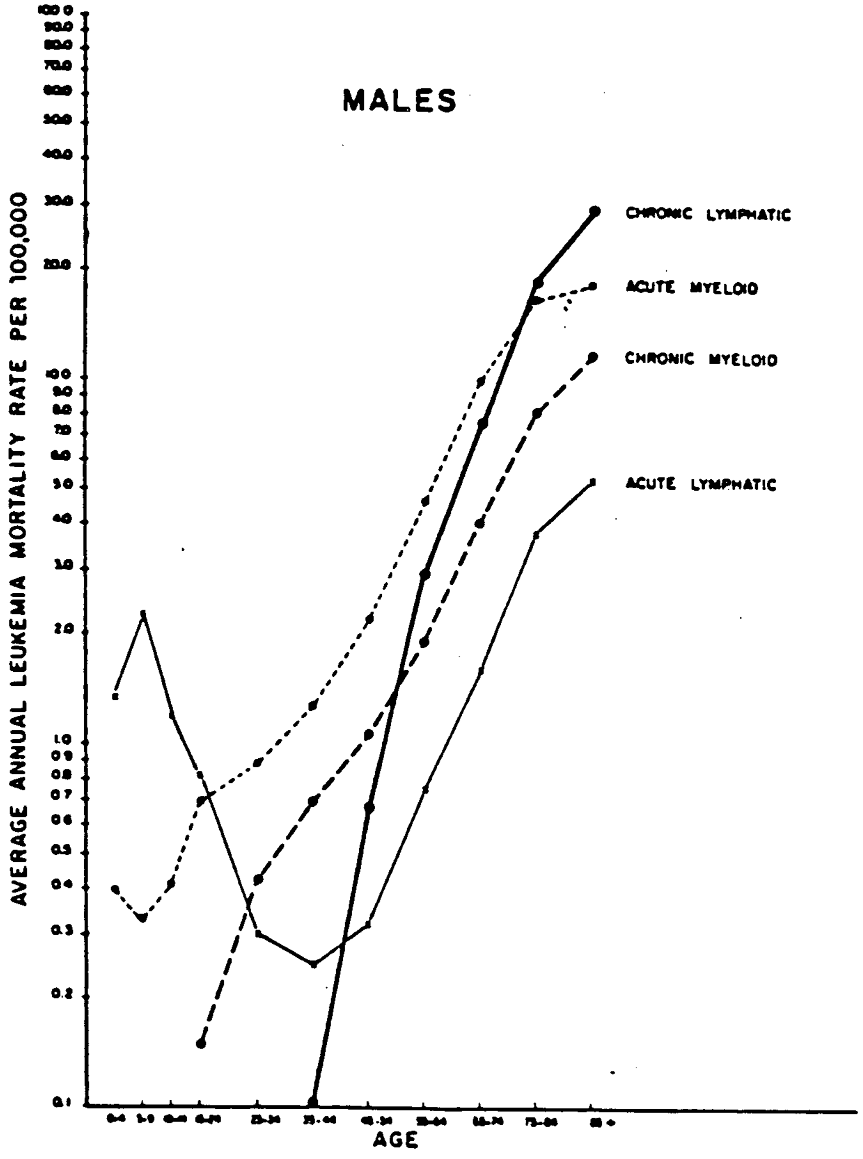
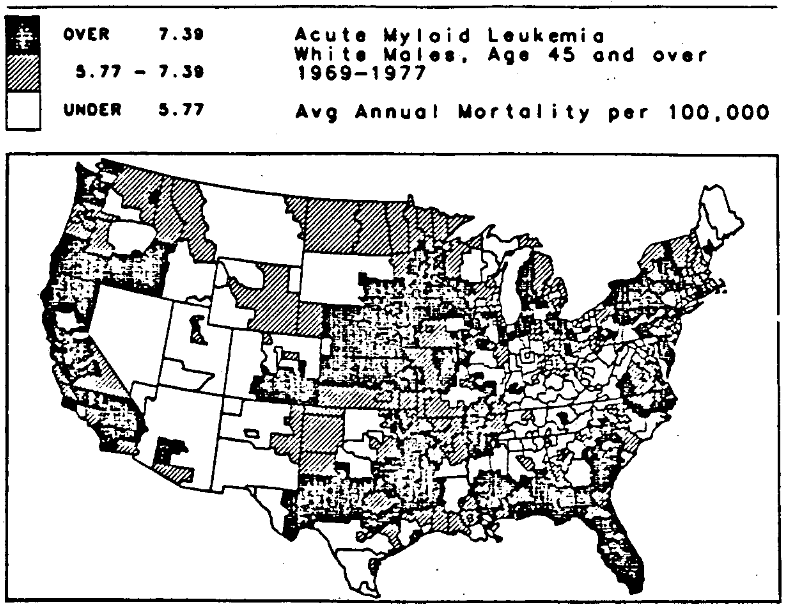


FIGURE 5

5) Average annual leukemia mortality rates (age-specific, per 100,000) for white males in the United States, 1969-1977. Lymphatic and myeloid leukemia are distinguished by different cell types; attribution of death to either chronic or acute leukemia depends on the survival time after the onset of symptoms. Childhood leukemia is almost entirely of the acute lymphatic type; among adults the other three types dominate. With increasing age, the rate of chronic lymphatic leukemia increases most rapidly. Rates for females (not shown) follow the same patterns as for males but are consistently lower. This figure shows the importance of analyzing separately the four types of leukemia, which has not always been possible due to the lack of adequate data.

6a


OVER 7.39
5.77 - 7.39
UNDER 5.77
Acute Myloid Leukemia
White Males, Age 45 and over
1969–1977
Avg Annual Mortality per 100,000

6b

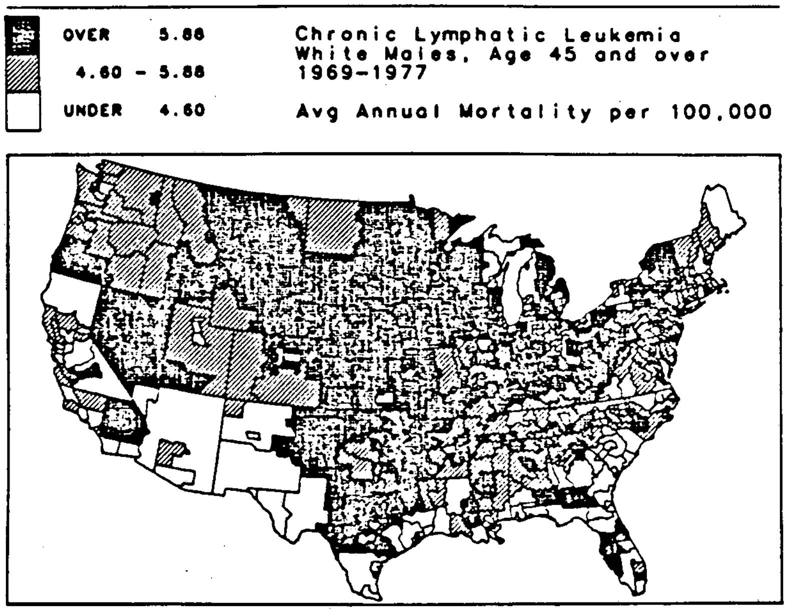

OVER 5.88
4.60 - 5.88
UNDER 4.60
Chronic Lymphatic Leukemia
White Males, Age 45 and over
1969–1977
Avg Annual Mortality per 100,000

FIGURE 6

6) Average annual leukemia mortality rates (per 100,000) for white males aged 45 and over in the United States, 1969-1977. Rates for acute myeloid leukemia and chronic lymphatic leukemia are shown in Figures 6(a) and 6(b) respectively. The geographic areas shown are Public Use Sample county groups of the 1970 Census, each of which has a minimum population of 250,000 and is composed of counties having similar socio-economic characteristics. The geographic patterns observed in 6(a) and 6(b) are clearly different; most striking are the below-average rates of acute myeloid leukemia in the southwest and in the Appalachian states, and the above-average rates of chronic lymphatic leukemia in the Great Plains states. Maps (not shown) for other leukemia types, for persons under 45 years, and for females exhibit different patterns. The differences, which are statistically significant, are not understood.

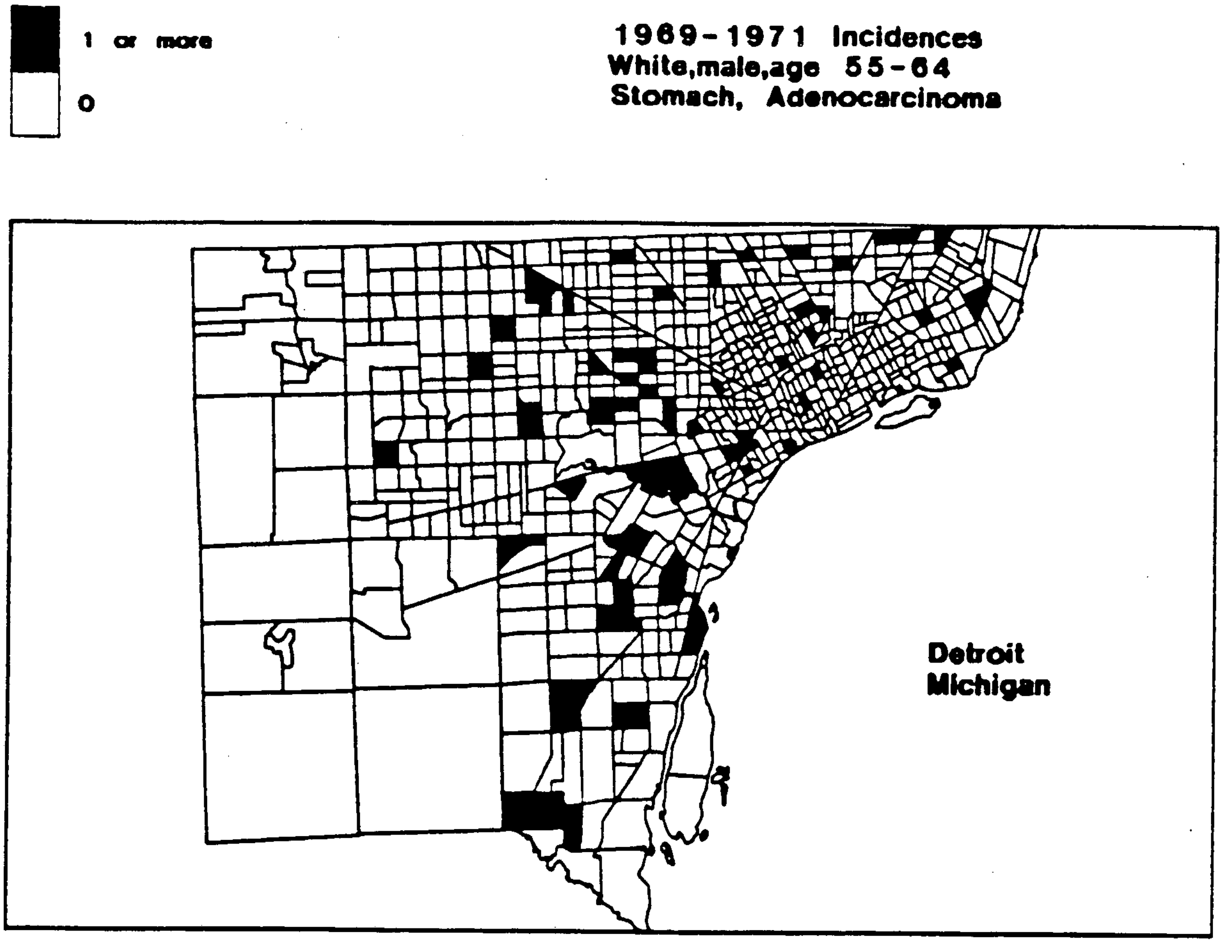


FIGURE 7

7) Number of cases of stomach cancer (adenocarcinoma) in white males aged 55-64, in the census tracts of Wayne county, Michigan in 1969-1971. This small data sample is illustrative of cancer incidence data available from the TNCS and SEER surveys of the National Cancer Institute, for a dozen urban areas of the United States. Census tracts contain roughly 4000 persons each, so the size of each tract gives an indication of its population density. Detroit is in the densely populated northeast portion of Wayne county. To the southeast lies the Detroit River and Windsor, Ontario. Major traffic arteries serving as census tract boundaries are visible. Of 711 census tracts, 651 tracts had no cases, 55 tracts had one case, three tracts had two cases each, and two tracts had three cases each. Central Detroit, which is predominantly black, has on the average fewer white males and correspondingly fewer cases per tract than do the western suburbs. For Wayne county as a whole, the 1970 white male population aged 55-64 was 91,987; the 67 cases correspond to an average annual incidence rate of 24 per 100,000. Taken by themselves, the data shown here are too scanty to permit any conclusions to be drawn.

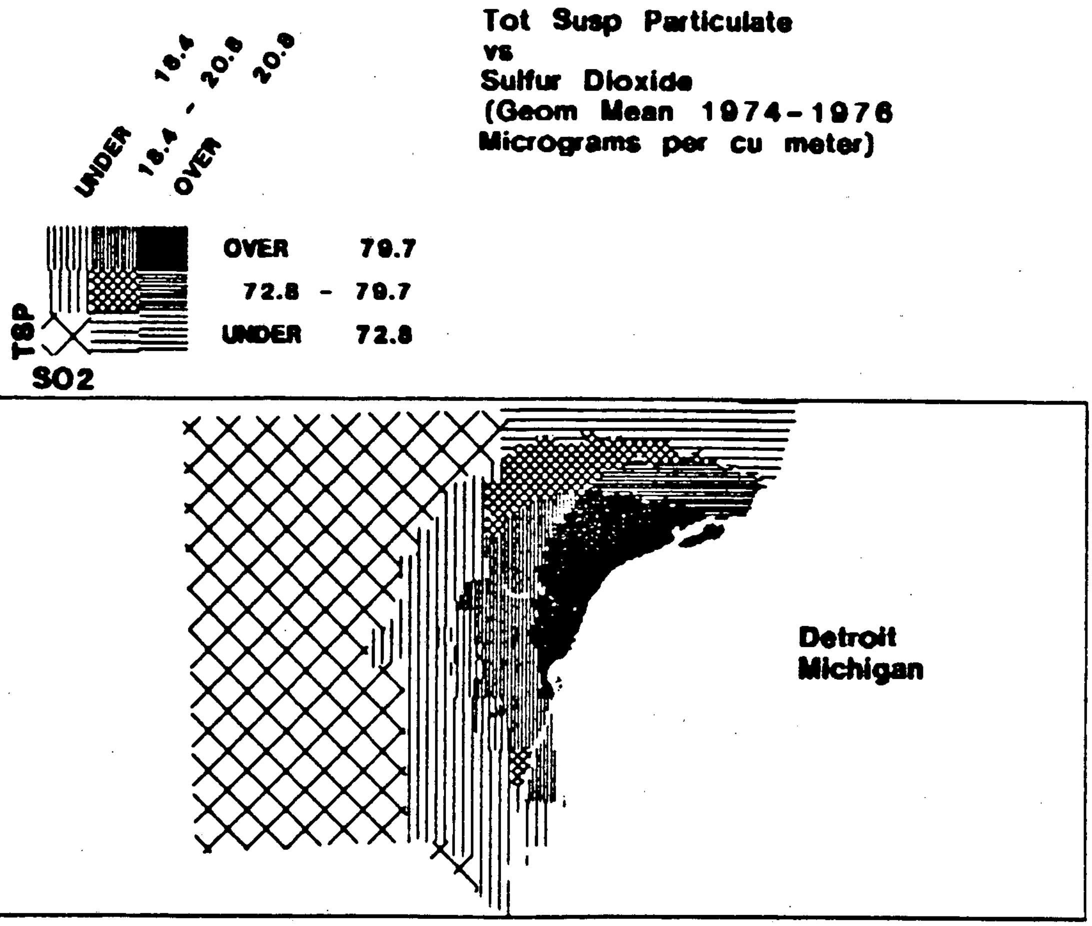

Tot Susp Particulate
vs
Sulfur Dioxide
(Geom Mean 1974-1976
Micrograms per cu meter)

FIGURE 8

8) Estimated concentration of total suspended particulate (TSP) and sulfur dioxide (SO2), for Wayne county, Michigan in 1974-1976. These data, along with air quality data from other time periods and corresponding socioeconomic indicators, are being studied in conjunction with cancer incidence patterns observed in the TNCS and SEER surveys. For each tract (boundaries not shown), concentration was estimated at its centroid as a weighted geometric mean of values from nearby monitoring stations. For a given tract centroid, the weight assigned to each station is f*exp(-.5*(d/d0)**2), where f is the fraction of time the station was active, d is the distance in kilometers between the tract centroid and the monitoring station, and d0 is 10 (kilometers). The form of the weighting function and the choice of d0 are the subject of ongoing study. For SO2, data from stations with 1-hour sampling intervals and 24-hour sampling intervals are combined; each 24-hour observation receives 24 times the weight of a 1-hour observation. For both TSP and SO2, concentrations are low in the western suburbs and high in central Detroit. The patterns are similar but not identical: high-SO2, moderate-TSP areas occur in the northeast, near Grosse Point, while high-TSP, moderate-SO2 areas occur just west of Detroit, near Dearborn. The differences, of the order of only 10 percent, may not be statistically significant.

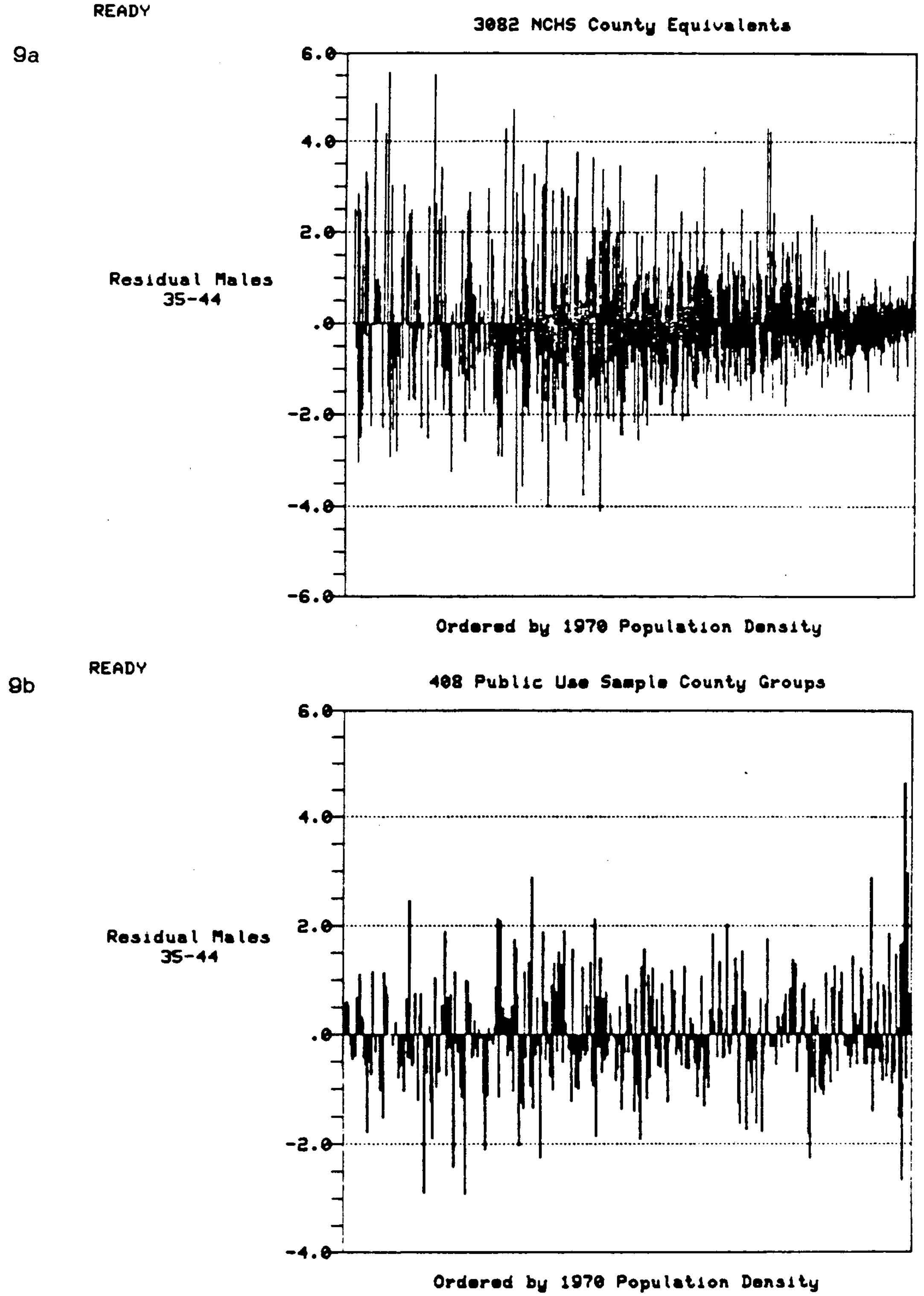


FIGURE 9

9) Residuals from a regression analysis of total mortality in white males aged 35 to 44 in the United States, 1968-1972. In Figure 9(a) the units of analysis are 3082 counties or county equivalents, in 9(b) the units of analysis are 408 Public Use Sample (PUS) county groups defined by the 1970 Census, each of which has a minimum population of 250,000 and is composed of counties having similar socio-economic characteristics. In a "well-behaved" regression analysis, one would expect to see in both 9(a) and 9(b) a uniform random distribution of residuals, with no dependence upon population density. Instead, one observes systematically smaller residuals for densely populated (urban) counties, and systematically larger residuals for the most densely populated PUS areas. These abnormalities are not understood and are representative of many similar difficulties that are observed when attempting to perform regression analyses on county-level data. The general observation is that results depend strongly upon initial assumptions that should not make any difference. As a consequence, the conclusions of many published analyses, which used similar data and methods, are cast strongly in doubt.

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the author(s) and not necessarily those of The Regents of the University of California, the Lawrence Berkeley Laboratory or the Department of Energy.

Reference to a company or product name does not imply approval or recommendation of the product by the University of California or the U.S. Department of Energy to the exclusion of others that may be suitable.

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