PAREP: POPULATIONS AT RISK TO ENVIRONMENTAL POLLUTION

D. Merrill and B. Levine
Computer Science and Applied Mathematics Department
S. Sacks and S. Selvin
Energy and Environment Division
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

I. Introduction

The project PAREP (Populations at Risk to Environmental Pollution) is being conducted at the Lawrence Berkeley Laboratory under funding from the Department of Energy. PAREP supersedes an earlier project PARAP (Populations at Risk to Air Pollution), funded by the Environmental Protection Agency (EPA).

An integrated data base has been assembled by the LBL Computer Science and Applied Mathematics Department (1). Analysis of the data is being performed by the LBL Energy and Environment Division, in collaboration with the UCB (University of California at Berkeley) School of Public Health.

II. Data Base Description

The PAREP data base covers the U.S. and territories at the county level, with subcounty detail for some data elements. The data include socioeconomic and demographic data, mortality data, and air quality data. Approximately 3000 data items are available for each county.

A. Air Quality Data

The PAREP project is unique in the way air quality was estimated at the county level. Previous nationwide analyses have averaged measurements from the monitoring stations within each county. Such a method is unsatisfactory because (a) many counties have no active monitoring stations; (b) many people live closer to stations outside their county than to stations within their county of residence.

A crucial task of the PAREP project was the creation of a nationwide directory, with reliable latitude and longitude coordinates for each active monitoring station. The existing EPA monitoring station directory file (with numerous errors) was combined with related files from several independent sources. Discrepancies were resolved by computer if possible, otherwise by consulting maps.

Yearly summary air quality data for 1974-1976, for nine pollutants, were obtained directly from the EPA SAROAD (Storage and Retrieval of Aerometric Data) data bank, and merged with corrected coordinates.

Figure 1 displays sample data from the resulting file: the three-year geometric mean value of total suspended particulate concentration is plotted as a circle at the station's location. The size of each circle indicates the relative pollutant concentration. Figure 1 illustrates the fact that air quality is poorer in the Los Angeles area than in other parts of California.

Next, air quality was estimated at the population centroid of each county as a weighted average of measurements from all nearby stations, whether in the county or not. The weight was taken to be

$$n(i) * \exp [-0.5 * (d(i)/d0)^{**}2]$$

where n(i) is the number of observations, d(i) is the distance from the county population centroid to station i, and d0 is an empirical parameter, taken to be 20 kilometers. The final choice for d0 awaits the results of further analysis.

Air quality estimates for larger geographic areas are calculated as population weighted averages of county values. The same method can yield estimates at an

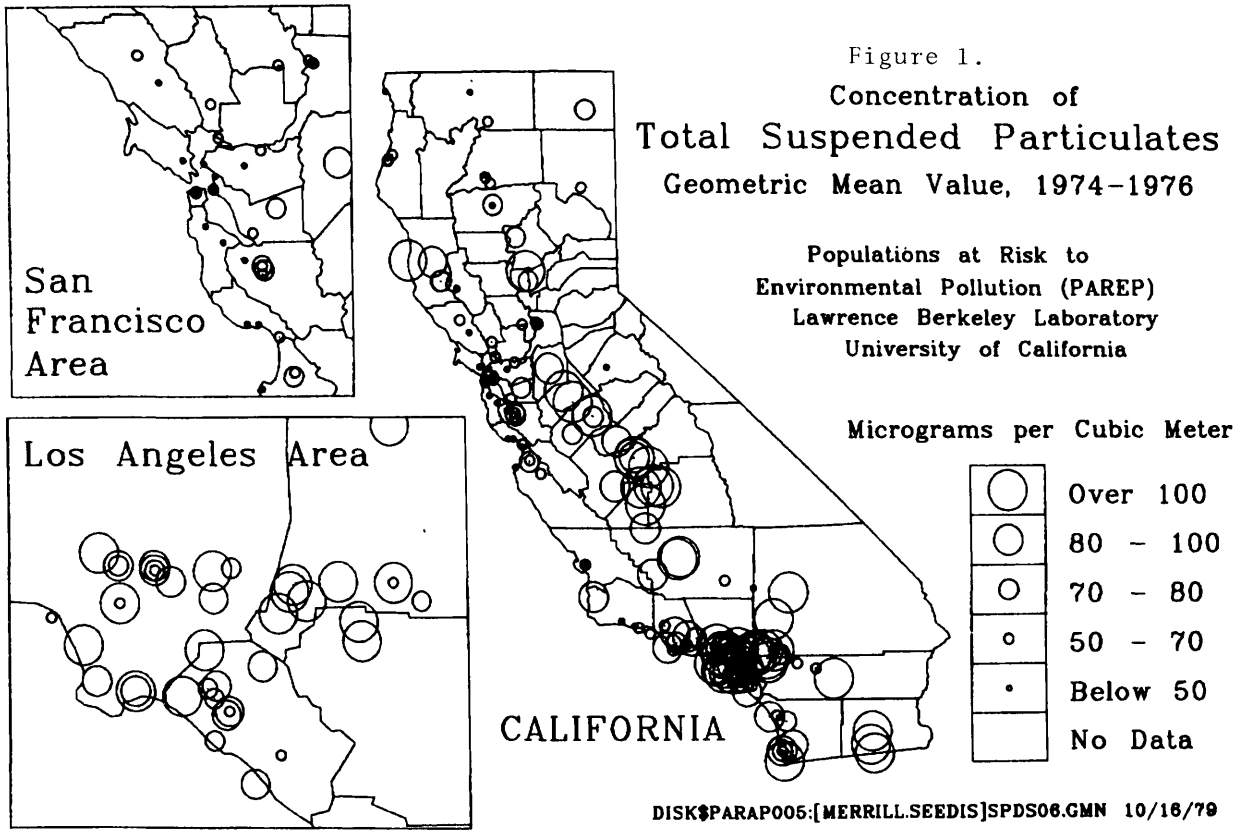


Figure 1.

Concentration of
Total Suspended Particulates
Geometric Mean Value, 1974-1976

Populations at Risk to
Environmental Pollution (PAREP)
Lawrence Berkeley Laboratory
University of California

arbitrarily fine level of detail, subject to the availability of real data. The density of measurements (effective full time stations per unit area) is provided in the data base, permitting the user to optionally discard estimates having large uncertainties. As a consistency check, conventional averages of the measurements within individual counties are also provided.

B. Mortality Data

The PAREP data base contains county level age-adjusted mortality rates by sex and race from two sources: (a) 1968-72 combined, 53 causes of death, from the University of Missouri; (b) 1950-69 combined, 35 cancer sites, from the National Cancer Insitutute.

"Standard scores" were calculated as

(county rate - U.S. rate)/(error)

where "error" is the absolute statistical error of the county rate, estimated as the county rate divided by the square root of the number of deaths (or the expected number of deaths, if no deaths occurred).

Table I.
Stomach Cancer in White Males
State of Arizona
U.S. Rate = 10.20

	1970 Population	1968-72 Annual Rate per 100,000	Standard Score
AZ PINAL	29983	15.02	1.43
AZ YAVAPAI	17982	15.26	1.16
AZ YUMA	28938	12.24	.65
AZ APACHE	3913	.00	-.00
AZ SANTA CRUZ	6429	10.11	-.02
AZ GRAHAM	7381	9.68	-.10
AZ GILA	12066	9.75	-.12
AZ PIMA	160936	10.10	-.12
AZ GREENLEE	5038	9.00	-.20
AZ MOHAVE	12588	8.36	-.49
AZ NAVAJO	11683	4.13	-2.18
AZ COCONINO	17359	3.99	-2.77
AZ COCHISE	30282	4.81	-2.89
AZ MARICOPA	448324	7.97	-3.61

Table I illustrates the results for stomach cancer among white males in Arizona for 1968-1972 (actually, four and a half years of data). Although the rate in Maricopa county (around Phoenix) is not as low as in four other counties, its deviation below the U.S. mean is statistically the most significant.

Figure 2, which also illustrates stomach cancer in white males, is typical of maps used to display geographic correlations in mortality statistics. Significant deviations above the U.S. mean are observed around Los Angeles, San Francisco, Sacramento, and Honolulu. Plotting standard scores rather than rates has the advantage that random fluctuations in small counties do not obscure statistically signficant trends in large cities.

C. Socio-Economic and other data

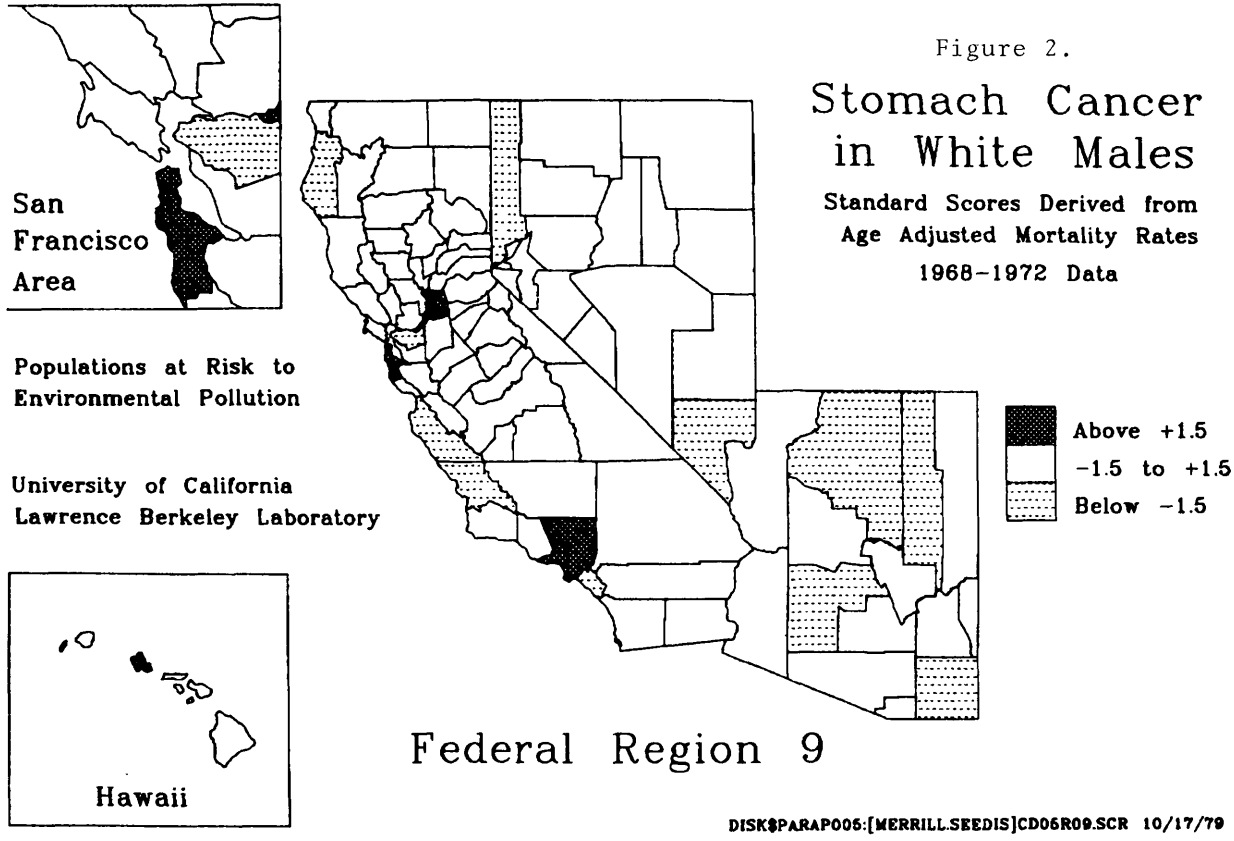
Indicators of socio-economic status (SES), including income, education, employment by industry and occupation, etc. were obtained from the 1970 Census. The PAREP data base includes corrected 1970 population counts by age, sex, race, and marital status for the purpose of normalizing mortality or morbidity rates. The Census Bureau's best available 1976 county population estimates are included. Related files provide subcounty detail on the geographic distribution of the U.S. population.

Most of the PAREP data base is being converted to a load format suitable for use in the SYSTEM 2000 Data Base Management System. The same data, and associated files, are installed in LBL SEEDIS (Socio-economic Environmental Demographic Information System), an interactive information system operating in a network of DEC VAX computers. In SEEDIS the PAREP data can be used in combination with files from other sources. Selected data, including data entered by the user, can be easily combined and displayed in tabular or graphic form, including maps.

III. Analysis

Analsis of the PAREP data base is in progress, in several areas.

Figure 2.


Stomach Cancer
in White Males
Standard Scores Derived from
Age Adjusted Mortality Rates
1968-1972 Data
San
Francisco
Area
Populations at Risk to
Environmental Pollution
University of California
Lawrence Berkeley Laboratory
Above +1.5
-1.5 to +1.5
Below -1.5
Hawaii
Federal Region 9
DISK$PARAP005:[MERRILL.SEEDIS]CD06R09.SCR 10/17/79

A. Estimation of Air Quality

This analysis considers the problems involved in geographic interpolation of annual average air quality measurements. Statistical and graphic techniques are used to investigate the validity of the averaging procedures described above. The basic criterion is that the model must accurately predict air quality at the position of a monitoring station, as a function of measurements from other nearby stations. The study will attempt to determine (a) the best value of the scaling parameter d0 described above (b) realistic standard deviation errors associated with the county air quality estimates.

B. Ecologic Patterns of Disease in the United States

The PAREP data base is an ecological data base - the basic unit is a group of individuals whose characteristics are known only on the average. Analysis of PAREP cannot be expected to identify cause and effect relationships between, say, air pollution and lung cancer. The most obvious fallacy is that 1975 air quality is being compared with 1970 mortality (an unavoidable choice imposed by the availability of data).

On the other hand, a large data base like PAREP can provide at relatively low cost a quantitative description of the relationships among a large number of variables. Any strong correlations not well understood would become candidates for analysis in a controlled case study.

A straightforward analysis technique is multiple regression, with cancer mortality as the dependent variable, and other variables (income, educations, air quality, etc.) as dependent variables. Patterns of disease are analyzed after removal or partial removal of the hypothesized linear influences of SES and air quality variables. Such an analysis was performed earlier (2,3) on a preliminary data base containing California data. Statistical limitations prevented any firm conclusions from being drawn. A similar analysis is being repeated for the entire United States.

Another project related to PAREP involves the analysis

of data from the Third National Cancer Survey, which recorded all incidences of cancer between 1969 and 1971, in nine areas of the United States. Individual case records, coded by census tract, have been merged with tract level SES data from the 1970 census, and tract level air quality estimates calculated as described above. Multiple regression and other techniques are being used to describe relationships between air quality, socio-economic status, and the incidence of certain histologic cancer types.

IV. References

1. Sacks, S.T.; Selvin, S.; and Merrill, D.W.; Building a United States County Data Base: Populations at Risk to Environmental Pollution; presented at the National Insitute of Aging, Bethesda, Md., June 26, 1979; LBL-9636, September 1979.

2. Merrill, Deane W. Jr.; Sacks, Susan T.; Selvin, Steve; Hollowell, Craig D.; and Winkelstein, Warren Jr.; Populations at Risk to Air Pollution (PARAP): Data Base Description and Prototype Analysis; LBL UCID-8039, August 1978.

3. Hollowell, C.D.; Sacks, S.T.; Selvin, S.; Levine, B.S.; Merrill, D.W.; and Winkelstein, W. Jr.; "PAREP: Populations at Risk to Environmental Pollution- Data Base Description and Prototype Analysis;" in Energy and Environment Division Annual Report 1978, pp. 152-155; LBL-8619 and UC-13; 1979.