LBL-9636 c.2

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

ENERGY & ENVIRONMENT DIVISION

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

BUILDING A UNITED STATES DATA BASE: POPULATIONS AT RISK TO ENVIRONMENTAL POLLUTION

Susan T. Sacks, Steve Selvin and Deane W. Merrill

October 1979

TWO-WEEK LOAN COPY

This is a Library Circulating Copy which may be borrowed for two weeks. For a personal retention copy, call Tech. Info. Division, Ext. 6782.

RECEIVED
LAWRENCE
BERKELEY LABORATORY

JAN 14 1980

LIBRARY AND
DOCUMENTS SECTION

LBL-9636 c.2

Prepared for the U.S. Department of Energy under Contract W-7405-ENG-48

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.

Presented at the Conference on Demographic and Health Information for Aging Research: Resources and Needs, National Institute on Aging, National Institutes of Health, Bethesda, Maryland, June 25-27, 1979

LBL-9636

BUILDING A UNITED STATES DATA BASE: POPULATIONS AT RISK TO ENVIRONMENTAL POLLUTION

Susan T. Sacks and Steve Selvin
School of Public Health
University of California

and

Energy and Environment Division
Lawrence Berkeley Laboratory

Deane W. Merrill
Computer Science and Applied Mathematics Department
Lawrence Berkeley Laboratory

October 1, 1979

The work described in this report was funded by the Office of Health and Environmental Research, Assistant Secretary for Environment of the U. S. Department of Energy under Contract No. W-7405-ENG-48.

INTRODUCTION

Scientists at the Lawrence Berkeley Laboratory (LBL), under contract to the U.S. Department of Energy (DOE), have undertaken a series of collaborative studies with researchers at the School of Public Health, University of California, Berkeley, on the effects of environmental pollution on health. A major study, Populations-at-Risk to Air Pollution (PARAP), was initiated in 1976 under funding by the U.S. Environmental Protection Agency (EPA). In October, 1978, this project was extended, under DOE funding, to consider other environmental hazards and was renamed Populations-at-Risk to Environmental Pollution (PAREP).

The PAREP project is divided into three main tasks:

1. Creation of an integrated data base containing socioeconomic and demographic characteristics, air pollution levels for several important pollutants, and disease specific mortality statistics for the U.S. on a county basis;

2. Determination of populations at risk to various pollutants; and

3. Analysis of possible associations between disease specific mortality and pollutant levels, taking into account socioeconomic and demographic variables.

CREATION OF THE PAREP DATA BASE

The integrated PAREP data base includes geographic, demographic, mortality and air quality variables collected and recorded for each county in the U.S. The scope of the data base is extremely broad, and a complete list of the variables is available upon request.

The county geographic and demographic data includes: (1) state and county code; (2) county area, geographic and population centroid; (3) vector of boundary points describing the county location by latitude and longitude; (4) total county population for 1970 and estimated total for 1975, race and age-specific (note: race = whites, blacks, non-whites); (5) a variety of U.S. Census variables; e.g., a total number of families, median school years completed, number of persons employed in industrial and occupational categories, age distribution, etc.

The data included in the PAREP data base is abstracted from the large quantities of data routinely collected by governmental agencies and is rarely available on an individual record basis. Most national data consist of tabulations for specific geographic areas, e.g., counties, census tracts, etc. The tabulated variables are comprised of aggregates of individuals and are usually called ecologic variables, since they reflect an average for some defined group. Several authors have discussed the problems of using ecological data. An often-quoted paper by Robinson* demonstrates that a product moment correlation can be misleading when calculated from pairs of ecologic variables, which are then interpreted as measuring the association among individuals. There

---

*Robinson, W.S., "Ecological Correlations and the Behavior of Individuals," American Sociological Review, 15, 351-357, 1950.

is general agreement that inferences drawn from ecological data lack the strength of studies based on individual records. However, if a striking association is noted between two ecologic variables, such as dollars spent on driver education and deaths from motor vehicle accidents among teenagers, it is difficult to dismiss the observation because the study is not based on individuals. Conversely, it is probably too strong to infer that spending more money on driver education would reduce the death rate among teenage drivers. The proper interpretation of an ecologically derived inference should lie somewhere between these two extremes. Historically, epidemiologists have hesistated to use ecological data partly because the conclusions are uncertain, and also because such studies lack the extensive technological methodology found in case-control studies or clinical trials.

Ecological data have several clear cut advantages over individual case samples. Ecological data are generally collected over a long duration and are usually coded and reported in a consistent manner. Normally, these data are easily obtainable at nominal costs in comparison to the high cost of other types of epidemiological data. In certain cases, ecological data are the only data available to investigate some types of phenomena. For example, air quality measurements are usually collected for geographic areas and not individuals.

DETERMINATION OF POPULATIONS-AT-RISK

Mortality rates are the most widely used ecologic variables in epidemiological investigations. Use of mortality data involves several well-documented problems including the diagnostic accuracy involved in recording the cause of death and the deceased designation of residence on the death certificate, both of which are potential sources of bias for the numerator of mortality rates. Use of population census estimates to provide the denominators of mortality rates can also be liable to biases such as underenumeration of specific subpopulations, e.g., young black males. Defining and estimating the size of "population-at-risk" is difficult for intercensal years. Conversely, mortality data reflect the aggregated health experience of a group typically defined by geographic area. In the case of mortality data aggregated for moderately large groups such as counties, rates will generally be stable, having a small sampling error, and will provide accurate estimates of rare disease frequency such as breast cancer in males. The precise interpretation of mortality rates as indicators of a community's health status has been widely debated, but little debate exists over the necessity of utilizing mortality data in an attempt to understand the disease process.

The mortality experience of each county is summarized in the PAREP data base by two sets of cause-specific average annual age-adjusted rates per 100,000 for males and females. The first set of mortality rates summarizes the years 1950-1969 for 35 site-specific causes of death due to cancer for whites and non-whites. The second set of mortality rates covers a 4 1/2 year period starting in 1968 and contains

the average annual age-adjusted mortality rates for 53 causes of death for whites and blacks. This data set was compiled from death certificates made available by the National Center for Health Statistics (NCHS). For both sets of mortality data, a calculated function of the age-adjusted rate is included, e.g., standard score. This translates the mortality rate into the number of standard deviations above or below the mean rate for the entire U.S. The purpose of this is to provide a statistical measure of mortality that essentially equalizes comparisons among counties regardless of population size.

For example, when no deaths occur in Alameda and Alpine counties in California, the cancer mortality rate for that site is zero for both counties. Nevertheless, these two rates do not accurately reflect the different risks of cancer since Alameda county has a population approximately 2000 times larger than that of Alpine County. In terms of standard deviations from the mean, zero deaths in Alameda county will be a substantial number of standard deviations from the U.S. mean for most cancer sites, but in Alpine county, which has a very small population, the standardized number will be small reflecting the fact that zero deaths in a small population is a likely event.

ANALYSIS

Yearly averages (1974, 1975, 1976) for seven air pollutants including total suspended particulates, SO_2, NO_2, CO, hydrocarbons, O_3, oxidants and non-methane hydrocarbons, are recorded in the PAREP data base for each county in derived summaries and for all active monitoring stations. For each station, the yearly averages are expressed using both arithmetic and geometric means; the standard deviations are included for

both. A frequency of measurement code, e.g., each hour, each day, is also included with an indication of the analytic measurement method.

Most analyses of air quality data, including standard published EPA reports, provide estimates by county or by Air Quality Control Region (AQCR) by averaging the estimates from all monitoring stations within the county or AQCR. Such an analysis ignores the actual locations of the monitoring stations as well as the distribution of the population. The PAREP data base contains mesurements of air quality that are derived using a different approach.

The county population centroid was calculated from the population distribution reported in the 1970 census. This same calculation can easily be performed for cities, census tracts, or any other political division. For the pollutant in question, the distance was measured from the population centroid to all active monitoring stations within 100 kilometers of the population centroid whether or not they were in the county. A weighted average was calculated in which each station i received a weight w(i) equal to:

$$w(i) = \exp(-1/2\ (d_i/d_0)^2)$$

Here d_i is the distance from the population centroid to station i and d_0 is a constant of the order of 20 kilometers. The empirical scaling distance of 20 kilometers was originally chosen based on annual average spatial variations of air quality. The "goodness of fit" of this scaling factor has recently been tested and has been found to work well for most pollutants. This weighted average is an indication of the air quality or pollution exposure experienced by the populations living in

each of the 3082 U.S. counties.

The completed data base contains not only the calculated values of air pollutant concentrations but also their corresponding weights. Thus, estimates of pollution exposures having a large uncertainty factor (i.e., no stations nearby and thus small values of w(i)) can be appropriately weighted in the statistical analyses. The choice of weights w(i) is equivocal. However, the individual station data values are maintained in the data base so that a user can combine station measurements into any desired summary measure.

Several errors in the data were encountered and had to be corrected. For example, errors were discovered in the latitude and longitude of air quality monitoring stations in the EPA Storage and Retrieval of Aerometric Data (SAROAD) site directory. Figure 1 shows the original monitoring station sites located in California. The same errors were propagated to the published EPA directories of air quality monitoring stations and to the Energy Data System (EDS). In order to correct these data, which are crucial to the PAREP project, computed routines were implemented to convert Universal Transverse Mercator coordintes to latitude and longitude.

The completed data base is currently being installed in a commercially available data base management system, SYSTEM 2000*. Implementation by means of a hierarchical data base management system makes the addition and retrieval of data elements relatively fast and uncomplicated, the only requirement for efficient access is knowledge of the

---

*produced by the MRI Systems Corporation in Austin, Texas.

System 2000 control language. Another important feature of the data base is its internal documentation. A description of each data element including definition, coverage, format, units, and data source is part of the data base.

A county level data base is somewhat problematical when focusing on interpretation of relationships. For example, there is no U.S. by county smoking data, which is an important factor in the study of disease, particularly cancer and heart disease. Another problem is the interpretation of the 1974, 1975, and 1976 air quality data in relation to 1968-1972 mortality data. Air quality measurements from the 1940's, 1950's, and the 1960's, should be used for study in relation to later mortality data. Such air quality data are not available for the entire U.S. or any large region. Consequently, the later air quality data have been used under the assumption that they reflect to some degree the environmental experience of most areas of the U.S. From this point of view the data base is certainly useful in "hypothesis generation."

PRELIMINARY RESULTS

Examples of some descriptions or "first looks" at the data base are included in Tables I and II. For all 53 causes of death in the 1968-1972 NCHS mortality data, all U.S. counties were ranked by standard score, separately for white males and white females. Tables I and II include state and county name, size of the white male or white female population and the standard score (see page 6), and average annual age-adjusted rate per 100,000. It is noteworthy in referring to the tables that for white males, 4 of the 21 counties in New Jersey appear in the top 50, and for white females, 5 of the 21 counties in New Jersey

appear in the top 50 of the more than 3000 U.S. counties. It should also be noted that Menominee county, Wisconsin, has an extremely high rate of stomach cancer among males. The importance of taking county size into account when comparing motality rates among counties becomes evident from these two samples. If the county size had not been taken into acccount, Menominee County would have been ranked first in Table I, while Kenedy, Texas, would have ranked first in Table II, since both counties have under 500 people.

Maps provide another descriptive tool, one that has been used extensively by the National Cancer Institute, the National Heart, Lung and Blood Institute, the National Center for Health Statistics, etc. Since the PAREP data base covers the entire U.S. by county, maps can be generated such as the one shown in Figure 2. Because the eye tends to focus on counties which have large areas when looking at maps of the entire U.S., a less deceptive presentation of PAREP data for the U.S. by counties is to show Federal Regions as shown in Figures 3, 4, and 5. These areas have a more uniform size and give a clearer picture of specific regions.

Prototype multivariate statistical analyses have been performed on California data in anticipation of completion of the entire U.S. by county data base. The principal technique in combining the major independent variables as predictors of disease is a multivariate regression equation. Monte Carlo methods are being used to study the validity of applying regression techniques to aggregated data such as county averages and medians. The combination of this theoretical work and the application of multiple regression analysis to the 3082 U.S. counties

will produce the first comprehensive look at national disease patterns while taking into account a series of socio-economic and environmental variables. Another approach which has been adopted by others and which will be used in analyzing the PAREP data is the strategy of "matching" counties on various demographic variables, a version of the case-control study, in an attempt to determine why certain counties are high for a specific cause of death. The results describing and analyzing the PAREP data base are in progress and will be published in the near future.

Table I. Stomach cancer mortality: white males, 1968-72.

	NAME		SIZE	SCORE	RATE PER 100,000
1	N.Y.	NEW YORK+ (3	906155.	9.699	15.100
2	N.J.	MIDDLESEX	274666.	9.196	18.610
3	ILL	COOK	2059414.	8.574	13.090
4	MICH	WAYNE	940960.	7.728	14.040
5	OHIO	CUYAHOGA	663764.	7.686	14.740
6	MASS	BRISTOL	208014.	7.257	17.830
7	N.Y.	ERIE	482892.	6.876	14.960
8	MINN	ST. LOUIS	106719.	6.623	19.910
9	IDAH	TETON	1166.	6.320	98.530
10	PA	ALLEGHENY	692856.	6.073	13.720
11	TEXA	REAGAN	1511.	5.981	83.640
12	CONN	HARTFORD	367347.	5.781	14.790
13	PA	CAMBRIA	88145.	5.583	19.210
14	PA	PHILADELPHIA	610654.	5.521	13.610
15	N.Y.	NASSAU	658343.	5.494	13.470
16	WISC	MENOMINEE	149.	5.446	223.050
17	MASS	ESSEX	300389.	5.136	14.710
18	OHIO	MAHONING	128719.	5.107	17.030
19	N.J.	PASSAIC	195831.	5.028	15.660
20	CONN	FAIRFIELD	354080.	4.890	14.160
21	R.I.	NEWPORT	49304.	4.855	20.670
22	CALI	SAN FRANCISC	246815.	4.676	14.730
23	N.J.	HUDSON	259176.	4.653	14.600
24	N.M.	SANTA FE	25935.	4.592	23.840
25	N.M.	RIO ARRIBA	10974.	4.579	31.090
26	N.D.	DUNN	2367.	4.555	54.900
27	HAWA	HONOLULU	142083.	4.544	15.990
28	R.I.	PROVIDENCE	265854.	4.529	14.430
29	N.J.	BERGEN	418486.	4.503	13.560
30	ALAS	ENTIRE STATE	131971.	4.493	16.140
31	COLO	SUMMIT	1361.	4.383	66.910
32	WISC	MARATHON	48213.	4.092	19.130
33	MASS	SUFFOLK	287268.	4.056	13.850
34	KY	HARLAN	16924.	4.015	24.960
35	MICH	MARQUETTE	32551.	4.002	23.820
36	OHIO	TRUMBULL	106778.	3.939	15.990
37	N.C.	BERTIE	4238.	3.888	38.730
38	WISC	DOUGLAS	21991.	3.852	22.630
39	TEXA	CROCKETT	1892.	3.792	51.820
40	N.M.	GUADALUPE	2475.	3.774	46.430
41	MASS	PLYMOUTH	159208.	3.756	14.730
42	N.H.	HILLSBOROUGH	107998.	3.658	15.550
43	MONT	BEAVERHEAD	4194.	3.504	36.050
44	MICH	MACOMB	305405.	3.487	13.250
45	GA	EFFINGHAM	5130.	3.485	33.450
46	N.D.	MERCER	3004.	3.441	40.190
47	IDAH	BINGHAM	13553.	3.429	24.290
48	MAIN	YORK	54382.	3.422	17.240
49	PA	WASHINGTON	98584.	3.410	15.420
50	CONN	NEW LONDON	111265.	3.322	14.990

Table II. Stomach cancer mortality: white females, 1968-72.

	NAME		SIZE	SCORE	RATE PER 100,000
1	ILL	COOK	2202565.	9.692	7.230
2	N.Y.	NEW YORK+ (3	762039.	6.952	7.710
3	N.M.	SAN MIGUEL	10801.	6.124	24.770
4	N.J.	HUDSON	283579.	6.338	8.840
5	MICH	WAYNE	993730.	6.004	7.060
6	N.J.	PASSAIC	212551.	5.806	9.260
7	TEXA	KENEDY	326.	5.742	111.520
8	N.Y.	NASSAU	699135.	5.560	7.270
9	KY	CLINTON	4095.	5.500	33.820
10	GA	CHATTAHOOCHE	3413.	5.364	35.780
11	KY	BRECKINRIDGE	7059.	4.692	23.740
12	WISC	FLORENCE	1583.	4.675	44.380
13	N.J.	MIDDLESEX	280498.	4.613	7.960
14	COLO	PITKIN	2996.	4.577	33.040
15	MINN	ST. LOUIS	111130.	4.308	9.370
16	IDAH	NEZ PERCE	14884.	4.112	16.330
17	OHIO	CUYAHOGA	721238.	4.049	6.640
18	N.D.	MCLEAN	5219.	4.028	23.710
19	N.Y.	ERIE	524798.	4.017	6.900
20	NEB	THOMAS	475.	3.960	65.880
21	KAN	COFFEY	3638.	3.841	25.800
22	KY	CUMBERLAND	3235.	3.791	27.360
23	W.VA	FAYETTE	22547.	3.743	13.390
24	NEB	DIXON	3712.	3.737	25.500
25	PA	WASHINGTON	104435.	3.655	8.830
26	N.D.	TOWNER	2247.	3.644	30.780
27	KY	MENIFEE	1962.	3.633	32.500
28	GA	WILKINSON	2557.	3.628	29.060
29	N.M.	TAOS	8360.	3.579	18.150
30	N.M.	QUAY	5482.	3.562	21.150
31	KY	GREENUP	16779.	3.523	14.140
32	N.J.	BERGEN	449811.	3.518	6.800
33	MASS	BRISTOL	229535.	3.521	7.490
34	PA	ALLEGHENY	763528.	3.468	6.380
35	MASS	MIDDLESEX	712884.	3.471	6.420
36	MICH	MARQUETTE	30622.	3.437	11.620
37	N.Y.	WESTCHESTER	421482.	3.425	6.810
38	N.D.	RAMSEY	6363.	3.416	19.380
39	PA	LACKAWANNA	123666.	3.369	8.270
40	N.M.	SANDOVAL	5317.	3.375	20.540
41	KY	POWELL	3825.	3.352	23.190
42	W.VA	CLAY	4610.	3.309	21.360
43	N.J.	UNION	249591.	3.307	7.260
44	PA	PHILADELPHIA	673523.	3.276	6.380
45	TEXA	CAMERON	73299.	3.264	9.080
46	N.M.	RIO ARRIBA	11317.	3.211	15.150
47	TEXA	LIBERTY	13300.	3.158	14.210
48	NEB	DAKOTA	6554.	3.145	18.250
49	TENN	BRADLEY	24829.	3.138	11.710
50	NEB	GRANT	508.	3.126	51.470

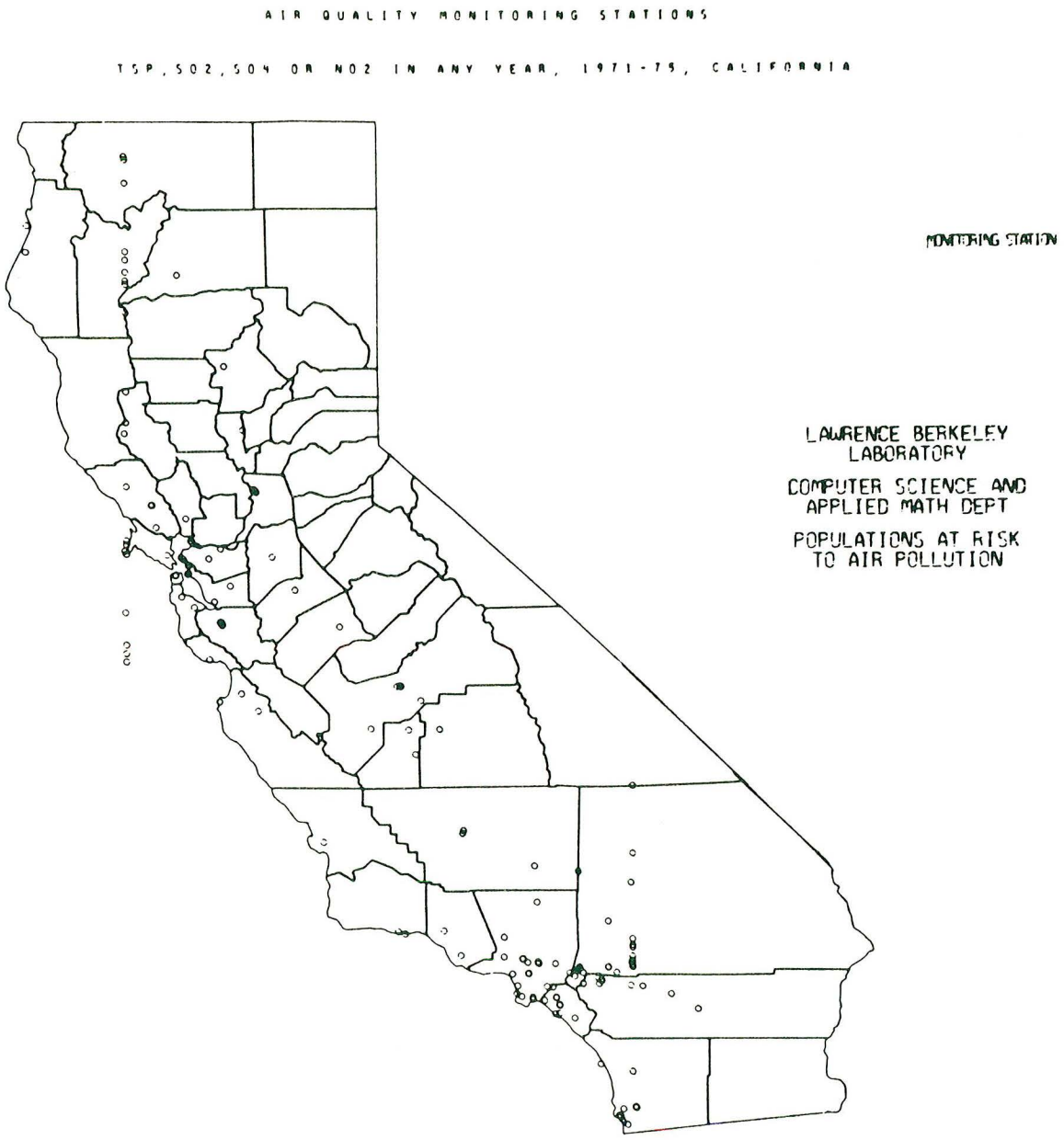

Figure 1. California Air Quality Monitoring Stations: TSP, SO_2, SO_4, O_x, NO_2 concentrations in any year, 1971-75.

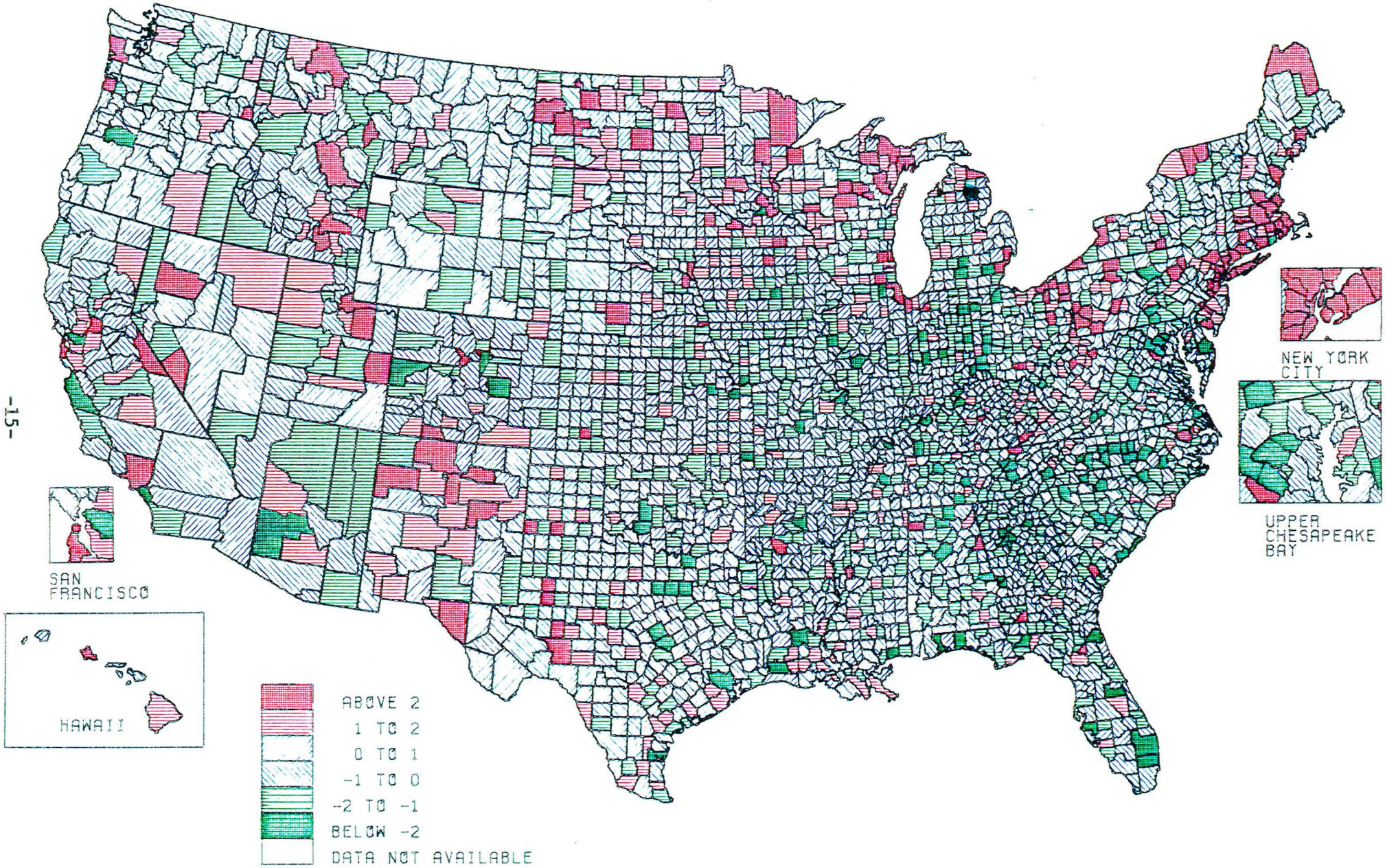

Figure 2. Stomach cancer mortality: white males, 1968-1972, United States by county.

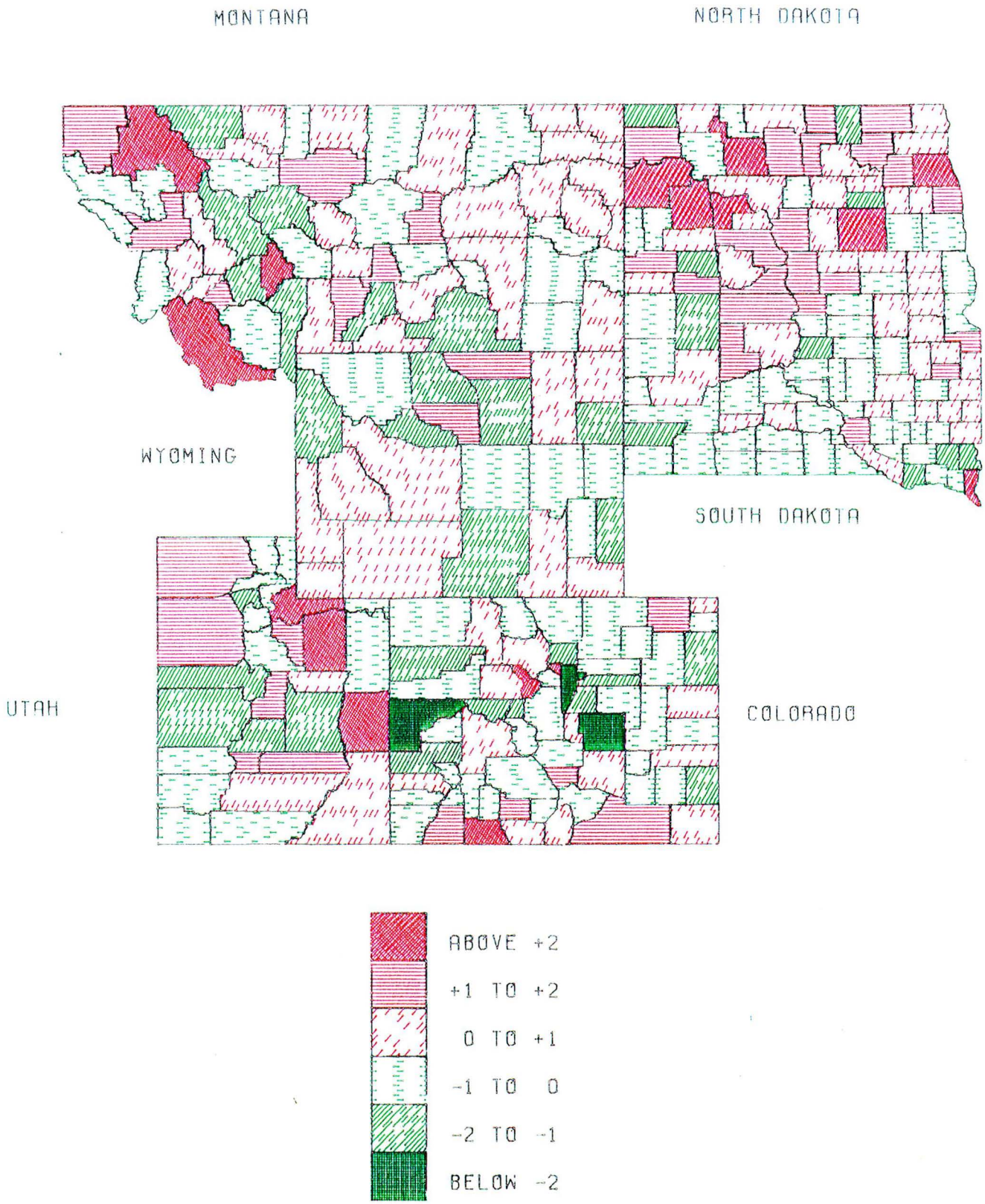

Figure 3. Stomach cancer mortality: white females, 1968-1972, Federal Region 8.

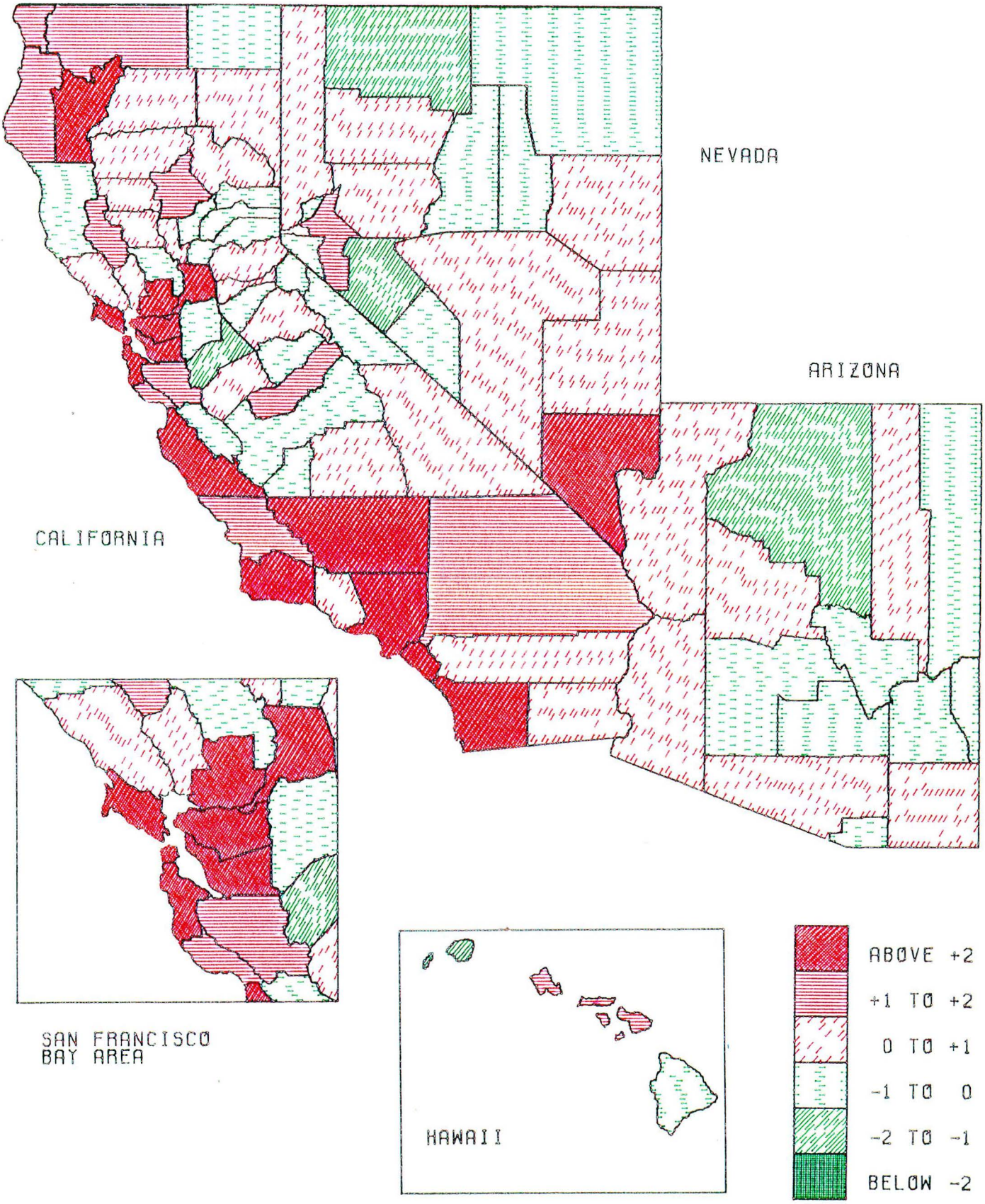

Figure 4. Respiratory cancer mortality: white females, 1968-1972, Federal Region 9.

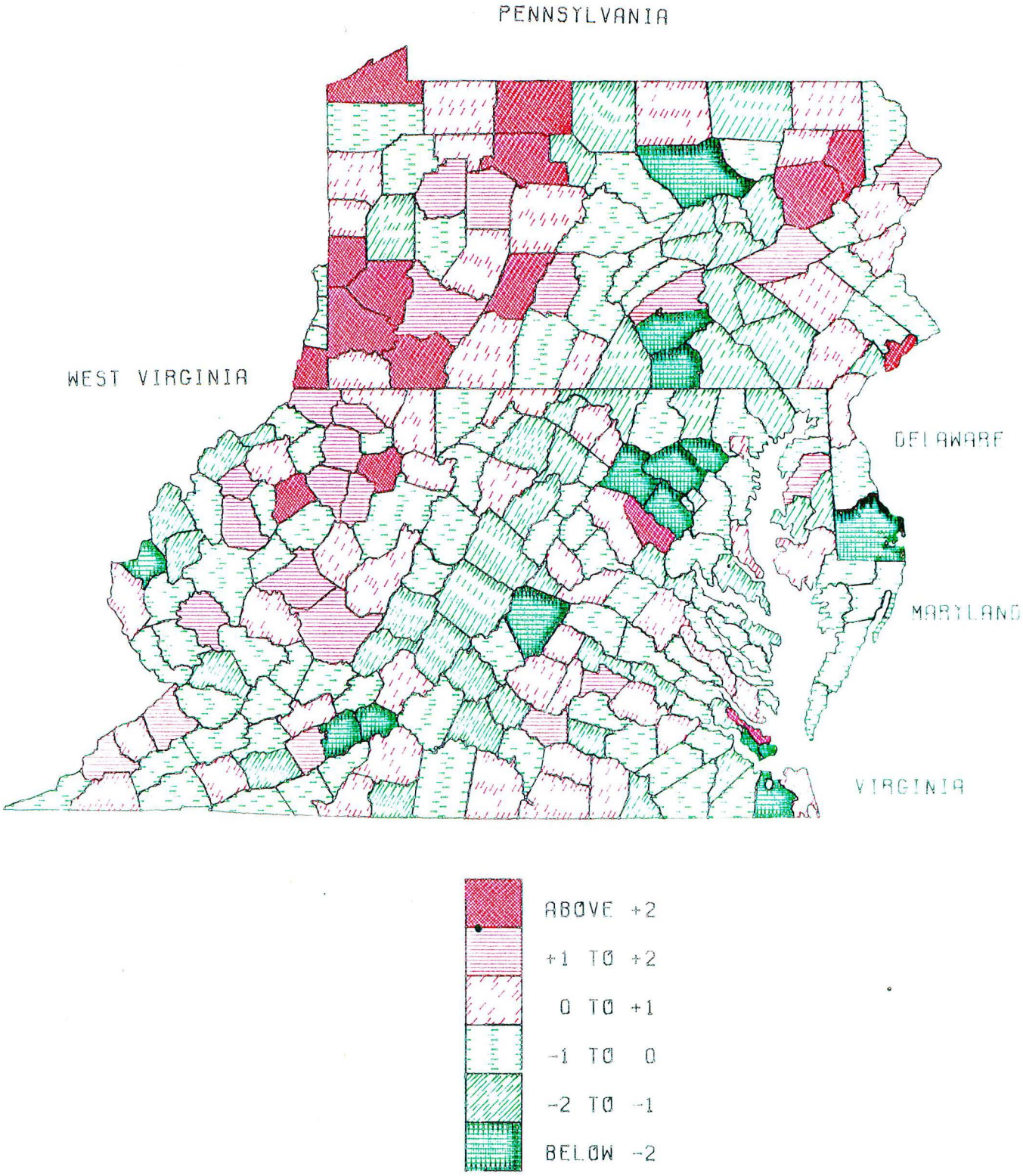

Figure 5. Stomach cancer mortality: white males, 1968-1972, Federal Region 3.