Lawrence Berkeley National Laboratory
Lawrence Berkeley National Laboratory

Title
ECOLOGIC REGRESSION ANALYSIS AND THE STUDY OF THE INFLUENCE OF AIR QUALITY ON MORTALITY

Permalink
https://escholarship.org/uc/item/1rw185tc

Author
Selvin, S.

Publication Date
1981-10-01

eScholarship.org Powered by the California Digital Library
University of California

LBL-12217
Preprint

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

Physics, Computer Science & Mathematics Division

RECEIVED
LAWRENCE
BERKELEY LABORATORY
NOV 4 1981
LIBRARY AND
DOCUMENTS SECTION

Submitted to the Journal of Environmental Economics and Management

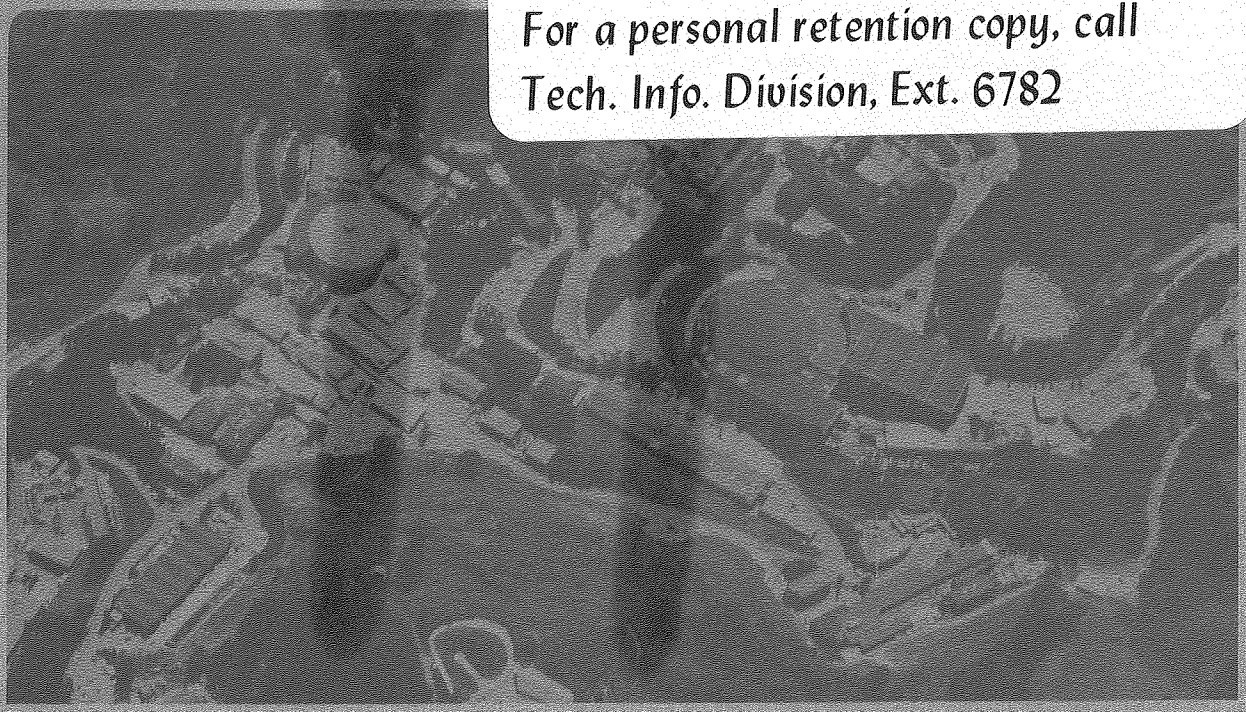
ECOLOGIC REGRESSION ANALYSIS AND THE STUDY OF THE INFLUENCE OF AIR QUALITY ON MORTALITY

S. Selvin, D. Merrill, L. Kwok, and S. Sacks

October 1981

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



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

LBL-12217
c.2

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-12217

ECOLOGIC REGRESSION ANALYSIS AND THE STUDY OF THE INFLUENCE OF AIR QUALITY ON MORTALITY

S. Selvin *, D. Merrill **, L. Kwok **, and S. Sacks ***

October 2, 1981

* University of California School of Public Health and Lawrence Berkeley Laboratory, Univ of Calif, Berkeley, CA

** Lawrence Berkeley Laboratory, Univ of Calif, Berkeley, CA.

*** University of California Medical Center and Lawrence Berkeley Laboratory, University of California, Berkeley, CA.

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, and by the Electric Power Research Institute under Contract No. EPRI/DOE 790702/800410.

ABSTRACT

This paper discusses the use of regression analysis applied to ecologic data for the study of the relationship between air quality and mortality. The first five sections describe the available ecologic data (mortality, age distributions, air pollution, and socio-economic status). The next four sections treat the application of regression techniques, where air quality and ecologic socio-economic measures are the independent variables and total mortality is the dependent variable. The last section presents results from analyses employing these same independent variables, with accident and stomach cancer mortality as the dependent variables.

Our purpose is to critically inspect the utility of regression methods to assess the health effects of air quality. Several, possibly acute, problems are noted. The existence of incomplete model bias, which potentially distorts regression coefficients, is discussed. Results from analyses of the same data at different levels of geographic detail show little consistency with each other or with those of previously published analyses. Separate analyses of different regions of the United States show no consistent pattern. Accident mortality, which has no known relationship with air quality, does in fact show some associations; stomach cancer, which has been shown to be associated with air pollution, yields no evidence of strong associations.

The results indicate that ecologic regression analysis techniques do not provide easily interpreted results, particularly with respect to the individual person. Potential biases are identified that could cause statistical summaries obtained from ecologic analyses to be useless or misleading. Certainly, regression analyses on nationally collected ecologic data cannot be used to usefully infer causal relationships between air pollution and mortality.

LEGAL NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or 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.

DISCLAIMER

This report was done with support from the Department of Energy. Any conclusions or opinions expressed in this report represent solely those of the authors 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.

TABLE OF CONTENTS

INTRODUCTION

The relationship of air quality to disease has been extensively studied and still remains a fundamental health issue. One approach to studying this relationship combines measures of pollution, rates of mortality and linear regression analysis applied to a series of defined geographic regions. Over a dozen research efforts fall into this class (e.g.,[1] and [2]; see Ricci [3] for a complete review). The present work attempts to use regression techniques to reproduce the results of others, particularly the work of Mendelsohn and Orcutt [1], and at the same time to delineate and discuss relevant methodologic issues. These issues include variable selection and bias, data completeness, data accuracy, and most importantly the validity of regression analysis applied to ecologic data (data aggregated on a geographic basis).

The assumptions and techniques of ordinary regression analysis are described in various places (e.g., [4]). The classic regression analysis postulates that a dependent variable is linearly related to a series of independent variables all of which are measured on the same observational unit. However, data are often not available on a unit basis but exist as statistical summaries of collections of units such as means, medians, percentages and rates. A natural extension of regression techniques is to analyze these collections of units with the same methods developed for linear regression analysis. The typical ecologic approach to the study of the influence of pollution on

disease consists of analyzing a series of geographic units, using total mortality rates, air quality data and census-derived socio-economic variables. Whether these types of data can be usefully employed to study air quality and health, and whether a regression model adequately reflects the complex relationships under study, are open questions.

The following report consists of a series of more or less independent sections concerning the problems surrounding ecologic regression analysis applied to the question of pollution and mortality. The exploration of issues relating to the ecologic data themselves, described in the first five sections, is followed by a critical application of regression techniques to two sets of ecologic data.

DATA AND THE LINEAR REGRESSION MODEL

Three governmental agencies, which are required to routinely collect specific types of information, provided the principal data used in this report. Mortality data were tabulated from death certificate files at the National Center for Health Statistics (NCHS). Air quality data were extracted from records maintained by the Environmental Protection Agency in the SAROAD (Storage and Retrieval of Aerometric Data) system. Each county in the United States was characterized with the use of variables from the U.S. County and City Data Book [5] of the U.S. Census Bureau. Some data concerning elevation and weather patterns were obtained from other sources [6].

The smallest possible common geographic area for comparing the three sources of data is the county, since the NCHS mortality data contain only the county of residence for each death certificate. This aggregation produced 3082 county records. (Not all government agencies use exactly the same county definitions. For example, independent cities in Virginia are sometimes considered separately and sometimes included with adjacent counties.) An ecologic analysis of this county level data set is presented in this report. In addition, it was desirable to analyze these same data at a geographic level other than the county, in order to compare our results with those of previous investigators [1]. The second geographic level chosen was the 1970 Census Public Use Sample (PUS) area, an aggregation of counties.

A complete description of the Census Public Use Sample is found in [7]. The geographic units of the PUS are groups of counties which are fairly homogeneous with respect to socio-economic status, and which divide the continental U.S. into 410 areas. Each area is defined so that its population exceeds 250,000 residents. For example, many large and sparsely populated counties in the western states like Montana, Colorado, Nevada, and Utah are aggregated to form single PUS areas, whereas large urban counties such as Los Angeles, Cook (Chicago), St. Louis, and Baltimore are themselves PUS areas. Comparisons between the county-level and PUS-level data are made throughout this report.

Total mortality is our fundamental variable of interest. An average annual mortality rate for each county is calculated by taking the number of deaths in each county for the period 1968 through 1972 (only half the deaths were recorded in 1972) and dividing by 4.5 times the 1970 county population. Rates for sex-, race-, cause-, and age-specific categories were similarly calculated. The analyses focused primarily on total mortality rates, rather than cause-specific rates, so that direct comparisons could be made with the other major ecologic investigations of air quality and mortality. (However, we present in addition some results from the analysis of stomach cancer and accident mortality rates.) Mortality rates are subject to bias from a variety of sources and these biases have been adequately discussed elsewhere (e.g., [2] or [8]). Although mortality rates have shortcomings in the study of disease, they are

the only population-based health data that exist on a national basis and as such are a valuable resource.

Seventeen variables reflecting the 1970 socio-economic situation in U.S. counties, four variables concerning county weather patterns, and one other variable (county elevation) serve as the measurements of variation associated with mortality rates that is not directly related to air pollution. A list of these variables is found in appendix A. The essence of a multivariate approach is the isolation of effects of specific variables (e.g. air pollution) from the influences of variables not of primary interest ("control" variables). The 22 variables listed in appendix A serve this "control" function.

The air quality data consist of measurements on three pollutants -- total suspended particulate (TSP), sulfur dioxide (SO2), and nitrogen dioxide (NO2). These values were extracted from data collected at 6625 monitoring stations operating during the three-year period 1974 through 1976. County-level air pollution estimates were interpolated from average values at individual monitoring stations. The air pollution estimates for the 410 PUS areas are population-weighted averages of the county level values. A detailed discussion of the air quality data and interpolation methods is contained in the section of this report entitled "Air Pollution Data."

The analysis of the total mortality rates employing 22 "control" variables and 3 air pollution measurements is

restricted to white males and white females 45 to 54 years of age. The analysis of other racial groups is not practical since too few deaths occurred during the period 1968-72 to calculate stable county-level mortality rates nationwide. Analysis of other age-specific categories provides little additional information since mortality rates are highly correlated among age categories. Furthermore, the 25 independent variables have the same values for each county or PUS level analysis regardless of the age category being considered. The only new information contained in an age-specific analysis comes from the age-specific mortality rates themselves, which obviously differ within a geographic area. However, the average annual age-specific mortality rates for the four age categories 35-44, 45-54, 55-64, and 65+ increase fairly linearly (actually geometrically), which implies that the age-specific analyses will differ very little in statistical significance.

The basis for a multivariate regression analysis is a linear model represented symbolically as

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + e_i$$

where b_1, b_2, ... b_k are regression coefficients associated with each independent variable x_1, x_2, ...,x_k. The dependent variable y_i is stochastic since the value represented as e_i is assumed to be a random variable with specific properties. For valid inferences to be made from the multiple regression structure, the following must be true:

1) The e_i values are uncorrelated

(correlations $(e_i, e_j) = 0$);

2) The e_i values have a normal distribution;

3) The variance of the e_i values is constant (variance $(e_i) = \sigma^2$).

The underlying structure of a regression analysis applied to mortality data (the dependent variable) can be investigated by checking for violations of the assumptions. Statistical procedures do not generally exist to prove that the assumptions are fulfilled. Statistical summaries of the male and female mortality rates for both county and PUS data sets are given in Table 1. The measures of skewness and kurtosis reflect the structure of the population under investigation and for normally distributed data have expected values of zero. The 99.5% critical values are also given in Table 1. As can be seen, the observed values of skewness and kurtosis in all four columns are large in the sense that they are extremely unlikely to represent random deviations from zero. The measures of skewness and kurtosis indicate that mortality rates for ages 45-54 are skewed to the right (skewness >0), and the probabilities associated with extreme rates are larger than expected from normally distributed data (kurtosis >0). However, the assumption of a normally distributed dependent variable is generally considered as not very critical to a regression analysis. Inferences made from approximately normally distributed data such as these mortality rates are not likely to be extremely misleading. It should be noted that the assumptions about the structure of the dependent variable do not affect the

estimates made from the regression analysis but rather influence the statistical interpretation of these estimates (e.g., significance probabilities or p-values).

A more critical issue involves the assessment of the residual values (i.e., the observed dependent variables minus the values predicted from the estimated regression equation [$y - \hat{y}$]). Data that reflect independent, normally distributed fluctuations about a linear model with the same variance (i.e., which satisfy the basic assumptions of a linear regression analysis) produce residual values that vary randomly about a mean of zero and have negligible relationships to the independent variables. Figures 1 through 4 show in standard deviation units the residual values from the regression analyses of males and females (discussed in detail later) plotted against the rank of county and PUS population sizes. The county-level analyses yield residual values that appear to be random deviations from a linear model except for the least populous rural counties, where extreme (beyond ± 3 standard deviations) residual values are observed. That is, no evidence exists to reject the hypothesis that a linear model adequately describes the county level mortality patterns, except for a few counties with the smallest populations. The PUS residual values show no trends with population size, which indicates that no strong evidence exists for violation of the basic linear regression assumptions for the 410 PUS areas. Extreme values (beyond ± 2 s.d.) occur with expected frequency (about 5%) and have no geographic pattern for county and PUS

regression analyses. Similarly, residual values plotted against other independent variables (not presented here) also appeared to be randomly distributed. Since no strong evidence exists that mortality rates do not adequately fulfill the requirements for a dependent variable in a multiple regression analysis, transformations were not considered necessary, and each mortality rate was weighted equally in the following analyses.

AGE DISTRIBUTION

Mortality rates increase geometrically after age 35, which complicates any comparison of groups with different age patterns. The 3082 U.S. counties do indeed differ with respect to age distribution. Most dramatically, large numbers of older individuals reside in Florida. Charlotte, Citrus, Highlands, Lake, Manatee, Martin, Pinellas, and Sarasota are Florida counties where the average age of the male and female residents exceeds 40 years of age (U.S. 1970 average = 33.1 years). The counties with the youngest average ages are found in Colorado (e.g., Adams -- males = 26.8 and females = 27.3; Lake -- males = 27.5 and females = 27.6). The total U.S. age distribution for males and females is shown in Table 2, along with the standard deviations associated with the county level age distributions.

The correlations (Table 3) among the percentages of male residents in each of eleven age categories reflects the strong interrelated patterns of the U.S. county level age distributions (only males are presented since the correlations for females are essentially the same). Consecutive age categories are highly correlated (r large and positive) for both males and females. The proportion of older county residents ($\geq$ 65) is negatively correlated with the proportion of younger residents ($\leq$ 14) for both sexes. For example, the proportion of residents greater than 85 years old is negatively correlated with the proportion of children under one year of age ($r = -.173$ for males and $r = -.205$ for females). These high correlations confirm the typical U.S.

pattern of single family units with children and older individuals without children residing to some extent in different geographic areas.

The eleven categories reflecting the county age distributions for both males and females can be adequately summarized by two principal components, which are defined in appendix C. These two components reflect 75.5% (male) and 78.1% (female) of the observed variation in nationwide age distributions. That is, the values of these two linear indices capitalize on the interrelationships in the county age structure to efficiently reflect differences in age distribution. The coefficients associated with these two indices show that the first principal component is dominated by the influence of the proportions of individuals under age 14, whereas the second component essentially measures the population over 55 years old. In the following analyses the percentage of children less than 5 years old and the percentage of individuals over 65 years old are used to summarize the county and PUS level age distributions. These two measures, in particular the proportion of individuals over 65, have been consistently used in other ecologic regression analyses (e.g., [2]). The principal components indicate that little significant information is gained by adding other measures of age distribution.

The bivariate nature of the U.S. age distribution is not widely recognized. Two previous ecologic regression analyses relating air pollution to health ([1] and [2]) employed the percentage of older individuals residing in a geographic

area as an independent variable and were criticized for a "rather casual choice" [9]. However, this type of simple summary appears to be suprisingly efficient and, with inclusion of measures of the proportion of both young and old residents, adequately summarizes the entire age distribution of a geographic area.

URBAN VERSUS RURAL MORTALITY

Urban/rural differences in mortality experience have been observed for specific areas of the United States [10] and for specific diseases [11]. Urban excesses in overall mortality are consistent with the hypothesis that air pollution has a measurable and general effect on the health of city dwellers where air quality is the poorest. This possible relationship has been noted by others investigating the influences of air quality on health (e.g. [1] and [2]). County level data provide an opportunity to explicitly study urban/rural differences in total mortality.

The total mortality pattern of each county is summarized by the expectation of life at birth, which is calculated from age-specific total mortality rates. Expectation of life is defined as the average number of years that would be lived by a cohort of people experiencing a specific pattern of mortality. The age-specific total mortality rates (1968-1972) for each county produce a single expectation of life. (For a general development, see [12].)

A brief statistical summary of the U.S. expectation of life for males and females is given in Table 4. The counties with the longest expectation of life (lowest overall mortality) for both males and females are located in midwestern states (Wisconsin, Minnesota, North Dakota, South Dakota, Nebraska, and Iowa (Figures 5 and 6). This pattern of low total mortality has been repeatedly observed and is thought to be primarily related to the low frequency of

heart disease among specific ethnic groups.

It is sometimes assumed that the low mortality of the rural midwest implies high mortality in the urban parts of the nation. For the years 1968-1972 the counties with the lowest expectation of life (highest mortality) are also small and rural. Males have the shortest expectation of life in the southeast (particularly the Atlantic coast counties of North Carolina, South Carolina and Georgia) and in the Appalachian mountains. The counties with the lowest expectation of life for females are scattered more or less randomly throughout the U.S. but are also small and rural (Figures 5 and 6).

In summary, no strong association is found between life expectancy and two measures of urbanization -- population density and percentage of urban area within a county. The correlation coefficients are:

	population density	% urban
males	r = .141	r = .007
females	r = -.025	r = -.060

A more direct evaluation of counties with low expectation of life is achieved by ranking the U.S. counties from low to high with respect to life expectancy. The ten lowest ranked counties are listed in Table 5A (males) and Table 5B (females) along with the county populations and expectation of life. For both sexes, no urban county appears in the lowest ten counties. In fact, only two counties with large

urban populations are observed in the 200 lowest ranked counties (St. Louis City, MO and Schuylkill, PA). Furthermore, among the 100 counties with the lowest expectation of life, more than 50% have no "urban centers" (defined by the U.S. Census Bureau as a place with more than 2,500 inhabitants), indicating that these counties are extremely rural.

Taking into account the economic status of the U.S. counties produces a more meaningful picture of the relationship between expectation of life and degree of urbanization. When counties are stratified by median family income (Table 6), only the wealthiest counties (those with less than 10 percent of the families earning less than $3,000) show no significant association between mortality and urbanization. The correlation coefficients measuring the association between expectation of life for these wealthier counties and population density are $r = -.070$ (males) and $r = .003$ (females). For counties with greater than 10 percent of the families earning less than $3,000 (Table 6), a negative correlation is observed between expectation of life and population density. This negative correlation becomes stronger as the percentage of families earning less than $3,000 increases. That is, shorter life expectancy is associated with more densely populated counties when the obscuring effect of economic status is held fairly constant. The strongest association is observed when more than 30 percent of the families earn less than $3,000 ($r = -.287$ for males and $r = -.258$ for females). An almost identical relationship (not shown here) is observed between expectation of

life and the percentage of urban area within a county.

The urban/rural differences in life expectancy among the U.S. counties are largely a function of economic status. This fact demonstrates the critical necessity of controlling for confounding variables such as economic status, in order to achieve a clear picture of any influence of air quality.

AIR POLLUTION DATA

Data from 6625 air quality monitoring stations active during the years 1974-76 produced the basic pollution data. These station values were combined in weighted averages to provide a summary air pollution measurement for each county. The weight for a specific station i is

$$w_i = f_i \exp(-0.5\, d_i^2 / d_0^2)$$

where f_i is the fraction of time (proportional to the number of observations) that station i was active, d_i is the distance from the monitoring station to the county geographic centroid, and d_0 is a scaling parameter taken to be 20 kilometers. The geometric mean level for a county is estimated as

$$\exp \frac{\Sigma\, w_i \log x_i}{\Sigma\, w_i}$$

where x_i is the geometric mean of measurements from station i. Note that stations contributing to the county estimate are not necessarily located within that county. The choice of the weighting function and, particularly, the selection of the scaling parameter d_0 result from subjective evaluations involving data availability, known data inconsistencies, average county size, and pollution dispersion effects. Various choices of the scaling parameter were considered; a detailed analysis is still in progress. The choice of d_0 = 20 kilometers permits consistent and reasonable estimates of air quality levels for all counties with, and many counties without, active monitoring stations located within the

county boundaries. It appears that a smaller value, d_0 about 10 kilometers, is appropriate for small area studies where monitoring station activity is rather high [13]. Estimates are calculated only if an active monitoring station is present within 60 kilometers of the county centroid (i.e. d_i < 60 km). It was not possible to completely investigate the validity and accuracy of the EPA pollution data, although many errors were found and corrected. However, averaging three years of data removes much of the variation and bias from cyclical sources such as weather patterns and seasonal trends. These three-year averages are constructed from the best available data and, hopefully, reflect the general air quality level of a county.

To quantify the amount of data available, the monitoring density, evaluated at the population centroid of each county and expressed as effective full time stations per unit area, is calculated as

$$density = \frac{\Sigma\ w_i}{2\pi\ d_0^2}$$

where w_i and d_0 are the same as defined previously. (The monitoring density is considered to be zero if no active stations are present within 60 km of the county centroid.) The monitoring density, a continuous function, is normalized so that its area integral over the entire U.S.,

$$\int_{US} density\ dA$$

is equal to the total number of effective full time stations,

$$\sum_{i=1}^{i=6625} f_i$$

The monitoring density for many counties and pollutants is zero; no estimates of pollutant concentration are available for these counties. Only three pollutants (TSP, SO2, and NO2) yield sufficient coverage of the United States to be included in the analysis without recourse to statistical missing value procedures. The total number of counties with non-zero monitoring densities for the three pollutants (TSP, SO2, NO2) is 1763 of a possible 3082 (57.2%). The inclusion of other pollutants would have caused the loss of a larger number of counties. The measurements of SO2 and NO2 levels are recorded either for one-hour sampling intervals or 24-hour sampling periods depending on the sampling methods used at each monitoring station. For the purpose of this analysis, it is assumed that regardless of the sampling interval or method, the air quality measurements are unbiased estimates of the pollution levels, and the two sampling intervals are treated as equivalent. Each station's mean value is weighted by the fraction of time the station was active, implying that each 24-hour measurement receives 24 times the weight of a one-hour measurement. Figures 7 through 11 show for each pollutant and sampling interval the distribution of the monitoring stations for the contiguous 48 states. There were a total of 6625 stations active during 1974-1976. Of these, 5473 measured TSP, 3491 measured SO2 (1051 in one-hour and 2440 in 24-hour intervals), and 2149 measured NO2 (353 in one-hour and 1796 in 24-hour

intervals).

The effective monitoring density is high in densely populated areas and low in sparsely populated areas. Table 7 gives the effective monitoring density (full-time stations per 1000 square kilometers, evaluated at the county population centroid) associated with large (> 500,000 persons), medium (10,000 to 500,000 persons), and small (< 10,000 persons) counties. The monitoring density between counties with the large and small populations differs by a factor of 40. Although the geographic area of the U.S. is not well covered by air quality monitoring stations, the vast majority of the U.S. population lives within 60 kilometers of one or more active stations. More precisely, only 1.5 percent of the U.S. population lives more than 60 kilometers from a station that measured TSP during 1974-76 (i.e., in counties where no estimate of TSP level is available). The same figures for SO2 and NO2 are 7 percent and 11 percent respectively. The geometric mean concentrations of the three pollutants for each population size class are also given in Table 7, along with average minimum and maximum values. TSP levels do not differ much between counties with large or small populations, whereas the SO2 and NO2 measurements show a more than twofold difference.

A tendency for levels of different pollutants to be positively associated is not surprising. For the county level data, the correlations are: r = .312 (TSP-SO2), r = .168 (TSP - NO2), and r = .424 (SO2 - NO2). The same correlations for the PUS data are similar: r = .290 (TSP - SO2),

r = .160 (TSP - NO2), and r = .442 (TSP - SO2). One implication of this rather strong association among pollutants is that these three air pollution measurements can be viewed collectively as an index of general air quality.

The estimates of air pollution levels for the PUS areas are weighted averages of the values for the counties located within the boundaries of the PUS areas. Two choices of weights were considered. The county values can be weighted by their effective monitoring densities, reflecting the amount of data associated with each pollutant value. Alternatively, the county values can be weighted by the populations potentially exposed to each pollution level. The justification for the latter strategy is that in a study of human health, exposure is a critical element and should be taken into account as best possible. In fact, the results obtained from these two approaches hardly differ. The correlations between the data-weighted estimates and the population-weighted estimates are r = .900 (TSP), r = .921 (SO2), and r = .923 (NO2). The population-weighted values were used in the following analyses.

There are 410 PUS areas. Employing weighted averages of county values produced 386 areas (94.1%) with valid estimates for levels of all three pollutants. The monitoring densities and the average geometric means are shown in Table 8. The monitoring densities are intermediate between the values observed for counties with medium and large populations (Table 7). The pollution levels are also in the intermediate range and are somewhat less variable than in

the county level data. PUS areas are defined to be rather homogeneous areas with more than 250,000 inhabitants, which implies in most cases that the air quality levels should resemble those of counties with large populations.

PUBLIC USE SAMPLE DATA

The authors of [1] used SES (socio-economic status) variables derived from the 1970 Census Public Use Sample (PUS). Use of the PUS is a potential source of efficiency but also of bias and statistical error. The Public Use Sample is a 1% sample consisting of roughly two million U.S. census records; data are available only down to the level of the 410 PUS areas, which are county aggregates. The PUS data were sampled in a complicated stratified manner so that computation of sampling errors is not easily accomplished. (The user's guide [7] for the PUS contains a series of expressions for approximating the sampling errors.)

In this section we compare two corresponding data files derived separately from the PUS and the 1970 Census Fourth Count summary tabulation. Direct comparison between the PUS data and the Fourth Count data aggregated into PUS areas will show any bias that may exist, and the magnitude of sampling errors. The Fourth Count tabulation, derived from a 20% sample, provides data down to the census tract level, but due to limitations in the mortality data there is no advantage in going below the county level.

In Table 9 we compare values of eighteen variables, at the geographic level of the 410 PUS areas. Column 1 contains data derived from the 1% PUS sample; column 3 contains corresponding data from the Fourth Count tabulation, aggregated from the county to the PUS level. The mean values and overall percentages show extremely close agreement for the

410 PUS areas. To further describe the differences among the PUS and aggregated Fourth Count data, Table 10 gives the absolute differences, relative differences and slopes (PUS values plotted against aggregated Fourth Count values). A slope of 1.0 indicates that no consistent bias exists between these two measures of the same quantity. The correlation coefficients summarizing the variability and the linear association between the two sources of data (not shown) are all greater than .98 except for the percentage of children less than one year of age ($r = .86$). No slope deviates from 1.0 by more than .04 units except, again, the percentage of children less than one year of age. Comparison of these summaries leaves little doubt that the aggregated county data are faithfully reproduced by the sample values from the PUS. Although no appreciable differences exist between PUS data and aggregated Fourth Count data, the following PUS analyses involve aggregated Fourth Count data since they are more easily accessible and have a less complicated origin.

The preceding analysis confirms only that the PUS file provides accurate and unbiased estimates of SES variables at the level of PUS areas. Elsewhere in this report we discuss the much more serious objections to analyzing data at the PUS rather than the county level; namely, the "ecologic fallacy" and biases incurred through aggregation.

INCOMPLETE MODEL BIAS

Estimates of regression coefficients from a linear regression analysis are affected by the choice of variables included in the linear model and excluded from the analysis. Only when all relevant variables are measured (full model) are the estimates of the regression coefficients unbiased. For example, if a three-variable linear model represented by

$$y_i = a + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + e_i$$

describes the data under consideration but a two-variable version of the model

$$y_i = a + b_1x_{1i} + b_2x_{2i} + e_i$$

(with x_3 omitted) is used in the analysis, then the estimates of the coefficients b_1 and b_2 associated with x_1 and x_2 will be biased (i.e., the expected values of the estimates are not equal to b_1 and b_2 in the full model). This phenomenon is called incomplete model bias.

Claims have been made that this type of bias should be conservative [1], which implies that most estimates of the regression coefficients are always smaller when models are incomplete. Furthermore, differences found to be significant using conservative estimates from incomplete models would be even more significant if more complete models were used. Others claim that incomplete model bias exists only when the variables not included in the analysis are related to both the dependent variable and the independent variables [2]. For example, variables related to disease outcome but not to

air quality, such as cigarette smoking, will not affect the regression coefficients associated with air quality since smoking and air qualify are not directly related. The issue of incomplete model bias is important since most applied regression analyses necessarily omit some important sources of variation.

The following numerical example, for which an algebraic proof is given in appendix B, illustrates incomplete model bias. This example is constructed to demonstrate that incomplete models do not necessarily produce conservative estimates and that generally all regression coefficients are affected when a variable is not included in the analysis.

Consider three variables (x_1, x_2, and x_3) with the following variance-covariance structure:

	x_1	x_2	x_3
x_1	$\sigma_1^2=16$	$\sigma_{12}=6$	$\sigma_{13}=1$
x_2	-	$\sigma_2^2=4$	$\sigma_{23}=2$
x_3	-	-	$\sigma_3^2=9$

If the linear model is given by

$$y_i = 10 + x_{1i} + 2x_{2i} + 4x_{3i} + e_i$$

then the covariances $\sigma_{y1} = 32$, $\sigma_{y2} = 22$ and $\sigma_{y3} = 41$ result. If the variable x_3 is omitted from the analysis, the biased estimates of the regression coefficients are $b_1 = -.14$ and $b_2 = 5.7$. Note that b_1 is substantially reduced and b_2 is substantially increased relative to the true values $b_1=1$ and $b_2=2$. The variances of the estimates made from the two-variable model and a sample of size n are

$\sigma^2_{b_1}$=20.5/n and $\sigma^2_{b_2}$=82.2/n (assuming σ^2_e=25) which are about five times larger than the variability under the full model ($\sigma^2_{b_1}$=3.8/n and $\sigma^2_{b_1}$=17.2/n). An increase in variability associated with incomplete models will always occur (see Appendix B) as long as the variable omitted is related to any of the independent variables or the dependent variable. An exact expression for the amount of increase is given in appendix B. If x_3 is unrelated to x_1 but remains associated with the dependent variable (e.g., σ_{13}=0 and σ_{y_3}=40), then the following estimates result (with x_3 omitted): b_1= -.71 and b_2= 6.6, with variances $\sigma^2_{b_1}$=21.1/n and $\sigma^2_{b_2}$=95.8/n. Again, these estimates hardly resemble the true parameters.

This numerical example demonstrates two points:

1) Incomplete model bias associated with the regression coefficients depends on the covariance structure and can either increase or decrease when variables are omitted from the analysis. Furthermore, statistical tests (t-test or F-test "to remove") of the regression coefficients depend on the estimated variability which always increases when a model is incomplete.

2) The fact that a variable is not related to some of the independent variables does not remove the potential for other estimated coefficients being affected by incomplete model bias.

In general, substantial incomplete model bias occurs when the variable omitted from the analysis is highly variable, unrelated to all independent variables, and strongly related to the dependent variable. The magnitude of incomplete model bias in many applied situations will be small but nevertheless exists at an unknown level.

THE SQUARED MULTIPLE CORRELATION COEFFICIENT

Fundamental to regression analysis is the squared multiple correlation coefficient, typically symbolized by R^2. This single number indicates the reliability of a linear model as a summary of a set of data. If $R^2 = 1$, a linear function totally explains the variation in the data. If $R^2 = 0$, a linear model is useless as a summary. For values of R^2 between zero and one the interpretation is not as simple. The difficulty lies in the fact that R^2 is a complex composite measure of the fit of the data to a linear model which combines the variation of the dependent variable, the variation of the independent variables, and the values of the regression coefficients into a single number.

In the case of an ecologic regression analysis with mortality rates as the dependent variable, a choice must be made for the definition of the mortality rate. For example, mortality rates can be calculated for a variety of age- or sex-specific categories. This choice affects the value of R^2. Table 11 gives the squared multiple correlation coefficients for fifteen choices of the dependent variable, for the same set of independent variables -- namely the 22 control variables listed in appendix A plus the three air quality measurements. The smallest values of R^2 are associated with rates classified by ten-year age intervals among females (i.e., $R^2 = .424$ for females 55-64). At the other extreme, if the crude mortality rates (combining both sexes and all ages greater than 35) are used, the R^2 value is

.859, which implies that a linear model "explains" over 85% of the observed variation in the crude mortality rate.

Not surprisingly, general mortality patterns are more predictable from ecologic variables than are specific patterns, since mortality rates are known to depend upon a variety of general characteristics. Increased values of the regression coefficients are the principal reason for elevated R^2 values, despite the opposing influences from increases in variability of the dependent variable (column 3 of Table 11).

A value of the multiple correlation coefficient close to one (or at least greater than .5) often leads to the explicit or implicit conclusion that the mathematical structure indicates causality. The reasoning behind the well-worn but often forgotten phrase "one can never prove anything with statistics (mathematical models)" is demonstrated by the following example.

Let x and y represent the members of six pairs of numbers:

x	10	8	5	5	2	1
y	4	5	8	13	20	29

A linear regression analysis of x and y yields a squared multiple correlation coefficient of $R^2 = .856$, indicating that the linear model based on x predicts the data rather well. However, the "goodness of fit" is deceptive since the y values were exactly generated by the expression $y=4+z^2$

where z = 0, 1, 2, 3, 4, and 5, which does not involve x. The fact that x is associated with z makes x a good predictor of y but not a direct "cause" of y. In multivariate as well as the simple univariate case, it must be kept in mind that large values of R^2 indicate only predictability. The observed R^2 values in many of the regression analyses described here exceed .50. Similar work by others (e.g. [1] and 2]) achieved R^2 values in the neighborhood of .80 and even a specific case of R^2 = .98 [2]. Nevertheless, a high degree of predictability does not guarantee that useful information can be extracted by analyzing the components of a mathematical model.

"CONTROL" VARIABLES

The primary function of the 22 "control" variables (items 1-22 in Appendix A) included in the analysis is to statistically isolate the influences of air pollution by means of a multiple regression analysis from known sources of variation that affect mortality rates. Median family income, for example, is associated with mortality ($r = -.541$) and also with air quality ($r = .530$ for NO2) which spuriously decreases any observed association between air quality and mortality ($r = -.131$). Without a strategy to disentangle the influences of variables such as education, income and age, the effects of pollution on health would not be assessable. However, the direct influence on mortality from these 22 variables is of some interest. The ecologic regression approach should at least reproduce the well known relationships between mortality rates and variables characterizing socio-economic status.

The principal measure of the influence of a specific independent variable in a regression analysis is the regression coefficient associated with that variable. A basic problem in the analysis of regression coefficients is that the magnitude of the coefficient depends on the units of measurement, which means that regression coefficients reflecting variables measured in different units are not directly comparable. For example, income in dollars is not commensurate with temperature measured in degrees, so that direct comparison of the regression coefficients associated with these two variables is useless. A commensurate

measurement that makes possible the direct comparison of the contributions of specific variables to the total variation in mortality is the path coefficient invented by S. Wright (see Ref. [14]). A path coefficient p_i is a standardized variable defined as

$$p_i = b_i \sigma_i / \sigma_y$$

where b_i represents the regression coefficient associated with the ith variable, σ_i the standard deviation of that independent variable, and σ_y the standard deviation of the dependent variable y. The numerator $b_i \sigma_i$ is the change in the dependent variable measured in standard deviations expected from a change of one standard deviation in the independent variable. The rate of change $b_i \sigma_i$ divided by σ_y is unitless and indicates the change expected in y per standard deviation of y. The path coefficient for income (% $\geq$ \$15,000) is -.214, and the path coefficient for July temperature is .123. This shows that mortality varies inversely with income and directly with temperature, and that the influence of income on mortality is slightly less than twice the influence of July temperature. Since this report involves a large number of regression analyses, only the essential pieces of the many analyses relating the 25 independent variables to total mortality will be presented.

Tables 12A (males) and 13A (females) contain the path coefficients for the 22 "control" variables for county and PUS level analyses. Among males (Table 12A) the major contributors to the regression equation for both county and PUS

analyses are variables relating to several dimensions of socio-economic status (SES). Four SES variables (percent black, percent over $15,000 income, percent four or more years college, and percent in owner-occupied housing) have associated path coefficients greater than .15 in absolute value and are statistically significant ($p < .02$) predictors of mortality rates. The divorce rate shows a smaller but also statistically significant contribution ($p = .073$ for county and $p = .109$ for PUS). Three of the four variables characterizing climate (January temperature, January precipitation, and July precipitation) have positive associations with male mortality, whereas July precipitation has a negative association. The same trends are observed in the county and PUS analyses; three of the eight coefficients are statistically significant. The variables reflecting the age distribution (percent under 5 years old and percent over 65 years old) have expectedly little influence since age-specific mortality rates were used. The measure of urbanization (percent urban) is statistically significant but shows opposite influences for county level and PUS level analyses. The other "control" variables have small individual influences on male mortality rates and are generally not statistically significant.

The identical analyses for female mortality (Table 13A) yield, more or less, the same associations as those observed in males. Income-related variables again give the strongest associations, although not as consistently as for males. The weather-related variables make significant contributions

to the variation in mortality (five of eight path coefficients are statistically significant). Once again, the county and PUS analyses show the opposite associations with the measure of urbanization (percent urban).

The values of the regression path coefficients observed in nationwide data are averages over regional variation. Tables 12B-12E (males) and 13B-13E (females) give the path coefficients from regression analyses for four separate regions. (West = federal regions 8, 9, and 10; Midwest = federal regions 5 and 7; South = federal regions 4 and 6; Northeast = federal regions 1, 2 and 3.)

The results observed for males nationally (Table 12A) are repeated regionally (Tables 12B-12E) for some variables and not for others. Income, July temperature, and housing ownership have fairly consistent associations for the four regions studied for both PUS and county data. Other variables do not have such consistent associations with mortality rates. For example, the percentage of foreign residents has a positive association in the western region (p_i = .372 for county and p_i = .066 for PUS) and a negative association in the southern area (p_i = -.124 for county and p_i = -.369 for PUS) for the male mortality rates.

The path coefficients resulting from the four regional analyses of the same 22 "control" variables and the female mortality rates (ages 45-54) are given in Tables 13B-13E. The associations noted for males remain in the female regression analyses with a few exceptions, but more random

variation is apparent. Other comparisons can be made among the 10 analyses (Tables 12A-12E and 13A-13E), but the primary focus is on the association between mortality and air quality with the variation from the 22 "control" variables statistically held constant.

AIR QUALITY REGRESSION COEFFICIENTS

If an ecologic regression analysis of air quality and mortality rates is meaningful, then the central issue is the interpretation of the regression coefficients associated with the air pollution measurements -- TSP, SO2, and NO2. The combined influences of these three variables on the overall variation in nationwide total mortality rates is small but statistically significant both in the county and PUS analyses ($p = .009$ for males and $p < .001$ for females at the county level; $p = .035$ for males and $p = .010$ for females at the PUS level). The squared multiple correlation coefficient increases from $R^2 = .312$ to $R^2 = .317$ (males) and from $R^2 = .105$ to $R^2 = .114$ (females) when the three air quality variables are added to the regression equation based on the 22 "control" variables for county data. Similarly, increases for the PUS analyses are from $R^2 = .705$ to $R^2 = .712$ (males) and from $R^2 = .406$ to $R^2 = .432$ (females). Although air quality has a statistically significant influence, its magnitude is small. To gauge the "size" of this effect, it is helpful to assess the influence of two other sets of variables. For example, income (% $\leq$ \$3,000 and % $\geq$ \$15,000) when added to the regression equation increases R^2 from .304 to .317 (males; PUS level data). When the five weather-related variables are added to the analysis, the R^2 values show an increase from .692 to .712 for males in county level data, and from .300 to .317 for males in PUS level data. The influences associated with the three air quality measurements show smaller but similar

changes in R^2 values.

It is important to note that statistical significance is achieved by variables that add only small amounts to the R^2 value. For the county level data, a variable that changes R^2 by more than .002 will be declared statistically significant at the 5% level; for the PUS data, a change of .003 will be declared statistically significant at the 5% level.

The direct comparison of the estimated regression coefficients (Tables 14A-14C) presents no special problems since all air quality measurements are in the same units (micrograms per cubic meter). The comparison of the three regression coefficients relating air pollution to male mortality rates (columns 1 and 2 of Tables 14A-14C) reveals a consistent but somewhat confused picture. The regression coefficients associated with the SO2 measurements show strong and statistically significant associations for both county and PUS analyses. However, three of four coefficients for TSP and NO2 are negative (one of these is statistically significant, with $p < .02$) which implies that lower mortality rates are accompanied by higher levels of TSP and NO2.

The same analysis for the female mortality rates (columns 3 and 4 of Tables 14A-14C) yields similar associations between the pollutant measurements and mortality, although the regression coefficients are reduced (in absolute value) in all six cases. Again, the SO2 coefficients are positive, and three out of the four TSP and NO2 coefficients are negative.

The same analyses were repeated employing a "dummy" variable to account (in a limited sense) for regional variation. For example, TSP in the western counties typically contains a high level of dust not found elsewhere. This type of regional variation is indeed expected and is incorporated in the regression equation by adding another independent variable (a somewhat simplistic solution to the issue but one which provides a direct comparison with other work [1]). This approach (rows 3 and 4 of Tables 14A-14C) yields a reduction in the magnitude of the regression coefficients for most categories, but leaves the associations observed in the unadjusted national analyses essentially unchanged for all 12 coefficients. That is, SO2 has a strong and positive association, and both TSP and NO2 predominately negative associations, for county and PUS analyses for both sexes.

When the national data are stratified into four regional analyses (the regions previously defined) at a loss of some statistical power, no consistent pattern between air quality and mortality emerges. The SO2 measurements show mostly positive associations with mortality rates with two exceptions (b_i = -4.24 for county data -- west and b_i = -14.36 for PUS data -- east). The coefficients associated with TSP and NO2 are not consistent between county and PUS analyses nor consistent between male and female analyses. Furthermore, as is the case with combined national data, many of the coefficients are negative and several of these are statistically significant.

One possible source of the inconsistencies among the regression coefficients computed for county and PUS data sets is related to the "ecologic fallacy" first discussed by Robinson [15]. Robinson demonstrated that the correlation coefficient calculated from a series of observations cannot be estimated by the correlation coefficient calculated from summary values that are aggregates or averages (ecologic variables) except under special circumstances. The problems of analyzing aggregated data as a unit of observation have been widely discussed, particularly in the social sciences (e.g. [16]). A similar "ecologic fallacy" applies to the calculation of regression coefficients. A summary regression coefficient representing the weighted average of the regression coefficients calculated from data within a series of groups is not, in general, equal to the regression coefficients derived from aggregated group values. Specifically, for the case of simple linear regression

$$b_{y/x}^{(w)} = \frac{R_{xx}}{R_{xy}} b_{y/x}^{(b)}$$

where $b_{y/x}^{(w)}$ is the regression coefficient derived from single record data, $b_{y/x}^{(b)}$ is the regression coefficient derived from aggregated data, and

$$R_{xx} = \Sigma\, n_i\ (\bar{x}_i - \bar{x})^2 \;/\; \Sigma\Sigma (x_{ij} - \bar{x})$$

$$R_{xy} = \Sigma\, n_i\ (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) \;/\; \Sigma\Sigma (x_{ij} - \bar{x})(y_{ij} - \bar{y})$$

The expressions represented by R_{xx} and R_{xy} are analysis of variance quantities -- namely, the ratios of the "between" to total sums of squares. The "within" regression

coefficient $b_{y/x}^{(w)}$ is equal to the "between" regression coefficient $b_{y/x}^{(b)}$ only when $R_{xx} = R_{xy}$, which is rarely encountered in applied situations.

To illustrate this type of bias the county is considered as a single data element (which it is not) and the PUS area as the aggregated unit of analysis. In this case, the regression based on county level data produces an estimate of the relationship between a variable x and mortality, namely the coefficients $b_{y/x}^{(w)}$, given in column 1 of Table 15. (Each row of Table 15 corresponds to a different variable x.) From the PUS data, the relationship between the same variable x and mortality can be represented by a set of aggregate coefficients $b_{y/x}^{(b)}$, given in column 2 of Table 15. The ratios, given in column 3 of Table 15, show the "ecologic" bias encountered from the aggregated values when a simple linear regression analysis is performed using PUS rather than county data. Absolute values of the observed ratios vary from >.1 to over 120. It should be reemphasized that county level data are not free from "ecologic" bias since the county values are themselves an aggregation of smaller units usually individuals. The magnitude of the county level bias cannot be estimated without the unit record data. The situation is more complex for the multivariable regression case, and similar biases exist.

ANALYSIS OF ACCIDENT AND STOMACH CANCER DATA

Two causes of death were analyzed using linear regression models, the 22 ecologic variables in appendix A, and the three air quality measurements previously employed, along with dependent variables stomach cancer and accident mortality rates for males and females aged 45-54. A demonstrated association exists between stomach cancer and air pollution [17]; accident mortality is not causally related to air quality. The county and PUS level data analyses should, therefore, yield the known associations or lack thereof, serving to reflect on the validity and sensitivity of the ecologic regression approach.

The county level analysis of stomach cancer mortality shows many of the expected associations between the 22 "control" variables and mortality in males. Income (% $\leq$ \$3,000 and % $\geq$ \$15,000), nationality (% foreign), age (% $\leq$ 5 years and % $\geq$ 65 years), and occupation (% professional) are significant ($p \leq .02$) contributors to the regression equation for stomach cancer. The identical analysis for stomach cancer rates among females yields less strong but similar associations between the 22 independent variables and mortality. Furthermore, as mentioned, highly specific dependent variables generally decrease the effectiveness of linear models based on ecologic data to summarize mortality rates. In the case of age-specific (45-54) stomach cancer, $R^2 = .027$ (males) and $R^2 = .009$ (females).

The PUS level analyses only generally reproduced the expected associations between the "control" variables and stomach cancer mortality. The squared multiple correlation coefficients are R^2 = .096 (males) and R^2 = .068 (females).

The county level analysis of accident mortality rates yields the generally expected association between the 22 "control" variables and mortality. Income, age, occupation and, additionally, temperature (average July and average January) are significant ($p < .02$) variables with regard to explaining the nationwide variation in accident mortality rates (R^2 = .172 for males and R^2 = .039 for females).

The PUS data reflect a somewhat different picture. Land area, urbanization (% urban), divorce rate, income (% $\leq$ \$3,000 and % $\geq$ \$15,000) and occupation (% manufacturing) are significant ($p < .02$) independent variables in the regression analysis for males. As before, these same associations are generally present but at a reduced level in the female analysis.

The fact that the "control" variables behave in a more or less expected manner in the ecologic regression analyses is reassuring. The role of air quality measurements in the explanation of stomach cancer and accident mortality is not as easily discerned. The coefficients given in Table 16 do not show any expected patterns. The coefficients from the stomach cancer regression analyses should be strong and positive for TSP measurements, since TSP has been found by

other investigators to be associated with increases in stomach cancer rates [17]. Of the four coefficients relating TSP to stomach cancer mortality, three are not statistically significant (two of these are negative) and one is positive and significant ($p = .025$). The other coefficients associated with SO2 and NO2 are not statistically significant and mostly negative (6 out of 8). The magnitudes of most of the regression coefficients relating air quality to accident mortality rates are no more than would be expected by chance variation (10 out of 12). However, two regression coefficients from the county analyses are statistically significant ($p = .002$ for NO2 - males and $p = .035$ for TSP - females) but negative.

DISCUSSION

Like ionizing radiation, high levels of air pollution are unquestionably toxic. The existence of effects at low doses is much more equivocal. Parallel to the debate surrounding ionizing radiation, it is argued that most air pollution levels are low and mechanisms exist which protect against any damaging effects. On the other hand, it is possible that no threshold level exists and any elevation of air pollution increases the risk of disease. Since large numbers of individuals are exposed daily to air pollution, evaluation of relevant data, methodologies and inferences bearing on the existence or non-existence of a dose-response relationship between low levels of air pollution and disease risk is critical.

This report is focused entirely on the evaluation of regression analyses of ecologic data to study the effects of ambient air pollution on mortality rates. The following discussion is divided into sections concerning the statistical issues underlying the use of linear models applied to aggregated data, the adequacy of ecologic data, and the interpretation of results (inferences) made from ecologic regression analyses. Data, analytic techniques, and conclusions are indeed inter-dependent but are presented separately for clarity.

STATISTICAL ISSUES

Regression analysis is rigorously derived from a set of mathematical assumptions. The application of regression

techniques is necessarily less precise. The assumption of a normally distributed variable with equal variance linearly related to a series of independent variables is never completely realized in any analysis. Some violations of these basic assumptions were detected in both the county and PUS data (e.g., mortality rates do not appear to be normally distributed). The analysis of the residual values shows moderate deviations from the expected values but no large, clear-cut trends. The fact that relatively large quantities of data are available for each analysis (>3,000 counties and >400 PUS areas) makes the analyses rather robust with respect to violations of some of the statistical assumptions. The questions of normally distributed data, equal variance, multicollinearity of coefficients, and adequacy of linear models could be investigated further but the present analyses indicate these purely statistical issues are not likely to be of fundamental importance.

However, interpretation of the regression coefficients is basic to the regression approach. In a classical regression analysis the regression coefficients estimate the expected response in a dependent variable for a unit change in an independent variable, while the other variables in the regression equation are held constant. The independent and dependent variables in an ecologic regression are summary values of aggregates of individuals. In this case problems arise in the interpretation of the regression coefficients when interest is on the individuals that make up the analyzed aggregate [18]. That is, no straightforward

interpretation exists of the ecologically derived regression coefficients with respect to the individuals (e.g., $b_{y/x}^{(w)}$ is not equal to $b_{y/x}^{(b)}$). The interpretation of ecologic regression analyses, in particular regression coefficients or correlation coefficients, as if they were derived from the classic regression assumptions, is often referred to as the "ecologic fallacy." We have illustrated biases due to the "ecologic fallacy" by comparing simple linear regression coefficients derived from county data with those derived from PUS level data. Although both regression analyses were estimating the same relationships, many coefficients differed rather strikingly. These comparisons do not help assess the biases incurred in the multiple regression situation but do show the potential for a disrupting and uncontrolled influence on inferences made from ecologically derived regression coefficients. Without a clear interpretation of the response of the dependent variable due to the isolated influences of specific independent variables (e.g., TSP, SO2 or NO2), the primary task of assessing the specific contributions to the variation in mortality rates from specific variables fails.

A technical point should be made with respect to the "multiple comparison" problem. The analyses presented in this report, as well as the similar work by others, involves large numbers of statistical tests, each with associated probabilities. These tests should be considered as exploratory tools and the stated significance levels as relative measures rather than accurate estimates of likelihood. When

a large series of ad-hoc, non-independent statistical tests is performed, the overall error rate increases and cannot be estimated without recourse to special multiple comparison procedures.

A statistical measure is declared significant if its value is unlikely to have occurred by chance variation. Analyses often yield statistically significant differences that have no consequential biologic influences, particularly when large amounts of data are involved. The county and PUS analyses of the influence of air pollution may fall into this category. Adding the three pollutants (TSP, SO2 and NO2) to the regression equations produced a statistically significant increase in R^2 values. The evaluation of these increases from a biologic perspective is more difficult. It is entirely possible that increases like those observed (e.g., 0.005 for county and 0.007 for PUS -- males) may be unimportant when assessed by other criteria. The question of "statistical" versus "biologic" significance is not unique to the study of air quality and disease, and should be kept in mind when evaluating the present results or those of other ecologic regression analyses.

DATA ISSUES

Mortality data are not ideal for statistical analysis since they are subject to a variety of biases [19] and problems [9]. A study of disease employing mortality data necessitates a choice of some measure of risk (usually rates). The use of total mortality rates minimizes the

problems of statistical instability due to small numbers of observations and avoids biases due to disease classification. This choice also maximizes the squared multiple correlation coefficient calculated in a regression analysis. General mortality patterns are more predictable from ecologic variables with linear models than are age- or cause-specific rates (e.g., among males aged 45-54, R^2 for total mortality is 0.859, while R^2 for stomach cancer is 0.096). This increase in predictability is paid for by a decrease in biologic specificity. If an association is established for total mortality, the question immediately arises as to which of the many widely varying causes of death are in fact involved. The possibility also exists that employing an overall measure of mortality obscures important interactions among the specific causes of death. The age- and sex-specific rates produce a more epidemiologically focused analysis but are not accurately summarized by a linear model (R^2 low). Total mortality leads to a high degree of predictability (R^2 high) but may yield more or less useless results since it is rare that a summary of a series of heterogeneous units is meaningful.

The 22 control variables present no technical problems. Furthermore, sampling errors and biases in the PUS data are non-existent or exist at extremely low levels. However, it should be emphasized that the important control variables may be missing. Measures of cigarette smoking and occupational exposures are not ecologic variables and are not tractable in the usual approach. The need to measure

smoking in studies of air pollution has been pointed out by many investigators (most recently [9]). The lack of smoking and occupational data in the ecologic approach is perhaps a fatal flaw.

The air quality data present concerns in several directions -- coverage, exposure, and timing. Only 57 percent of the US counties have adequate estimates of TSP, SO2 and NO2 (based on the 60 kilometer criterion in the interpolation algorithm). The estimates vary in accuracy (monitoring density) but measure at least to some degree the air quality surrounding most of the nation's population. Although the coverage may be adequate, the degree of exposure of county residents is not directly measured for at least two reasons. Air quality monitoring stations are often placed to record specific sources of pollution and the data may or may not be representative of the area. For example, a station might be placed near a coal burning utility company, so that air quality measurements from this station would not generally reflect the actual levels experienced by the county residents. Secondly, neither mortality statistics nor control variables incorporate into the analyses the important aspects of population stability. The fact that a person resides in a specific county does not necessarily imply that personal exposure levels are reflected by air pollution estimates for that county. An undetermined number of persons will be new residents, or work elsewhere, or for a host of reasons, spend little time in the county of residence that appears on the death certificates. To the degree that

this number is large, the county estimates will not accurately reflect exposure.

Whether the air quality measurements recorded in the EPA-SAROAD data base represent human exposure is one question. Another important question is when the air quality was measured. The present data involve mortality during 1968-1972 and air quality recorded during 1974-1976. Other studies are also forced to use rather recent air quality measurements since accurate nationwide data are available only for the last decade. For example, the work of Mendelsohn and Orcutt [1] used 1970 mortality and 1970 PUS data along with 1974 air quality measurements. Implicit in analyzing mortality data from a time prior to the air quality data is the assumption that relative air quality differences among geographic units are stable over time. This important and usually ignored assumption implies that overall pollution levels could change, but relative differences must remain stable to be useful in an analysis of antecedent mortality rates. In fact, it is not obvious when the air quality measurements should ideally be made. If pollution affects mortality largely by increasing cancer rates, then air quality measurements should be made 10-20 years prior to the mortality data since this time interval is thought to be the latency period for most cancers. Other causes of death have other latency periods and present a complicated picture for determining when air quality should be measured.

Another potentially severe problem with geographically based variables used in ecologic regression analysis is that these variables in many cases are averages of rather large and diverse units -- counties, PUS areas or Standard Metropolitan Statistical Areas (SMSA) [2]. Whether the ecologic variables are mortality rates, census summaries, or interpolated air quality measurements, they are averages of large numbers of observations. Analysis of this type of data does not address the basic concern that these averages may not be representative of any specific quantity. That is, they represent such a diverse set of measurements that they are relatively meaningless for understanding the nature of the relationships under investigation. For example, Los Angeles county has a population of over 7 million. Summaries such as median family income, percent black population, and percent owner-occupied homes may have little meaning since these values ignore the many extremely different subpopulations (some of the nation's richest and poorest populations live in Los Angeles county). Los Angeles county is an extreme case, but to a lesser extent the averaging of heterogeneous observations into a single set of numbers occurs in all county-, PUS-, or SMSA- based data.

INFERENCES

The ecologic regression analyses of both county level and PUS level data sets produce no strong or consistent evidence that a link exists between ambient air pollution and mortality. A few associations (positive regression coeffi-

cients) do occur. The association between mortality rates and SO2 levels is strong and positive for most analyses with a few exceptions. The interpretation of this result is complicated. Taken at face value, a positive coefficient reflects a direct influence in terms of a linear model, but the relationship between independent variables and dependent variables is undoubtedly more complex. Consider, for example, the observation that the coefficient associated with divorce rate is positive and in many cases significantly associated with total mortality. It is somewhat simplistic to conclude that the divorce rate directly influences mortality. That is, reducing the divorce rate alone is not likely to reduce mortality. Although a regression equation is easily used to estimate the change in the number of deaths that would result from a specific percentage decrease in divorce (elasticity), this number would not be very plausible, nor would any corresponding estimates made from these ecologic regression equations. Similarly it is indeed possible that the positive association between SO2 and mortality does not result from a direct causal relationship but rather from a complicated social/biological mechanism. Considering that TSP and NO2 levels have mostly negative coefficients for a majority of analyses, the most likely explanation of any observed relationship between air quality and mortality is that the coefficients are artificially produced by the analytic approach ("ecologic fallacy"). Protective effects (negative coefficients) from TSP and NO2 pollutants are biologically implausible and result either from indirect

associations with unmeasured variables (incomplete model bias) or are strictly the result of the fallacy of drawing inferences from ecologically derived regression coefficients. The difficulty of interpreting both county and PUS analyses is further demonstrated by the analysis of stomach cancer and accident mortality rates. The known association between stomach cancer [17] and TSP is not duplicated. In fact several (two out of four) of these regression coefficients are negative. Although most of the coefficients associated with accident mortality are not statistically significant (as expected), two are strongly negative ($p < .035$) again weakening the strength of any inferences made from this county/PUS approach to study air quality and mortality.

ACKNOWLEDGMENTS

The work described here is part of the ongoing PAREP (Populations at Risk to Environmental Pollution) project, a collaboration between the LBL Computer Science and Applied Mathematics (CSAM) Department and the University of California (at Berkeley) School of Public Health (SPH). Past and present funding was obtained from the Environmental Protection Agency (Bill Nelson), the Department of Energy (Walter Weyzen and John Viren), and the Electric Power Research Institute (Ron Wyzga).

Since the inception of the PAREP project in 1976, Warren Winkelstein (SPH) and Carl Quong (CSAM) have provided important guidance and support. Previous PAREP project managers were Craig Hollowell (LBL Energy and Environment Division) and Donald M. Austin (CSAM).

This project depends upon Seedis, the Socio-Economic Environmental Demographic Information System being developed by CSAM under an interagency agreement between the Department of Energy and the Department of Labor, Employment and Training Administration. Two Seedis modules - Chart, written by Bill Benson, and Carte, written by Peter Wood and Albert Yen - provided the tables and figures of this report.

Numerous people - primarily Barbara Levine, Simcha Knif, Fred Gey, Bob Healey, Edna Williams and Bill Hogan - helped prepare the data presented here.

The 1968-1972 mortality data, originally from the National Center for Health Statistics, were tabulated by

Herbert Sauer of the University of Missouri, and provided to LBL by Larry Milask, formerly with the UPGRADE project of the Council for Environmental Quality. The 1974-1976 air quality data were extracted from the SAROAD (Storage and Retrieval of Aerometric Data) data bank and provided to LBL by Carol Evans of the Environmental Protection Agency. Other related air quality data files were provided by Carmen Benkovitz of Brookhaven National Laboratory.

REFERENCES

1. R. Mendelsohn and G. Orcutt, An empirical analysis of air pollution dose-response curves, J. Envir. Econ. and Mang. 6, 85-106 (1979).
2. L. Lave and E. Seskin, "Air Pollution and Human Health," The Johns Hopkins University Press, Baltimore (1977).
3. P. F. Ricci and R. E. Wyzga, A statistical review of cross-sectional studies of ambient air pollution and mortality, presented at DOE Statistical symposium, Berkeley, CA (1980).
4. N. R. Draper and H. Smith, "Applied Regression analysis," Wiley & Sons, New York (1966).
5. U.S. Bureau of the Census, County and City Data Book, 1977 (A Statistical Abstract Supplement), U.S. Government Printing Office, Washington, D.C. 20402, 1978.
6. The Area Resource File, U.S. Department of Health, Education, and Welfare, Public Health Service, Health Resources Administration, Bureau of Health Manpower, Manpower Analysis Branch, DHEW Publication No. (HRA) 80-4, October 1979.
7. Public Use Sample Users Guide, U.S. government publication, U.S. Census Bureau (1975).
8. A. M. Lilienfeld, "Foundations of Epidemiology," Oxford University Press (1976).
9. W. W. Holland, A. E. Bennett, R. Cameron, et al., Health effects of particular pollution: reappraising the evidence, Am. J. Epid. 110; 527-659 (1979).
10. M. L. Levin, W. Haenszel, B. E. Carroll, et al., Cancer incidence in urban and rural areas of New York State, J. Nat. Can. Inst., 24: 1243-67 (1960).
11. W. Haenszel, D. B. Loveland, and M. G. Sirken, Lung cancer mortality as related to residence and smoking habits, J. Nat. Can. Inst., 28: 947-961 (1962).
12. C. L. Chiang, "Introduction to Stochastic Processes in Biostatistics," John Wiley & Sons, (1968).
13. S. Selvin, S. T. Sacks, D. W. Merrill and W. Winkelstein, The relationship between cancer incidence and two pollutants (total suspended particulate and carbon monoxide) for the San Francisco Bay Area, report LBL-10847 (June 1980).
14. C. C. Li, "Population Genetics," The University of Chicago Press (1968).
15. W. S. Robinson, Ecologic correlation and the behavior of individuals, Amer. Soc. Rev. 15: 351-357 (1950).
16. E. W. Borgutta and D. J. Jackson (editor), "Aggregated Data," Sage Publications, Beverly Hills, CA (1980).
17. W. Winkelstein and S. Kantor, Stomach cancer: positive association with suspended particulate air pollution,

Arch. Environ. Health, 18: 544-547 (1969).

18. L. I. Langbein and A. J. Lichtman, "Ecological Inference," Sage Publication, Beverly Hills, CA (1978).

19. S. Selvin, S. T. Sacks, and D. W. Merrill, Patterns of United States mortality for 10 selected causes of death, report LBL-10627, 1981, to be published.

20. J. L. McCarthy, D. W. Merrill, A. Marcus, W. H. Benson, F. C. Gey, and C. Quong, The Seedis Project: A Summary Overview, Lawrence Berkeley Laboratory Report LBID-379, April 1981 (revised version in preparation).

21. D. Merrill, 1974-1976 Air Quality: County versus PUS Area, Seedis file Z (CYAQSUM2), internal Seedis documentation, Lawrence Berkeley Laboratory, 1981.

TABLES

Fourth Count summary tabulation.

Table 10. Absolute differences, percent differences, and slopes summarizing the comparison between data derived from (a) the 1970 Census Public Use Sample (PUS) file and (b) aggregated county level data from the Fourth Count summary tabulation.

Table 11. Means and standard deviations of mortality rates (1968-1972), and value of the squared multiple correlation coefficient R^2, for several age and sex categories. The values of R^2 were obtained from regression analyses in which the 22 "control" variables and three pollution variables were the independent variables.

Table 12A. Males, U.S. total: path coefficients and R^2 from analysis of total mortality, ages 45-54, with 22 "control" variables.

Table 12B. Males, West: path coefficients and R^2 from analysis of total mortality, ages 45-54, with 22 "control" variables.

Table 12C. Males, Midwest: path coefficients and R^2 from analysis of total mortality, ages 45-54, with 22 "control" variables.

Table 12D. Males, South: path coefficients and R^2 from analysis of total mortality, ages 45-54, with 22 "control" variables.

Table 12E. Males, Northeast: path coefficients and R^2 from analysis of total mortality, ages 45-54, with 22 "control" variables.

Table 13A. Females, U.S. total: path coefficients and R^2 from analysis of total mortality, ages 45-54, with 22 "control" variables.

Table 13B. Females, West: path coefficients and R^2 from analysis of total mortality, ages 45-54, with 22 "control" variables.

Table 13C. Females, Midwest: path coefficients and R^2 from analysis of total mortality, ages 45-54, with 22 "control" variables.

Table 13D. Females, South: path coefficients and R^2 from analysis of total mortality, ages 45-54, with 22 "control" variables.

Table 1.
Total Mortality Rates:
Summary Statistics

	males		females	
	county	PUS	county	PUS
mortality rate per 100,000	935.20	895.80	435.60	447.40
standard deviation	252.00	146.10	123.40	56.00
skewness	1.21	.73	.37	.71
skewness 99.5% critical value	.01	.23	.01	.23
kurtosis	5.82	.80	2.95	2.29
kurtosis 99.5% critical value	.23	.50	.23	.50
maximum	12963.00	6306.00	5555.00	3225.00

Table 2.
U.S. Age Distribution in Whites, 1970

age	percent males	percent females	std dev (a) males	std dev (a) females
0	.02	.02	.00	.00
1-4	.07	.06	.02	.01
5-14	.20	.19	.04	.04
15-24	.18	.17	.05	.04
25-34	.12	.12	.03	.03
35-44	.12	.11	.02	.02
45-54	.12	.12	.02	.02
55-64	.09	.10	.03	.03
65-74	.06	.07	.03	.03
75-84	.03	.04	.01	.02
85+	.01	.01	.00	.01

(a) standard deviation indicates variation over 3082 U.S. counties

Table 3.
Age Correlations (*)
White Males, 1970

	age										
	0	1-4	5-14	15-24	25-34	35-44	45-54	55-64	65-74	75-84	85+
0	1.00										
1-4	.82	1.00									
5-14	.70	.85	1.00								
15-24	.41	.32	.22	1.00							
25-34	.70	.72	.57	.48	1.00						
35-44	.59	.69	.75	.25	.78	1.00					
45-54	.31	.41	.57	.06	.44	.73	1.00				
55-64	-.01	.09	.29	-.12	.03	.34	.74	1.00			
65-74	-.16	-.10	.10	-.17	-.18	.08	.48	.84	1.00		
75-84	-.18	-.13	.08	-.17	-.25	.00	.43	.75	.90	1.00	
85+	-.17	-.11	-.08	-.14	-.23	-.02	.35	.63	.74	.85	1.00

* Diameter of circle is proportional to correlation.
Shaded circles indicate negative correlations.

Table 4.
Expectation of Life at Birth:
Summary Statistics

	male	female
number of counties	3081.00	3081.00
expectation of life		
mean	66.78	74.25
std dev	2.07	1.51
standard error	.04	.03
minimum	54.50	65.53
maximum	78.98	85.00

Table 5A.
U.S. Counties with Lowest Expectation of Life
White Males, 1968-1972

	population	expectation of life (years)
MA Nantucket	3774	54.50
NB Loup	854	54.76
CO Hinsdale	202	57.18
WI Menominee	2607	57.71
GA Long	3746	58.81
CO Mineral	786	58.87
MT Treasure	1069	59.24
VA Nelson	11702	59.29
GA Atkinson	5879	59.61
CO Jackson	1811	59.74

Table 5B.
U.S. Counties with Lowest Expectation of Life
White Females, 1968-1972

	population	expectation of life (years)
MI Lake	5661	65.53
NV Storey	695	65.71
MA Nantucket	3774	65.89
MT Mc Cone	2875	66.96
GA Schley	3097	67.64
VA Charles City	6158	67.76
NV Lincoln	2557	67.93
WI Menominee	2607	67.93
VA Nelson	11702	68.04
MI Oscoda	4726	68.82

Table 6.
Correlation Between Expectation
of Life and Population Density

% < $3,000 (a)	counties	correlation coefficients females	correlation coefficients males
0 - 10%	728	.00	-.07
10% - 15%	818	-.14	-.13
15% - 20%	590	-.12	-.10
20% - 30%	657	-.18	-.10
30% - 100%	261	-.26	-.29
total	3081	-.06	.01

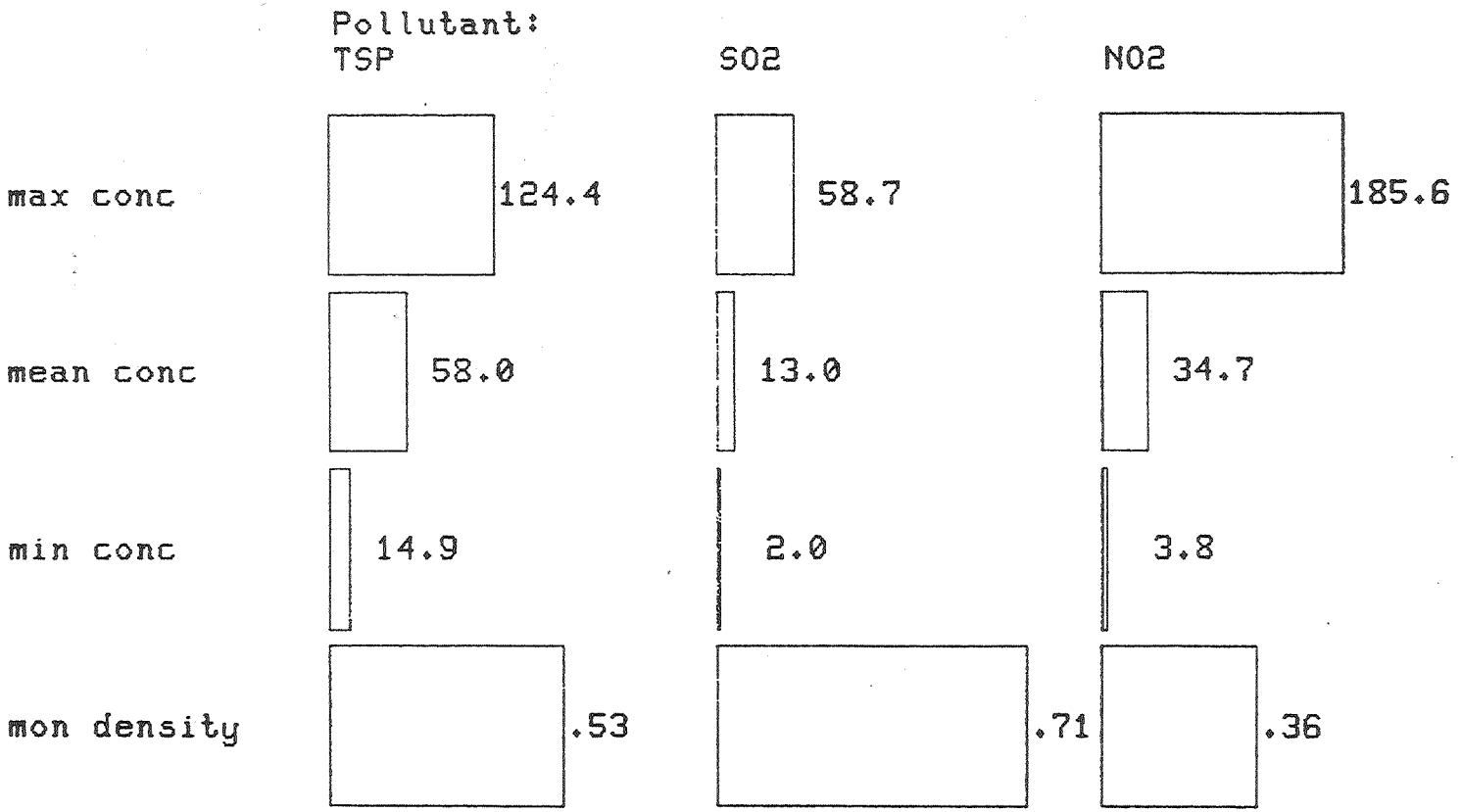
(a) percent of families with 1969 income under $3,000

Table 7.
County Level Air Quality, 1974-1976

	County Population: <10,000	10-500,000	>500,000	total
counties				
total	842	2144	73	3059
TSP	634	2049	73	2756
SO2	397	1736	73	2206
NO2	314	1467	73	1854
max conc				
TSP	127.9	115.1	174.2	174.2
SO2	55.1	51.7	67.6	67.6
NO2	169.8	75.4	201.8	201.8
mean conc				
TSP	52.9	55.5	61.9	55.1
SO2	7.5	10.1	17.8	9.8
NO2	20.1	27.1	48.3	26.7
min conc				
TSP	21.5	7.6	8.0	7.6
SO2	2.4	2.0	2.0	2.0
NO2	17.3	2.8	1.1	1.1
mon density				
TSP	.03	.16	1.23	.15
SO2	.04	.18	1.87	.18
NO2	.02	.08	1.02	.09

max/mean/min conc = micrograms per cu meter
mon density = full-time stations per 1000 sq km

Table 8.
PUS Level Air Quality, 1974-1976



max/mean/min conc = micrograms per cu meter
mon density = full-time stations per 1000 sq km

Table 9.
PUS File vs Fourth Count Tabulation:
Means and Standard Deviations

	PUS file		4th Count	
	mean	std dev	mean	std dev
1970 population	496620.1	517202.5	496269.6	515316.0
% black	9.8	11.7	9.7	11.6
% foreign	12.9	10.5	13.1	10.6
number of families	125114.3	130210.6	124656.3	129885.8
% <= 1 year old	8.5	.9	8.5	1.0
% <= 65 years old	10.1	2.9	10.1	3.0
% <= $3,000	11.4	6.0	10.7	5.8
% >= $15,000	17.9	8.6	18.5	8.4
population >= 25 yrs	268710.3	291686.2	267241.7	289994.2
% attended college	10.0	4.0	9.9	4.1
persons in labor force	195572.3	221281.5	195925.4	220806.6
persons employed	187028.0	210393.4	187240.7	209916.3
% employed, manufacturing	25.4	10.7	25.0	10.6
% employed, professional	22.2	4.7	22.0	4.7
occupied housing units	154988.3	173483.0	155096.2	173606.0
% incomplete plumbing	7.2	7.2	7.2	7.1
% > 1 person per room	8.3	3.6	8.4	3.6
% owner occupied	66.8	9.3	66.3	9.8

Table 10.
PUS File vs Fourth Count Tabulation:
Differences and Slopes

	absolute difference	percent difference	slope (a)
1970 population	-319.11	.00	1.00
% black	-.06	-8.20	1.01
% foreign	.32	3.30	.98
number of families	-454.30	-.40	1.00
% <= 1 year old	.04	.20	.86
% <= 65 years old	.04	-.10	.97
% <= $3,000	-.81	-9.50	1.04
% >= $15,000	.77	5.00	1.02
population >= 25 yrs	-1137.04	-.40	1.01
% attended college	-.07	-1.70	.98
persons in labor force	636.01	.50	1.00
persons employed	537.94	.40	1.00
% employed, manufacturing	-.08	-.60	1.01
% employed, professional	-.24	-1.40	.99
occupied housing units	135.25	.20	1.00
% incomplete plumbing	-.06	-3.10	1.02
% > 1 person per room	.02	-.50	1.00
% owner occupied	-.42	-2.10	.96

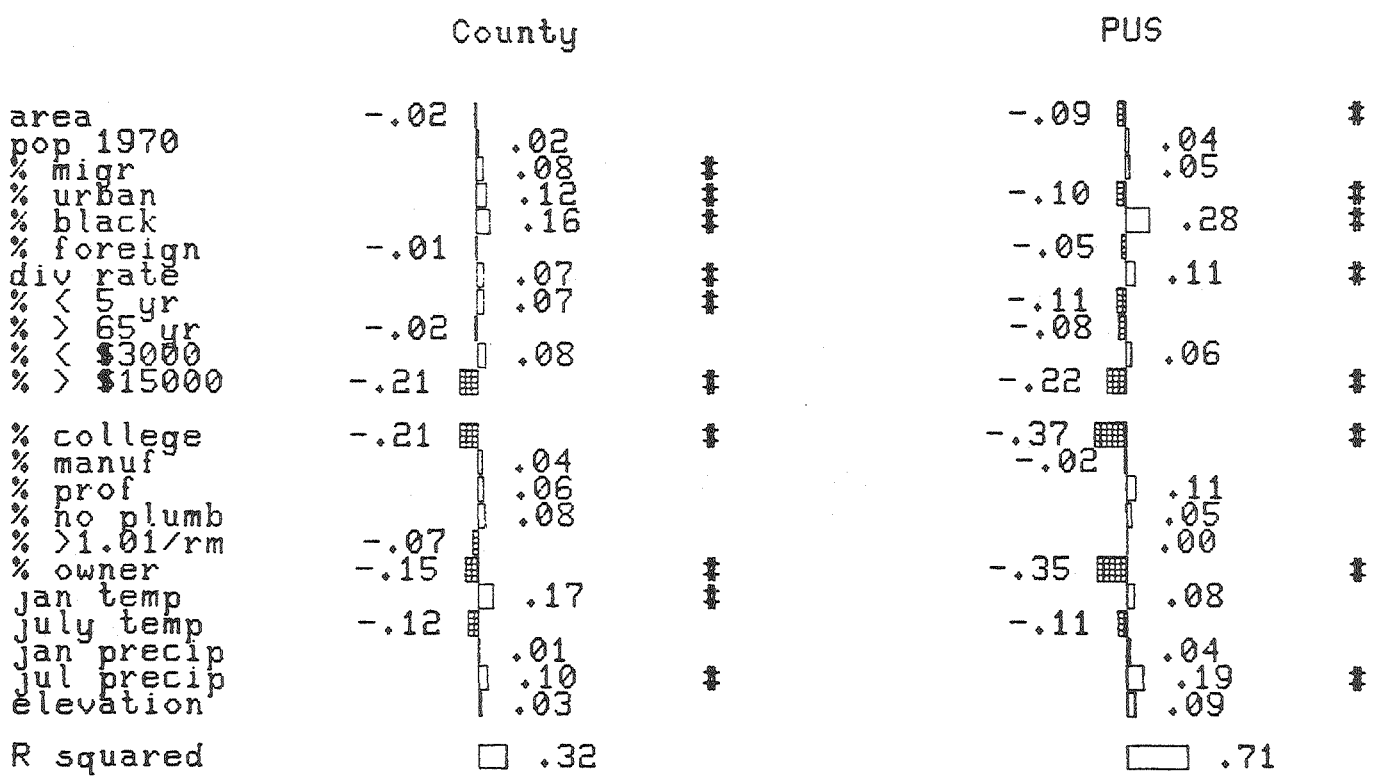
(a) slope = 1 implies no consistent source of bias between PUS and 4th Count

Table 11.
Mortality Rates and R Squared

	mortality rate (a) mean	mortality rate (a) std dev	R squared
males			
35-44	350	73	.69
45-54	896	146	.71
55-64	2197	255	.61
65+	7370	580	.63
females			
35-44	192	30	.48
45-54	447	56	.43
55-64	988	108	.42
65+	5002	461	.63
males			
35-54	623	110	.73
55+	4698	415	.67
females			
35-54	321	43	.53
55+	3164	324	.62
males +			
35-54	469	69	.69
55+	3849	335	.65
35+	2019	316	.86

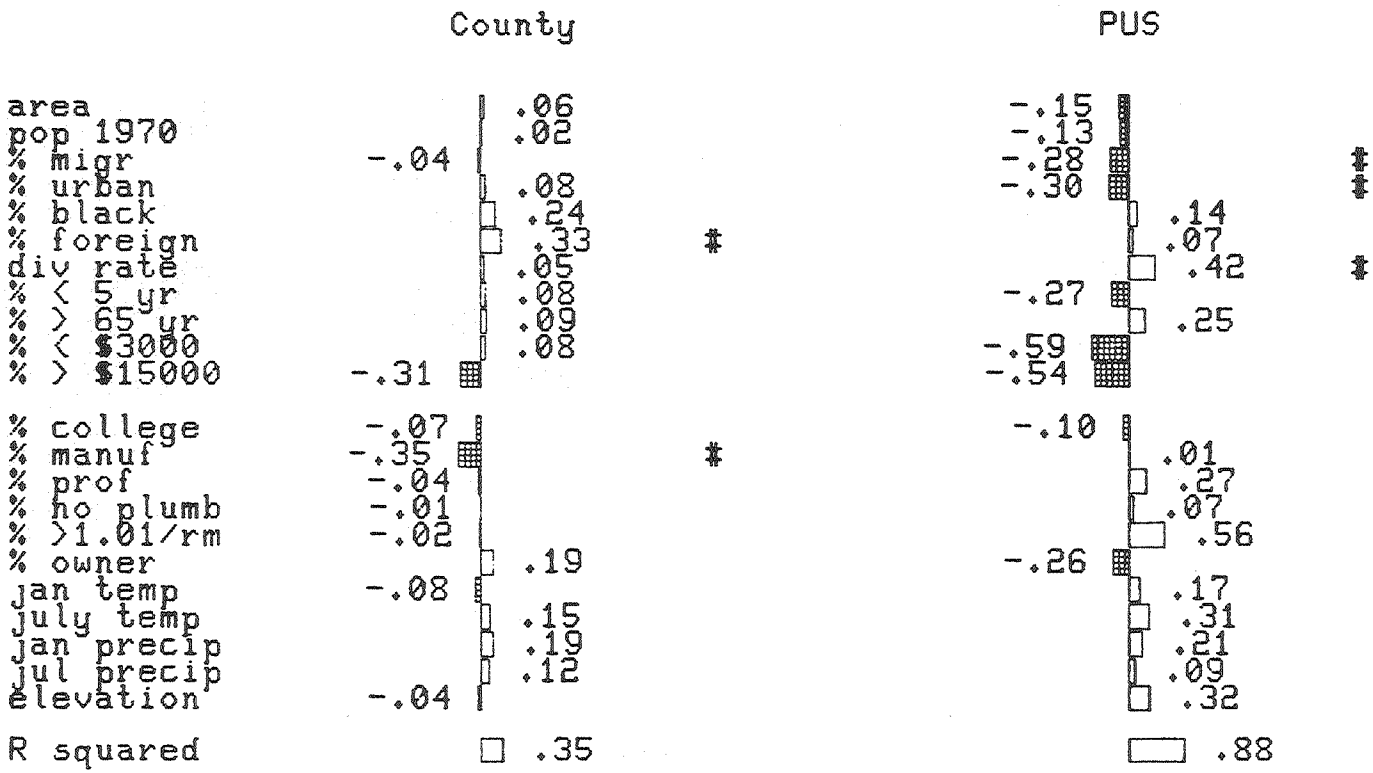
(a) avg annual total mortality per 100,000

Table 12A.
Males, U.S. total
Path Coefficients and R Squared



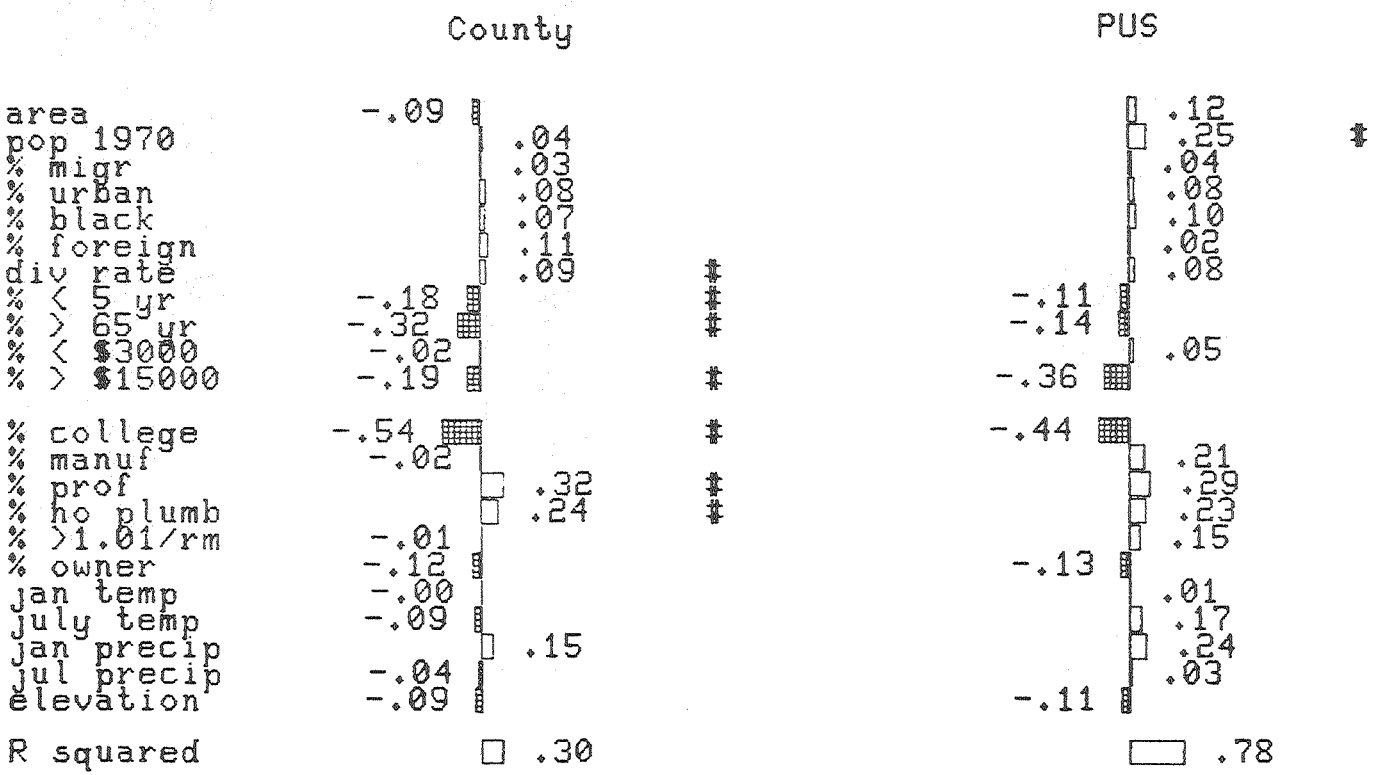
# indicates p-value <.02

Table 128.
Males, West
Path Coefficients and R Squared



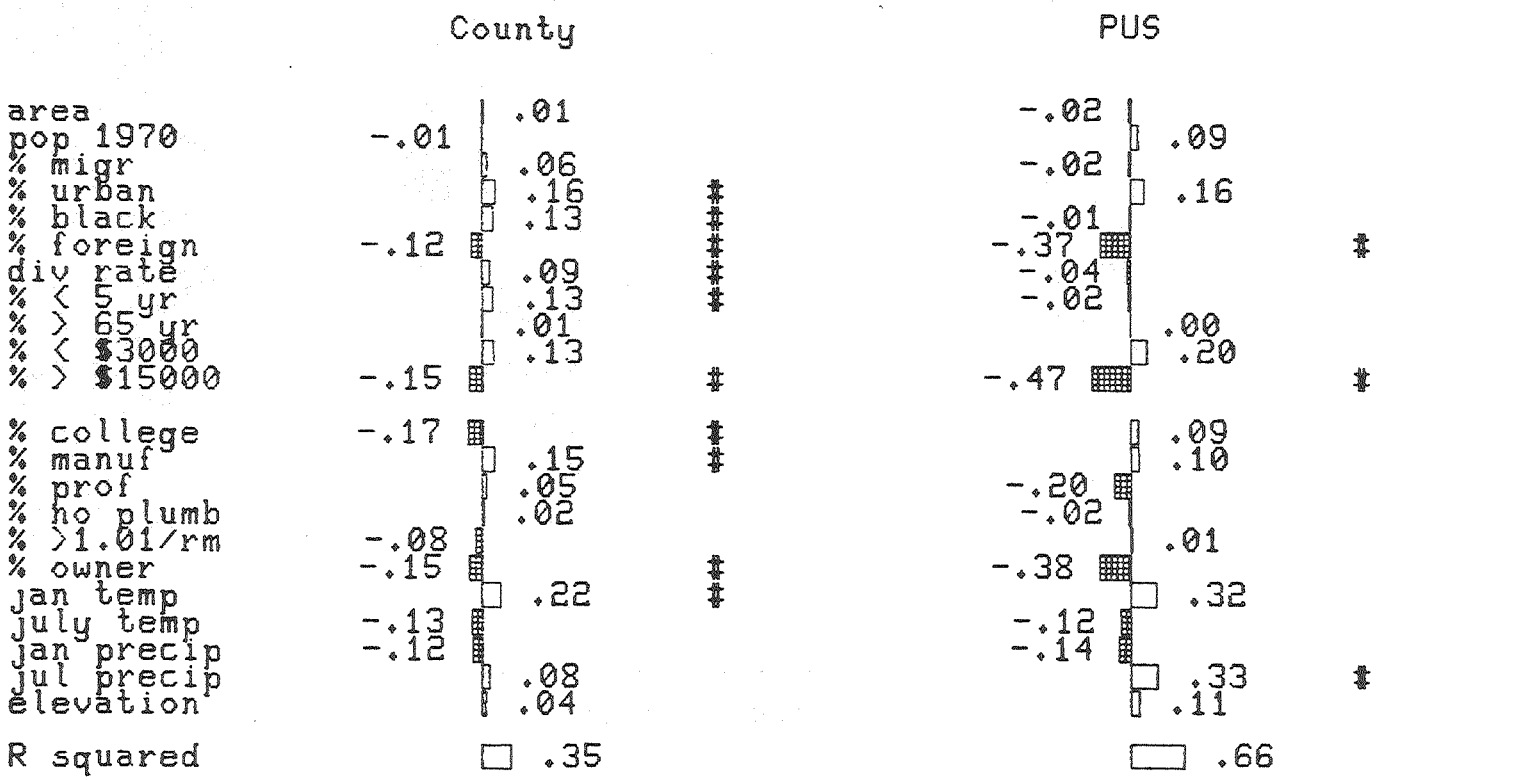
* indicates p-value <.02

Table 12C.
Males, Midwest
Path Coefficients and R Squared



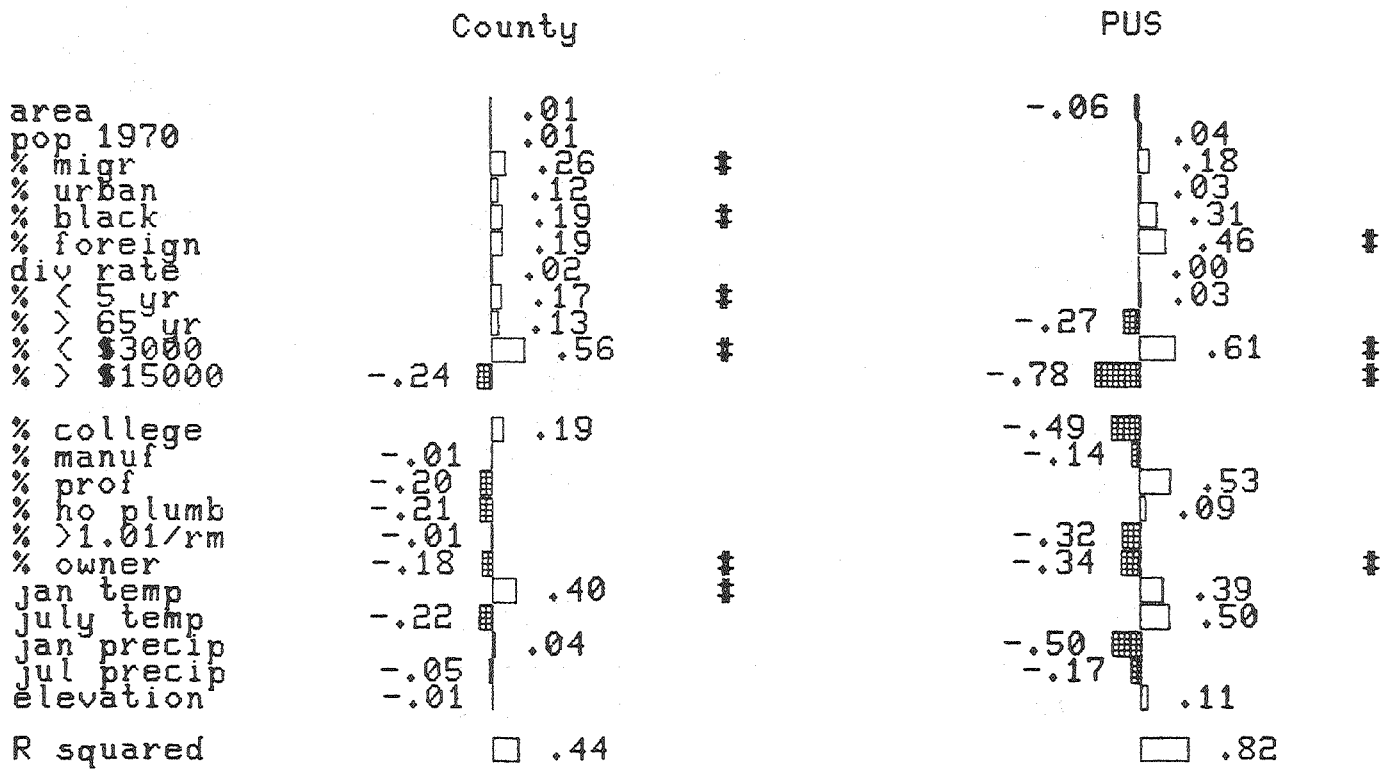
# indicates p-value <.02

Table 12D.
Males, South
Path Coefficients and R Squared



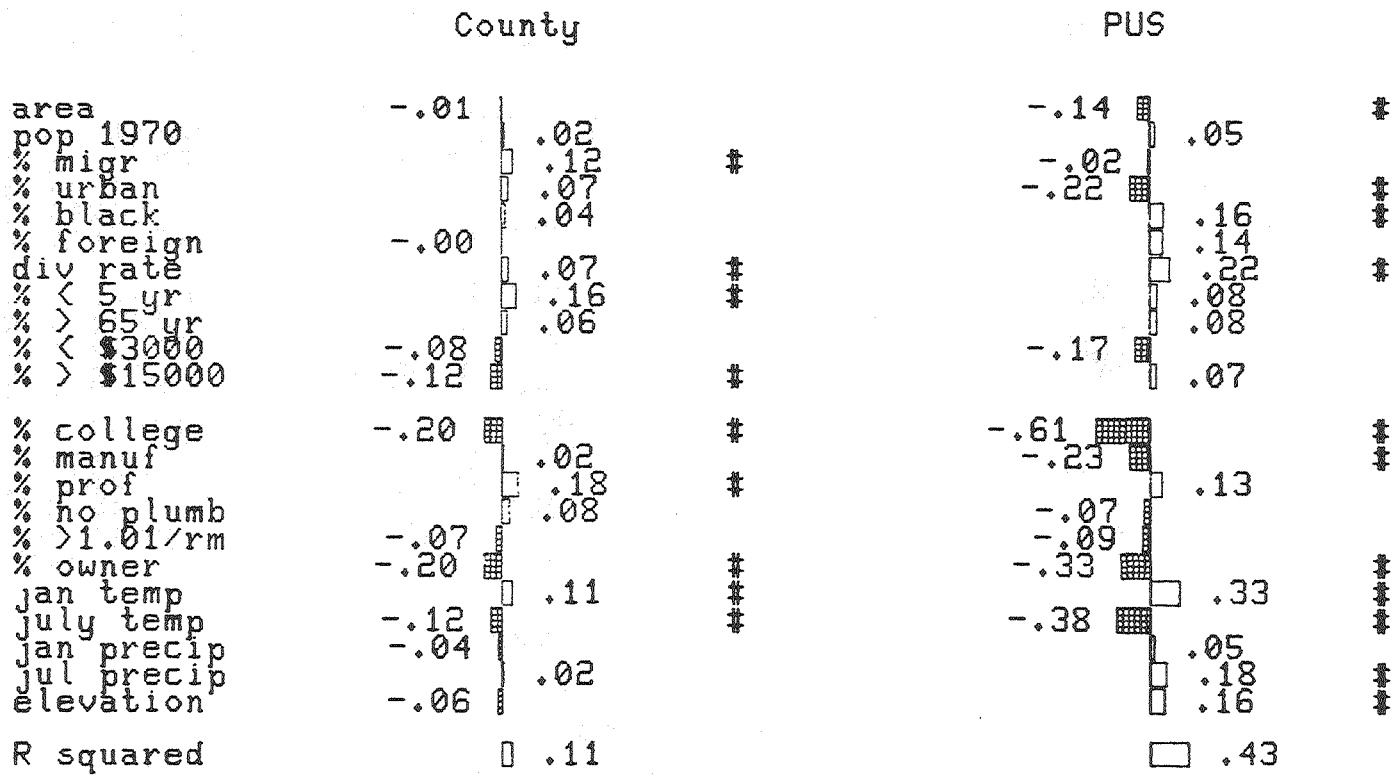
# indicates p-value <.02

Table 12E.
Males, Northeast
Path Coefficients and R Squared



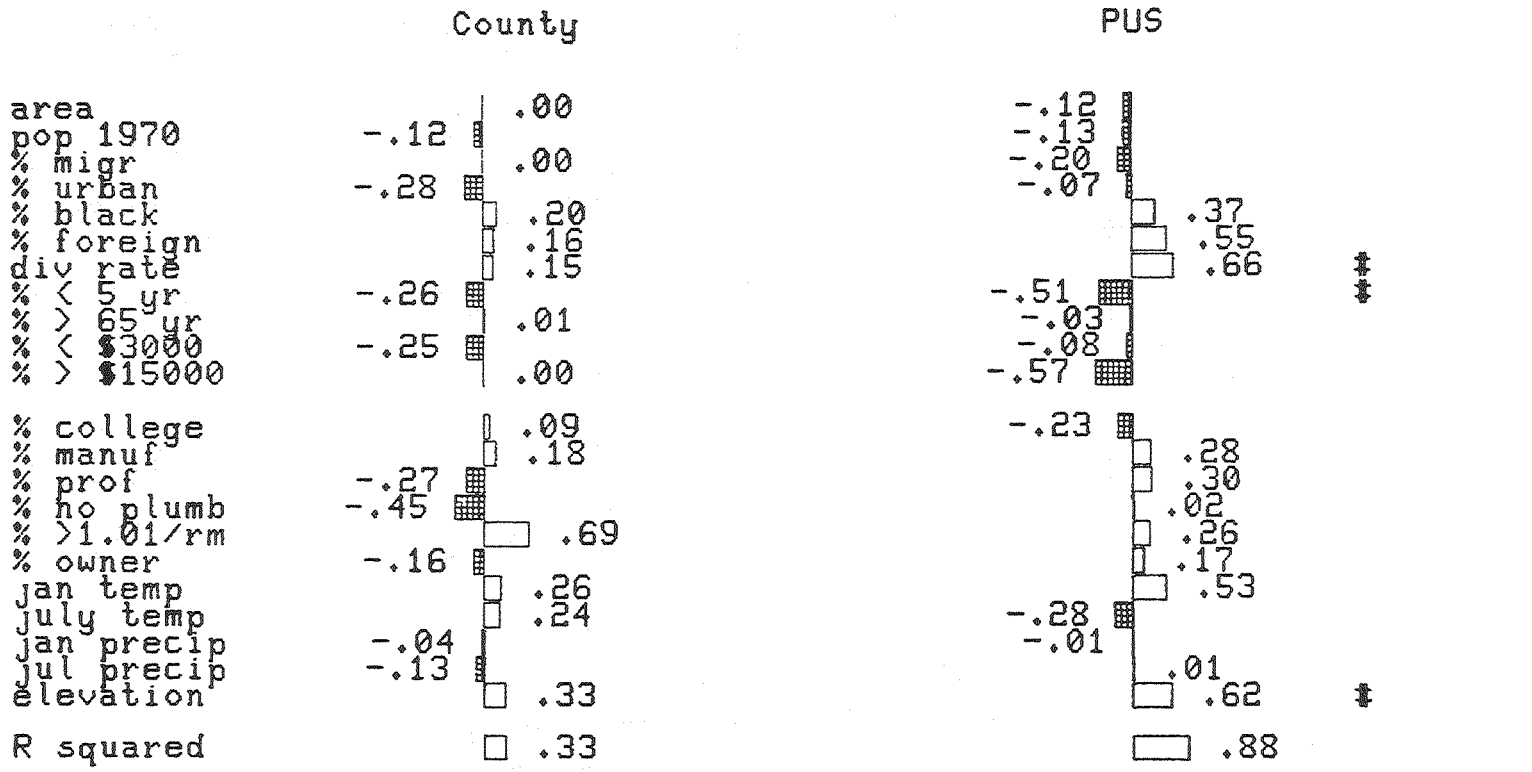
# indicates p-value <.02

Table 13A.
Females, U.S. total
Path Coefficients and R Squared



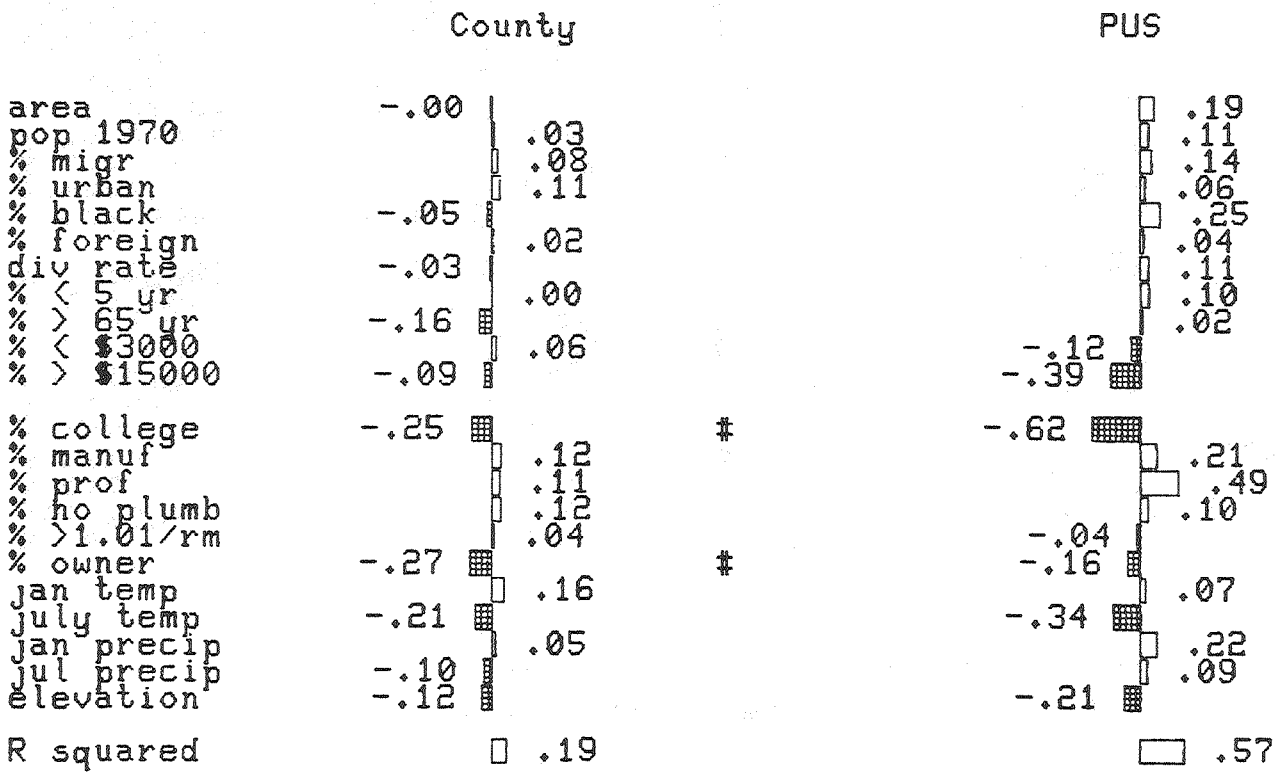
# indicates p-value <.02

Table 13B.
Females, West
Path Coefficients and R Squared



# indicates p-value <.02

Table 13C.
Females, Midwest
Path Coefficients and R Squared

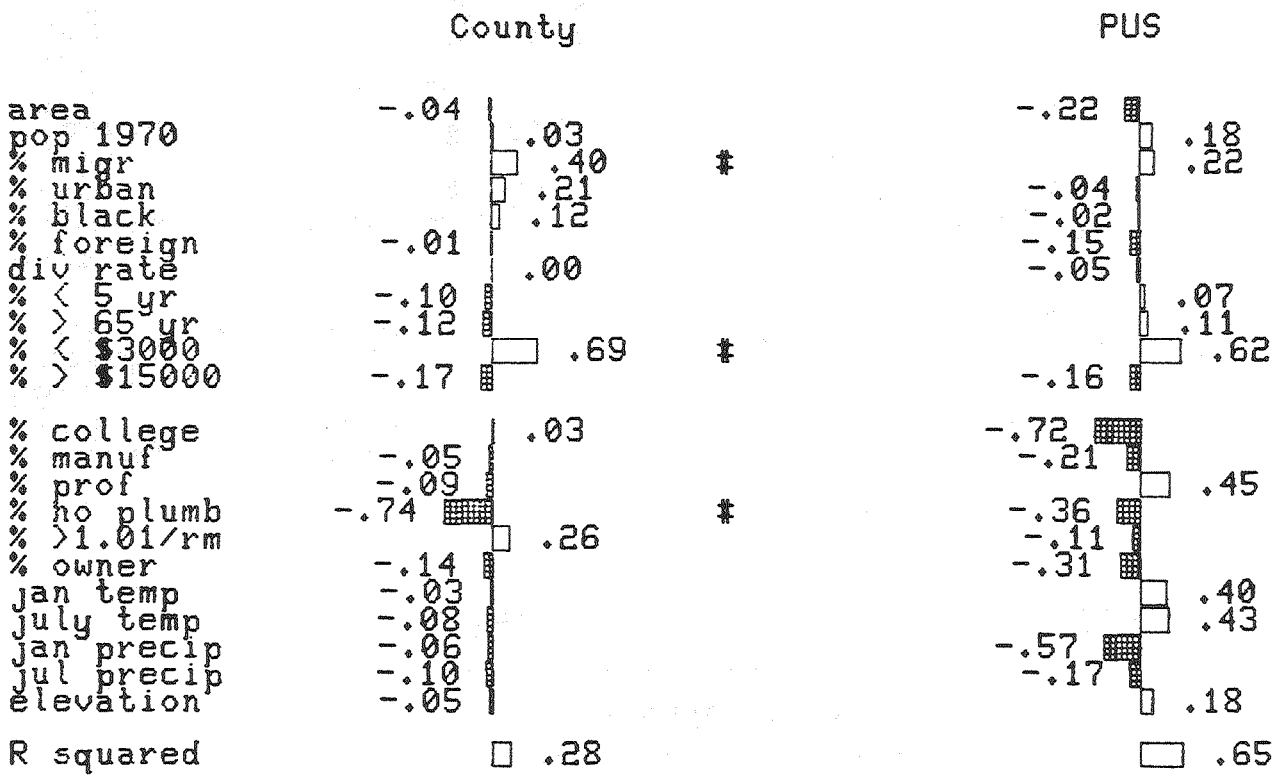


# indicates p-value <.02

Table 13D.
Females, South
Path Coefficients and R Squared

# indicates p-value <.02

Table 13E.
Females, Northeast
Path Coefficients and R Squared



# indicates p-value <.02

Table 14A.
Regression Coefficients
Total Suspended Particulate

	males		females	
US				
county	-3.7	#		.4
PUS	-2.6		-.7	
US(1)				
county	-2.4			.8
PUS	-1.1		-.5	
west				
county	-7.2		-9.2	#
PUS	2.5		-.2	
south				
county	2.3			.7
PUS	-6.0		-.6	
east				
county	8.5			1.9
PUS	8.0	#	-3.6	
midwest				
county	-1.4			2.6
PUS	2.5			2.4

# indicates p-value <.02

Table 14B.
Regression Coefficients
Sulfur Dioxide

	males		females	
US				
county	11.3	#	7.4	#
PUS	7.5	#	6.5	#
US(1)				
county	7.1		4.1	#
PUS	4.3		3.9	#
west				
county	-4.2		2.5	
PUS	18.7		5.8	
south				
county	24.8	#	2.0	
PUS	28.2		7.2	
east				
county	2.1		1.4	
PUS	-14.4		5.0	
midwest				
county	1.1		-.3	
PUS	-.7		-1.6	

# indicates p-value <.02

Table 14C.
Regression Coefficients
Nitrogen Dioxide

	males		females	
US				
county	-2.3		-2.0	*
PUS	1.4		-.3	
US(1)				
county	-2.6		-2.1	*
PUS	1.2		.0	
west				
county	10.1		3.6	
PUS	-5.6		-3.4	
south				
county	-1.6		-1.1	
PUS	1.6		1.0	
east				
county	-6.8		-3.1	
PUS	6.5		1.7	
midwest				
county	-6.8	*	-2.6	
PUS	2.1		2.8	

* indicates p-value <.02

Table 15.
Bias Incurred Through Aggregation

	county	PUS	bias ratio (a)
area	-.0	.0	4.8
pop 1970	.0	.0	1.6
% migr	9.1	4.5	2.0
% urban	4.5	-2.3	2.0
% black	12.0	15.7	.8
% foreign	-1.1	-3.0	.4
div rate	42.3	39.7	1.1
% < 5 yr	65.7	-81.3	.8
% > 65 yr	-7.3	-18.5	.4
% < $3000	11.4	7.0	1.6
% > $15000	-33.7	-17.0	2.0
% college	-57.5	-61.3	.9
% manuf	3.8	-1.3	2.9
% prof	12.9	15.1	.9
% ho plumb	7.7	5.0	1.6
% >1.01/rm	-16.6	.1	-120.8
% owner	-21.6	-28.2	.8
jan temp	-3.1	-4.7	.7
jul temp	1.8	3.6	.5
jan precip	.7	6.4	.1
jul precip	.1	1.4	.0
elevation	.0	.1	.5
TSP	-3.7	-2.6	1.5
SO2	11.3	7.5	1.5
NO2	-2.3	1.4	-1.7

(a) equal to 1.0 if no bias exists.

Table 16A.
Regression Coefficients
Counties

	stomach cancer		accidents	
TSP				
males	-.05		-.76	
females	.00		-.52	*
SO2				
males	.20		.00	
females	-.03		-.03	
NO2				
males	-.08		-1.31	*
females	-.03		-.25	

* indicates p-value <.02

Table 16B.
Regression Coefficients
PUS Areas

	stomach cancer		accidents	
TSP				
males	.19	#	.15	
females	-.01		.19	
SO2				
males	-.02		-.67	
females	.02		-.66	
NO2				
males	-.04		.01	
females	-.01		.08	

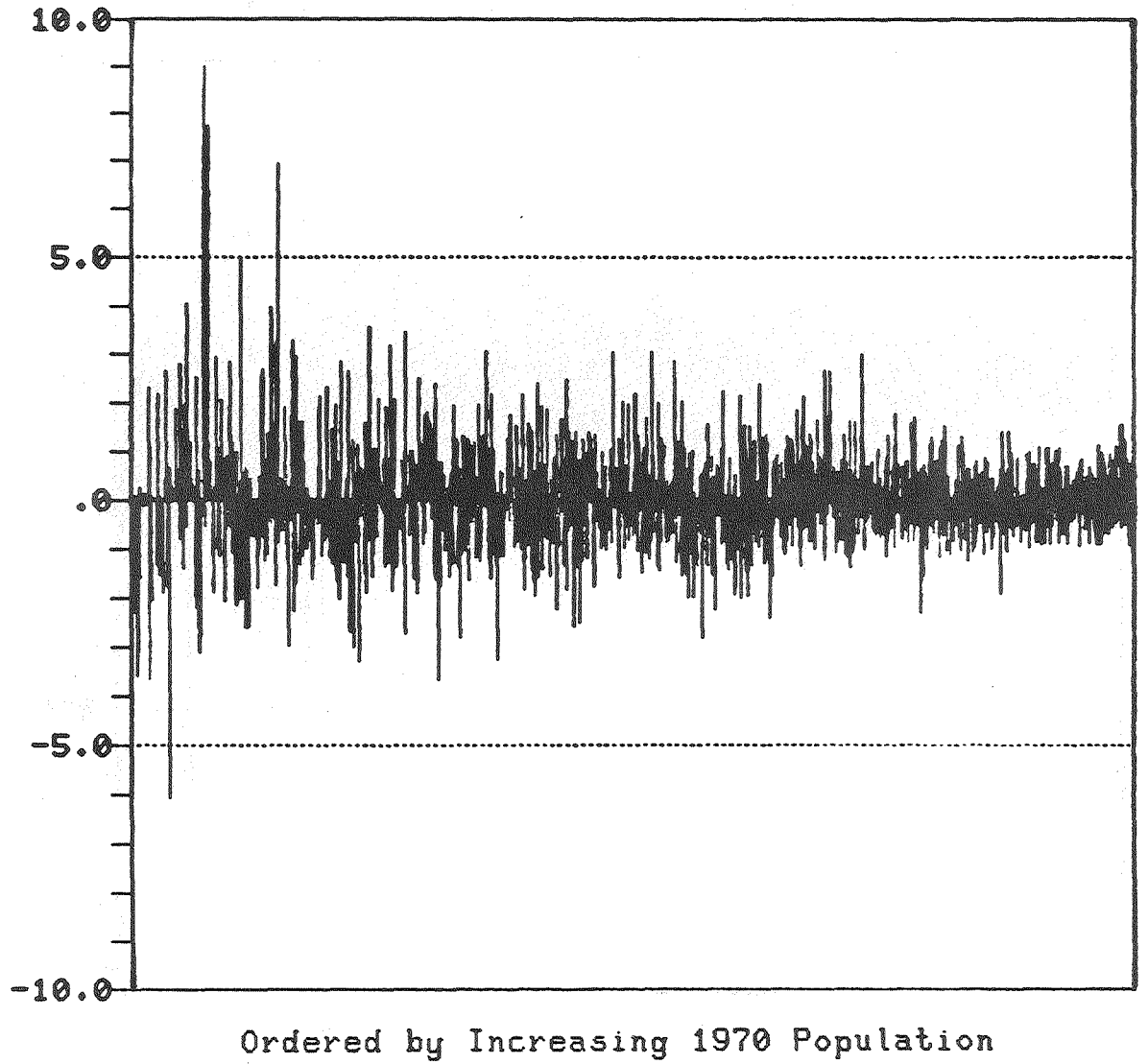
# indicates p-value <.02

FIGURES

Figure 1.

Residual for
Males, Age 45-54

Counties



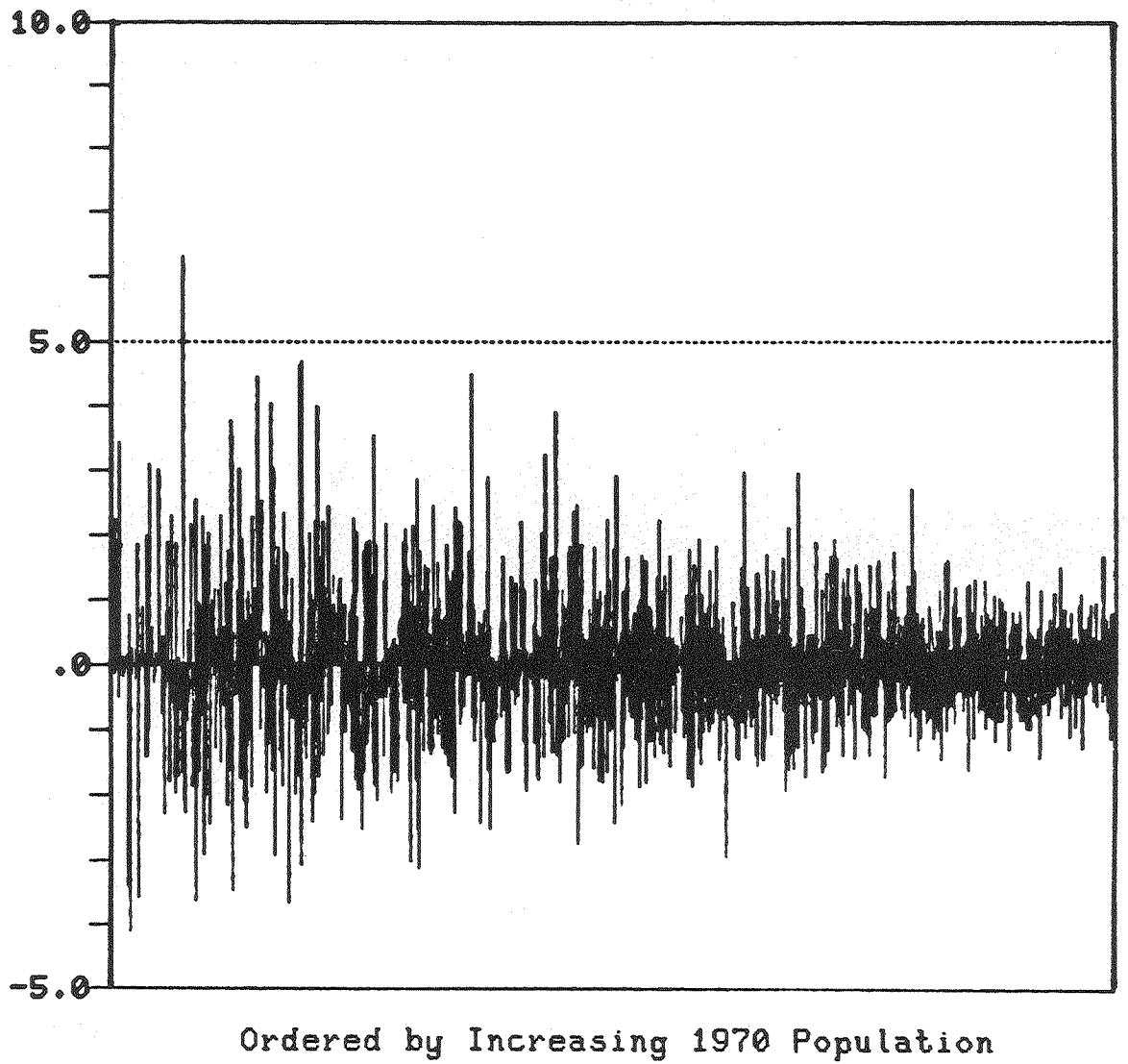


Figure 2.

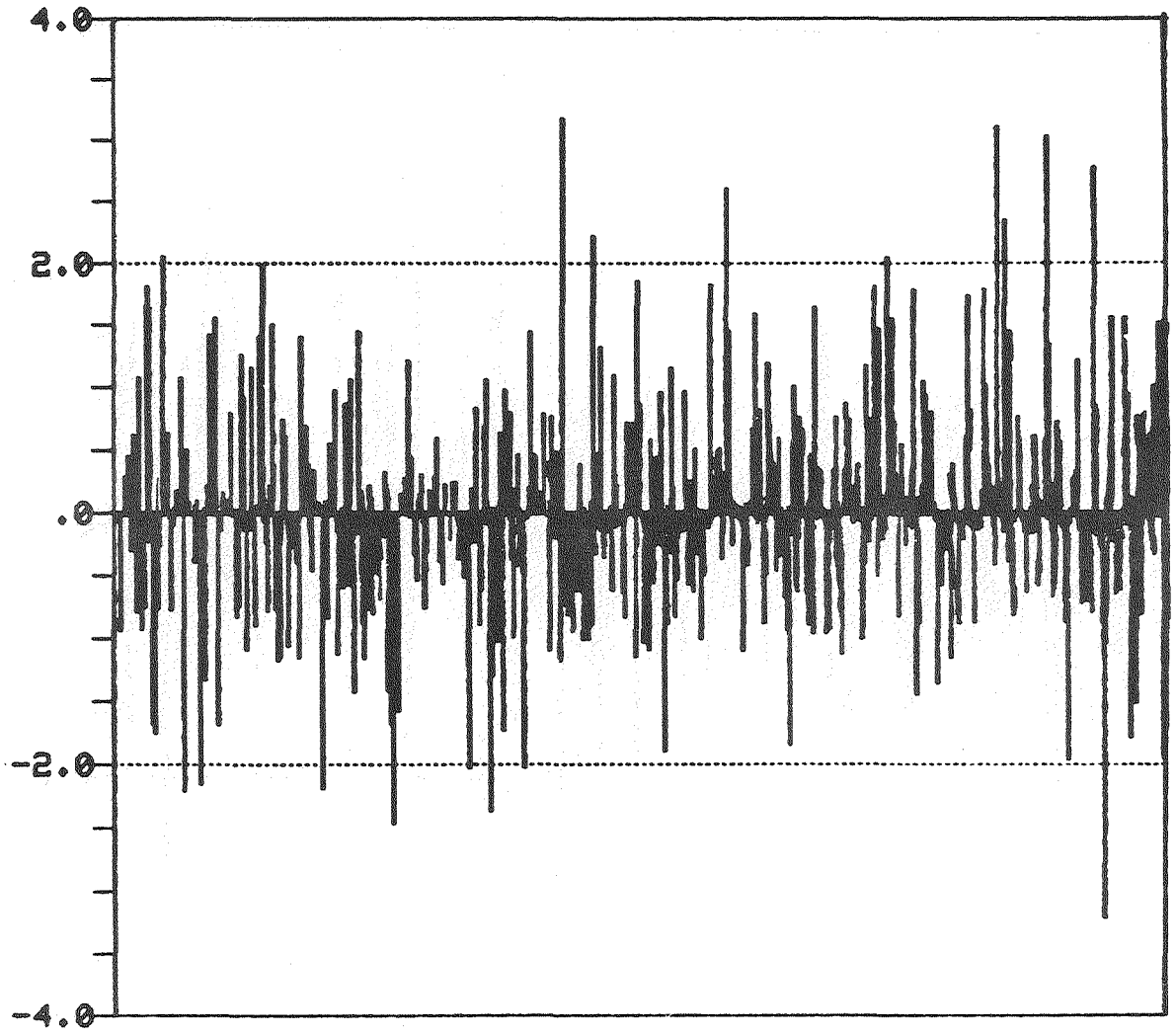
Residual for Females, Age 45-54

Counties

Figure 3.

Residual for
Males, Age 45-54

PUS Areas



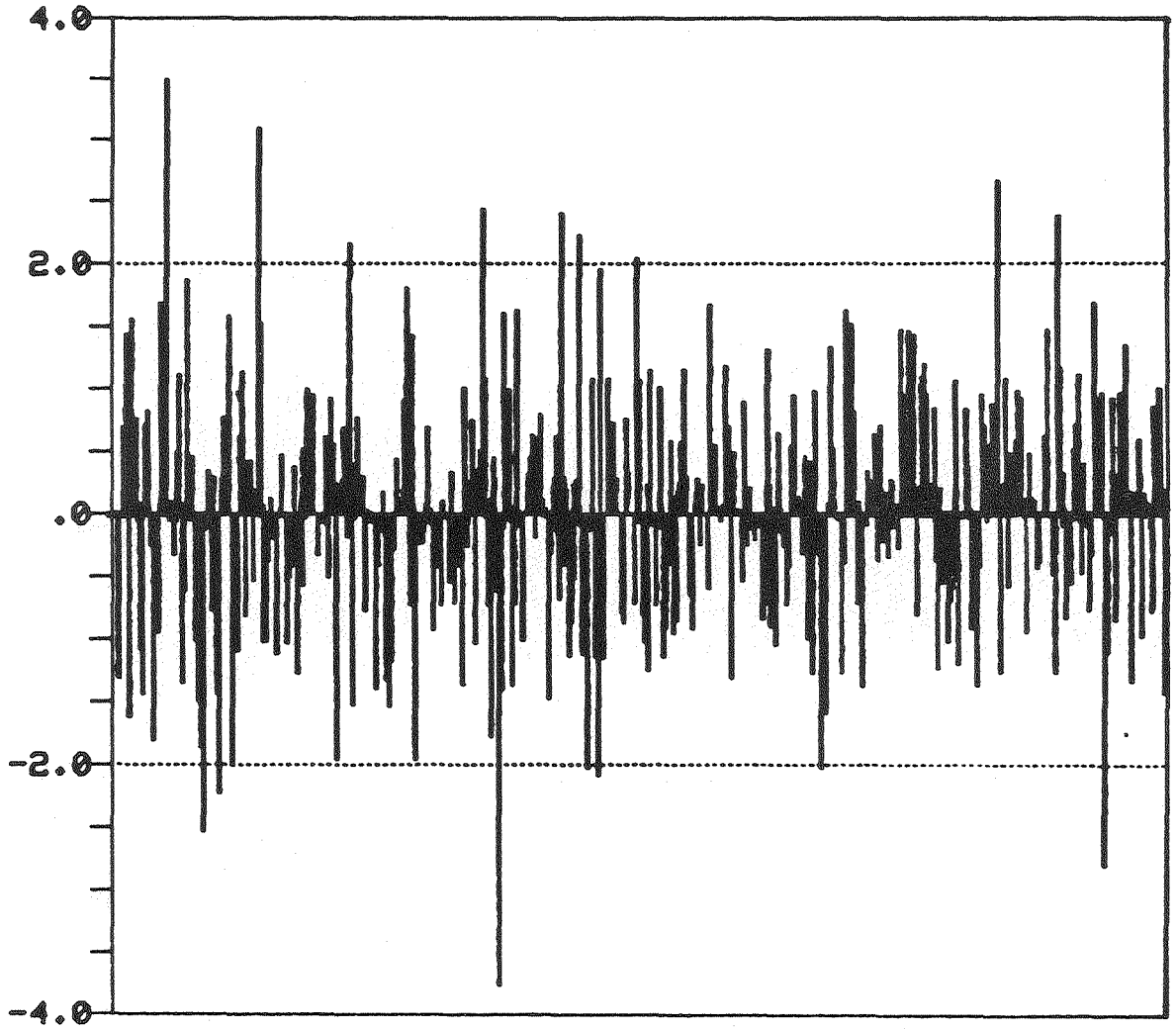


Figure 4.

Residual for Females, Age 45-54

PUS Areas

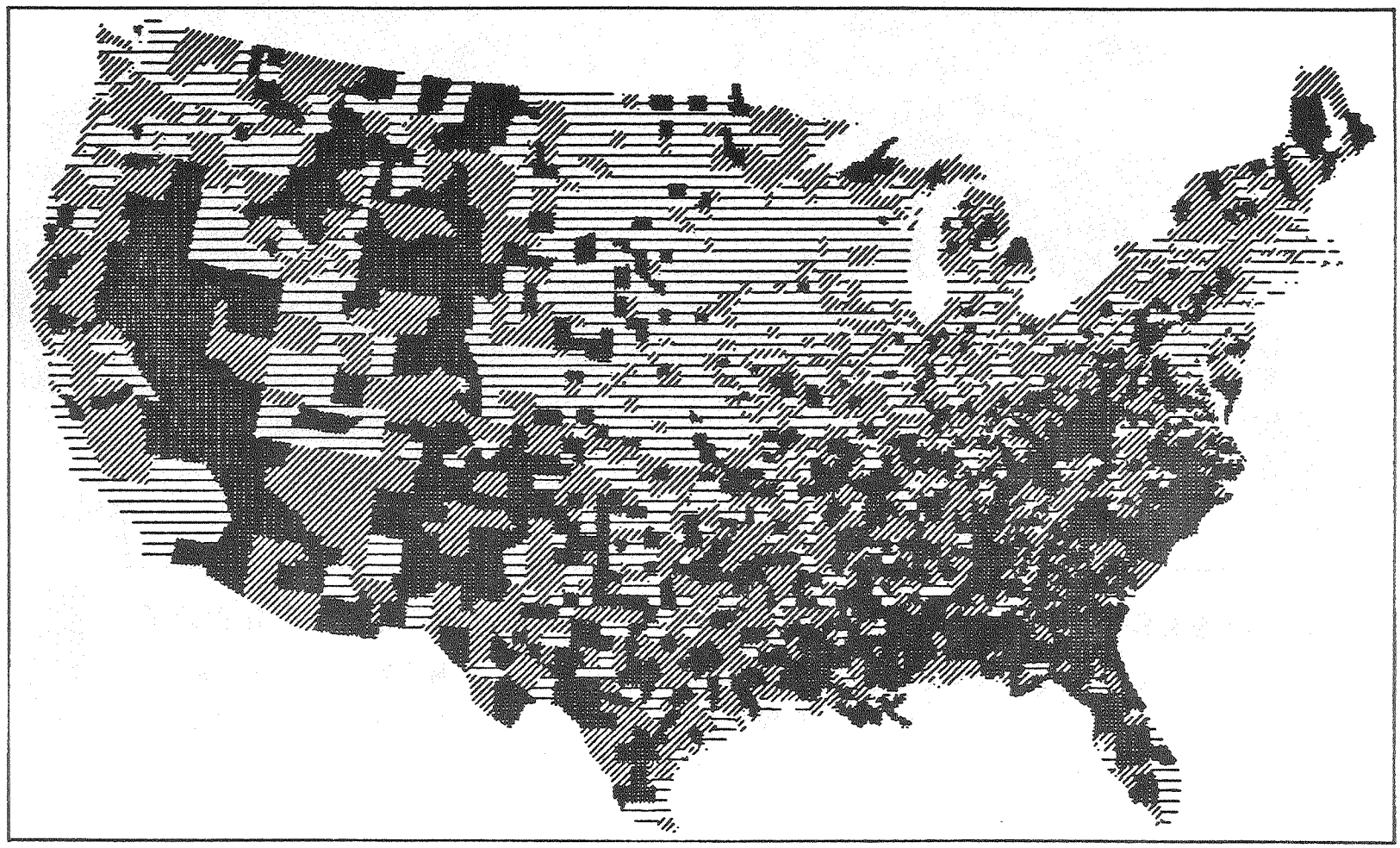


Figure 5.
Expectation of Life
at Age 0 (years)
White Males, 1968–1972

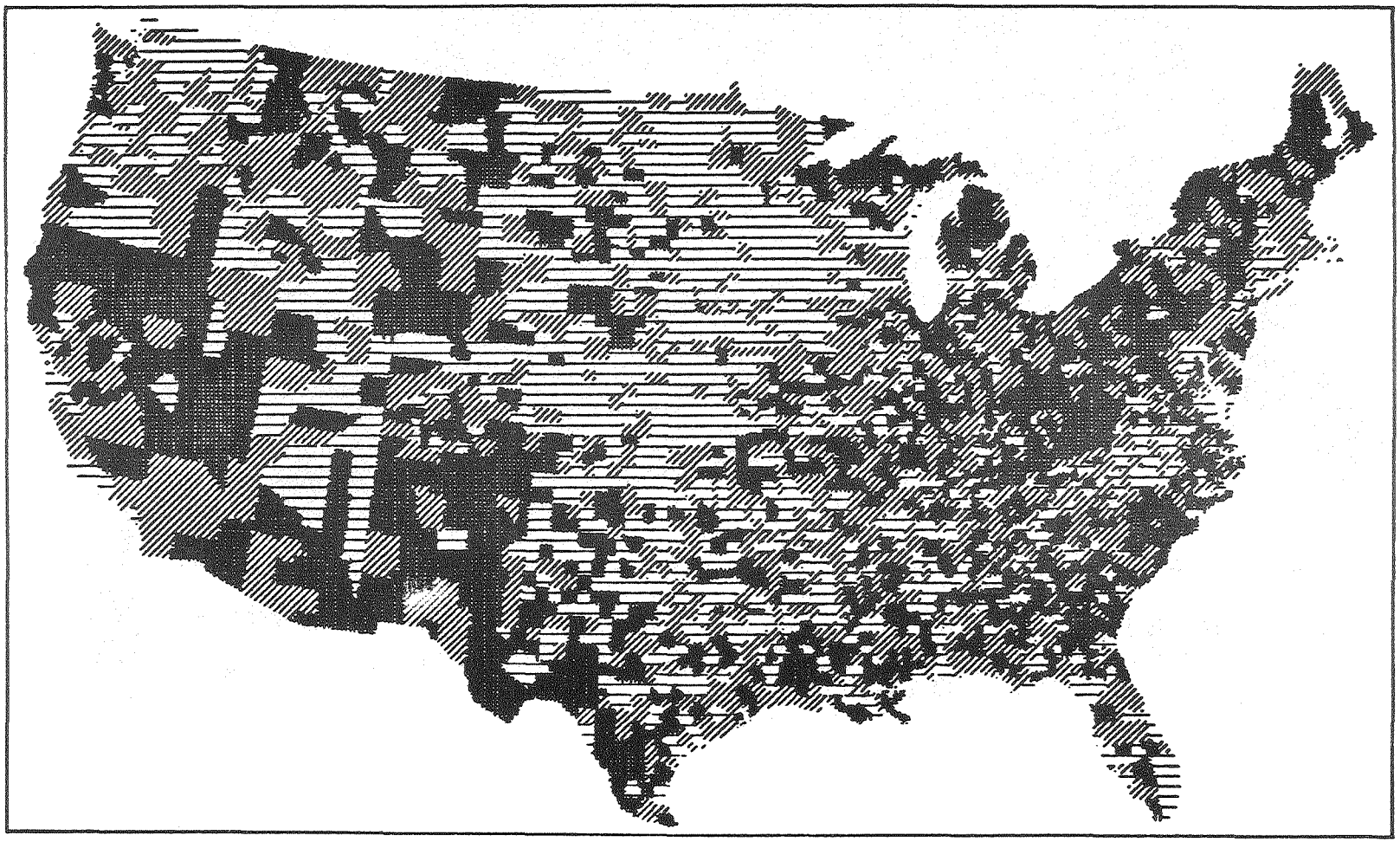


Figure 6.
Expectation of Life
at Age 0 (years)
White Females, 1968-1972

Figure 7.
1974-1976: Stations Measuring
Total Suspended Particulate
(24-hour sampling interval)

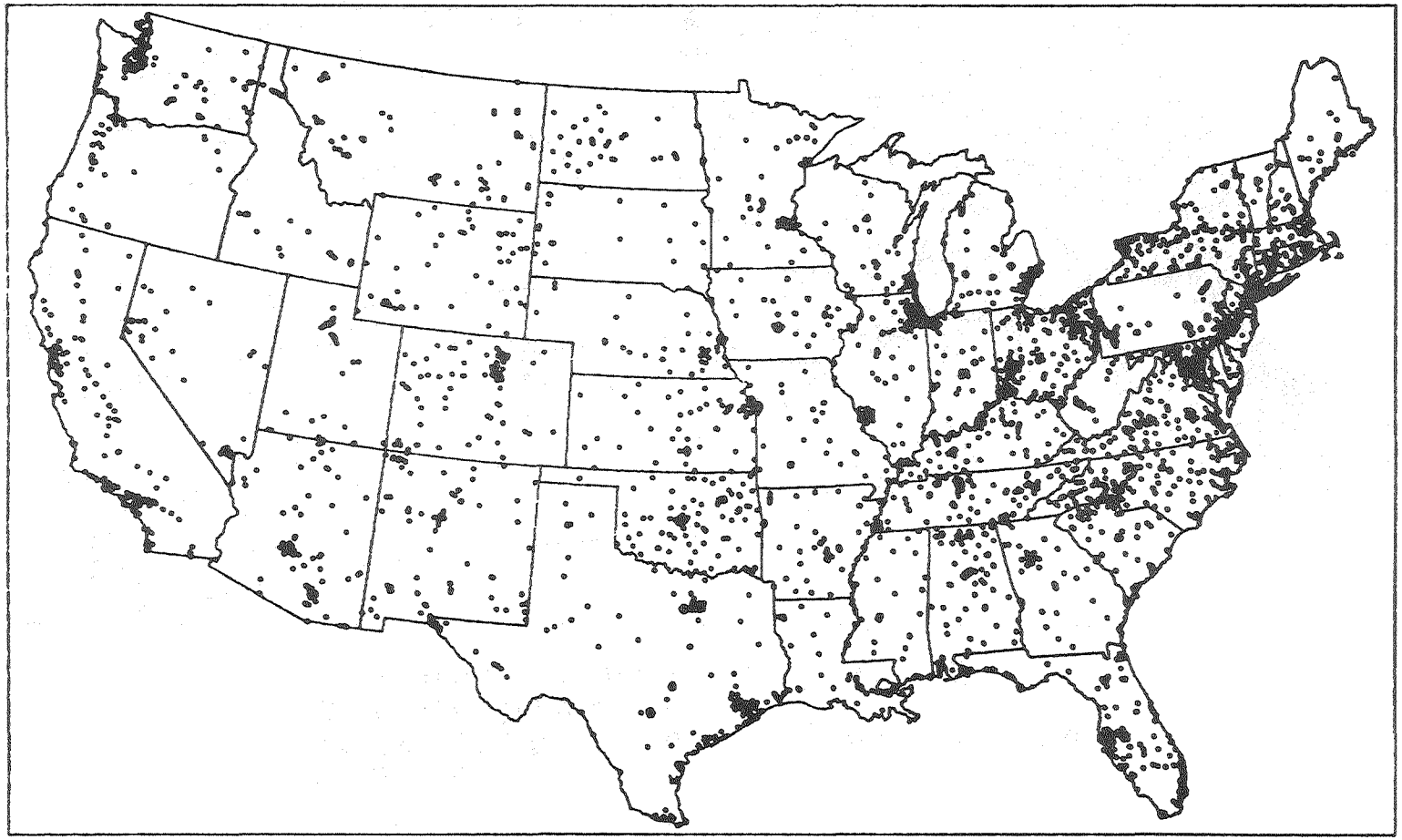


Figure 8.
1974–1976: Stations Measuring
Sulfur Dioxide
(24-hour sampling interval)

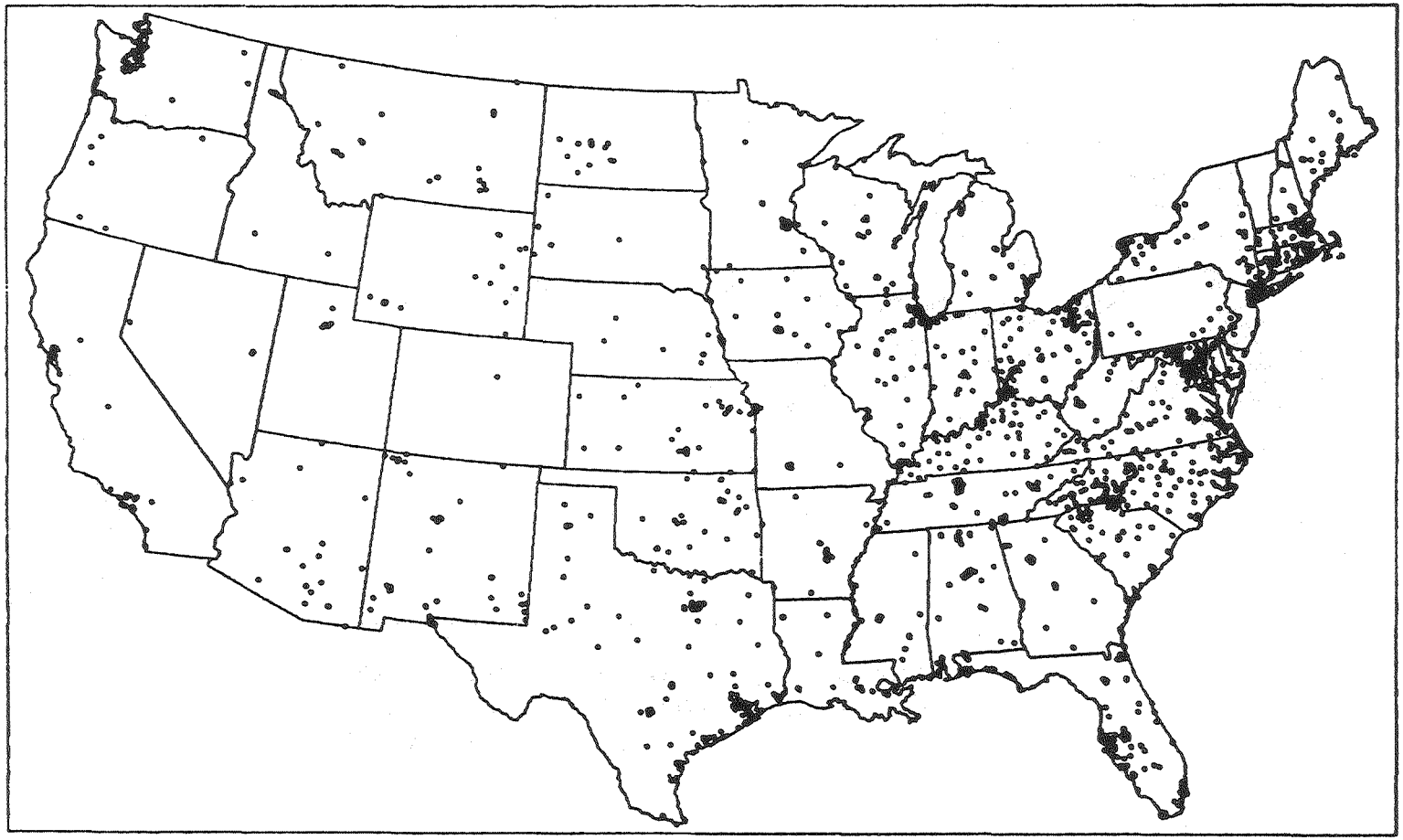


Figure 9.
1974–1976: Stations Measuring
Sulfur Dioxide
(1-hour sampling interval)

Figure 10.
1974–1976: Stations Measuring
Nitrogen Dioxide
(24-hour sampling interval)

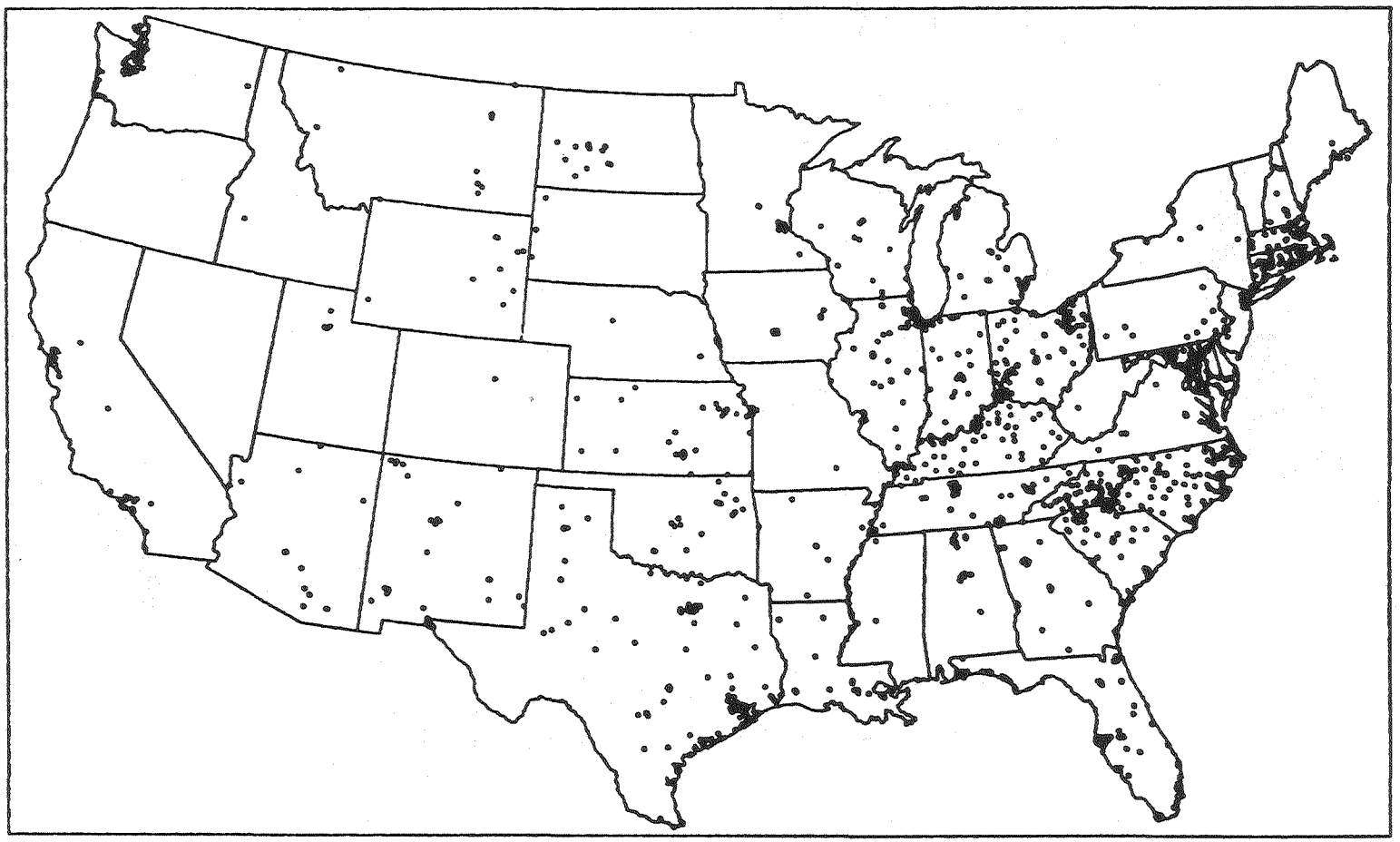


Figure 11.
1974–1976: Stations Measuring
Nitrogen Dioxide
(1-hour sampling interval)

APPENDIX A: SOURCES OF AIR QUALITY AND "CONTROL" DATA

The air quality and "control" data used in this analysis came from Seedis, the Socio- Economic Environmental Demographic Information System maintained by Lawrence Berkeley Laboratory [20]. The data in Seedis came originally from three separate sources:

Items 1-17 (Seedis file F):
1977 County and City Data Book [5]

Items 18-22 (Seedis file V):
1977 Area Resource File [6]

Items 23-25 (Seedis file Z)
1974-1976 Air Quality: County vs PUS Area [21]

Listed below for each data item are:

(a) the item number;

(b) a brief description;

(c) the Seedis database and data element code;

(d) where applicable, the corresponding item number on pages iv-vi of [5];

(e) where applicable, the starting column position on pages 27-82 of [6].

1. Land Area in Square Miles, 1970
   F.CCDBC0004 -- -----

2. Population, 1970
   F.CCDBC0016 7 -----

3. Net Migration Percent Change, 1970-1975
   F.CCDBC0031 13 -----

4. Population, Percent Urban, 1970
   F.CCDBC0041 8 -----

5. Population, Percent Black, 1970
   F.CCDBC0056 9 -----

6. Population, Percent Foreign Stock, 1970
   F.CCDBC0077 10 -----

7. Divorce Rate Per 1000 Population, 1970
F.CCDBC0117 27 -----

8. Persons, Percent Under 5 Years, 1970
F.CCDBC0060 -- -----

9. Persons, Percent 65 Years and Over, 1970
F.CCDBC0066 -- -----

10. Families, Percent With Income Less Than $3000, 1970
F.CCDBC0234 -- -----

11. Families, Percent With Income $15000 And Over, 1970
F.CCDBC0239 52 -----

12. Persons 25 Years or more, Percent With 4 Years College Or more, 1970
F.CCDBC0129 -- -----

13. Employed, Percent in Manufacturing, 1970
F.CCDBC0185 -- -----

14. Employed, Percent In Professional & Managerial Occupations, 1970
F.CCDBC0207 -- -----

15. Occupied units, Percent Lacking Some or all Plumbing, 1970
F.CCDBC0308 73 -----

16. Occupied units, Percent with 1.01 or More Persons/Room, 1970
F.CCDBC0310 74 -----

17. Occupied Units, Percent Owner Occupied, 1970
F.CCDBC0344 72 -----

18. January Temperature, 0.1 Degrees, 1976
V.C.6936 -- 12120

19. July Temperature, 0.1 Degrees, 1976
V.C.6940 -- 12124

20. January Precipitation, 0.01 Inches, 1976
V.C.6948 -- 12132

21. July Precipitation, 0.01 Inches, 1976
V.C.6952 -- 12136

22. Elevation, Feet, 1976
V.C.6970 -- 12154

23. Total Suspended Particulate, Geometric Mean Concentration, of County at Population Centroid, Micrograms per Cubic Meter, 1974-1976
Z.GMEAN.CNTRD.SPD -- -----

24. Sulfur Dioxide, Geometric Mean Concentration, of County

at Population Centroid, Micrograms per Cubic Meter, 1974-1976
Z.GMEAN.CNTRD.SDT -- -----

25. Nitrogen Dioxide, Geometric Mean Concentration, of County at Population Centroid, Micrograms per Cubic Meter, 1974-1976
Z.GMEAN.CNTRD.NDT -- -----

APPENDIX B: Incomplete Model Bias

The case of three independent variables is presented; the general case of k independent variables follows the same reasoning. Let the independent variable y have the following covariance structure:

$$\Sigma = \begin{vmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{vmatrix}$$

$$\Sigma_y = (\sigma_{y1} \ , \ \sigma_{y2})$$

and

$$\Sigma_3 = (\sigma_{13} \ , \ \sigma_{23})$$

The estimates of the regression coefficients when variable x_3 is omitted are

$$b = \Sigma_y \ \Sigma^{-1}$$

and the incomplete model bias caused by omitting x_3 is given by the expression

$$B = b_3 \ \Sigma_3 \ \Sigma^{-1}$$

where b_3 is the regression coefficient associated with x_3. Therefore, the unbiased estimates of the regression coefficients are $b-B$. The regression sum of squares can be expressed as a sum of the contributions from the included and excluded variables or regression sum of squares:

$$(b + B) \ \Sigma_y' + b_3^2 \sigma_3^2 [1-R_{(3)}^2]$$

The term

$$b_3^2 \sigma_3^2 [1-R_{(3)}^2]$$

represents the contribution from x_3 and will not be observed when x_3 is omitted from the model. The term $R^2_{(3)}$ is the squared multiple correlation coefficient where x_1 and x_2 are regressed on x_3. This form of the regression sum of squares demonstrates that the "explained" variation is always reduced when the model is incomplete. Furthermore, it should be noted that the incomplete model bias and the reduction in the regression sum of squares will be small when b_3 or σ^2_3 is small. The term $[1-R^2_{(3)}]$ shows that the analysis is relatively unaffected when the terms already in the model (x_1 and x_2 in this case) are highly correlated with the omitted variable (i.e., $1-R^2_{(3)}$ close to 1.0).

APPENDIX C: PRINCIPAL COMPONENTS OF AGE DISTRIBUTION

The age distributions for males and females can be represented by principal components P_1 and P_2, defined as follows:

$$P_1(m) = \sum_{i=1}^{i=11} c_{1i}(m)\ x_i$$

$$P_1(f) = \sum_{i=1}^{i=11} c_{1i}(f)\ x_i$$

$$P_2(m) = \sum_{i=1}^{i=11} c_{2i}(m)\ x_i$$

$$P_2(f) = \sum_{i=1}^{i=11} c_{2i}(f)\ x_i$$

where x_i is the fraction of the population in the ith age category. The definitions of the 11 age categories, and the coefficients c_{1i} and c_{2i} are as follows:

i	Ages	$c_{1i}(m)$	$c_{1i}(f)$	$c_{2i}(m)$	$c_{2i}(f)$
1	0	.77	.84	-.37	-.23
2	1-4	.85	.89	-.28	-.19
3	5-14	.88	.87	-.04	-.01
4	15-24	.38	.65	-.34	-.15
5	25-34	.80	.89	-.38	-.16
6	35-44	.89	.87	-.06	-.70
7	45-54	.75	.60	.45	.65
8	55-64	.42	.19	.83	.91
9	65-74	.18	-.01	.91	.91
10	75-84	.13	-.06	.93	.95
11	85+	.10	-.08	.85	.85