EPA 560/3-76-002

PILOT STUDY OF THE ASSOCIATION BETWEEN CANCER MORTALITY AND INDUSTRIAL CONCENTRATION IN 200 U.S. COUNTIES

Environmental Protection Agency
Office of Toxic Substances
Washington, D.C. 20460

December 1976

ABSTRACT

The Office of Toxic Substances is undertaking studies which will attempt to identify potentially regulable sources of health hazards in the environment. These studies involve statistical analyses of relationships between industrial concentrations and mortality rates from various diseases in all 3,069 U.S. counties. A pilot study, involving 200 counties, was undertaken to anticipate and resolve as many methodological problems as possible before encountering them in the full-scale effort. This report describes the results of that pilot study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PILOT STUDY OF THE ASSOCIATION BETWEEN CANCER MORTALITY AND INDUSTRIAL CONCENTRATION IN 200 U.S. COUNTIES

INTRODUCTION AND PURPOSE:

As part of an attempt to identify health hazards posed by potential regulatory targets, the Office of Toxic Substances (OTS) is undertaking a computerized statistical analysis of the relationship between industrial concentrations and mortality rates from various diseases in all 3,069 U.S. counties. Since this involves compiling and correlating several massive data sets heretofore not considered conjunctively, many unforeseen methodological problems can arise which, once the full-scale project is under way, would be expensive or impossible to correct.

To anticipate and resolve as many such problems as possible, OTS implemented a smaller, more manageable, 200-county pilot study. The experience gained in this pilot study, carried out by the same OTS staff members who developed the methodology for the full-scale project, was incorporated directly into that larger effort. This report describes the methodological problems discovered during the pilot study, and discusses how they are to be handled in the full-scale project. It is in four sections:

1. Introduction and Purpose
2. Analysis of 200 County Data

   Describes the difficulties encountered in defining the relationship between four broad categories of cancer and 30 selected industries in the 200 counties.

3. Methodological Problems

Describes the methodological problems discovered in the pilot study and the steps taken to avoid them in the full-scale project.

4. Methodological Analyses

Describes some additional analyses done to shed light on methodological issues.

ANALYSIS OF THE 200 COUNTY DATA:

This pilot study was limited to an analysis of the relationship between 30 industries and four broad categories of cancer mortality among the 200 most populous counties in the U.S. The mortality rates were adjusted to neutralize the effect of certain known non-industrial correlates of mortality. As shown by the following summary, the pilot study was far more modest in scope than the full-scale project will be.

Pilot Study	Full-Scale Project
200 most populous counties in the continental U.S.	All counties in the continental U.S.
30 selected industries in 1959.	Over 600 manufacturing, extractive and transportation industries in 1959, 1967, 1973.
Four broad categories of cancer.	All causes of death divided into 56 categories.
Crude mortality rates for total county population.	Age-adjusted mortality rates by race and sex for the age band 35-74.
10 demographic control variables.	Approximately 40 non-regulable control variables.

A. Description of data sets

1. Industrial

The index of industrial concentration for each county and specific industry was the percent of the total county workforce engaged in the specific industry in 1959 expressed as a multiple of the percent of the total U.S. workforce in that industry in the same year. Thus,

$$\text{Index} = 100\ \frac{CWSI/CWF}{USSI/USWF}$$

Where: CWSI = Number of county workers in the specific industry.

CWF = Total county workforce.

USSI = Number of U.S. workers in the specific industry.

USWF = Total U.S. workforce.

Thus, if the index for a specific industry in a specific county is 800, the percent of that county workforce employed by that industry is eight times the percent of the U.S. workforce employed by the industry.

For most of the statistical analyses discussed here, it makes no difference that the county percentage (CWSI/CWF) is divided by the U.S. percentage (USSI/USWF) for the same industry; the index is equivalent to the county percentage alone. However, when index values for several industries are added together to produce a group index, division by the U.S. percentage does make a difference. This procedure will be discussed in Section F.

Social Security employment figures published by the Census Bureau were used in computing the index (1). Although exact employment figures were withheld in some instances, they were estimated from other data available in the same publication.

Thirty specific industries, listed in Table 1, were selected for analysis because of their relatively high concentration of workers in counties with high age-adjusted cancer mortality rates among white males during the period 1950-1969. Table 2 lists the cancer mortality categories that were selected for study.

(1) U.S. Bureau of the Census and U.S. Bureau of Old-Age and Survivors' Insurance; Cooperative Report, County Business Patterns, First Quarter 1959, Part 1, U.S. Summary. Washington, D.C., 1961.

TABLE 1

Thirty Industries Selected for Study

Abbreviation	S.I.C. Number	Industry
IRONMIN	101	Iron ore mining
ANTHMIN	1111	Anthracite mining
PETROMIN	131	Crude petroleum and natural gas mining
OILSERV	138	Oil and gas field services
SULFMIN	1477	Sulfur mining
BEER	2082	Malt liquors
CIGARS	212	Cigars
NAROFAB	224	Narrow fabrics and other smallwares mills: cotton, wool, silk and man-made fibers
BROADFAB	2262	Finishers of broad woven fabrics of man-made fiber and silk
BLOUSES	2331	Women's, misses', and juniors' blouses, waists, and shirts
DRESSES	2335	Women's, misses', and juniors' dresses
COALTAR	2814	Cyclic (coal tar) crudes
DYES	2815	Dyes, dye (cyclic) intermediates, and organic pigments (lakes and toners)
ORGANICS	2818	Industrial organic chemicals, not elsewhere classified
PLASTICS	2821	Plastics materials, synthetic resins and non-vulcanizable elastomers
ELASTICS	2822	Synthetic rubber (vulcanizable elastomers)
DRUGS	283	Drugs
MEDCHEM	2833	Medicinal chemicals and botanical products
FRTLIZER	2871	Fertilizers
PETROREF	291	Petroleum refining
ASPHALT	2952	Asphalt felts and coatings
RUBRPROD	306	Fabricated rubber products, not elsewhere classified
GYPSUM	3275	Gypsum products
STLMILLS	3312	Blast furnaces (including coke ovens), steel works and rolling mills
STLFOUND	3323	Steel foundries
COPSMELT	3331	Primary smelting and refining of copper
COPRLNG	3351	Rolling, drawing, and extruding of copper
WIRE	3357	Drawing and insulating of nonferrous wire
DOCKS	4462	Piers and docks
STEVEDOR	4463	Stevedoring

Source: U.S. Bureau of the Budget, Standard Industrial Classification Manual, prepared by the Technical Committee on Industrial Classification, Office of Statistical Standards, Washington, D.C., 1957.

TABLE 2

Cancer Mortality Categories Selected for Study

Abbreviation	I.C.D. Numbers	Disease Category
TOTALCAN	140-209	Malignant neoplasms, including neoplasms of lymphatic and hematopoietic tissues
DIGSTCAN	150-159	Malignant neoplasms of digestive organs and peritoneum
RESPCAN	160-163	Malignant neoplasms of respiratory system
URINCAN	188-189	Malignant neoplasms of urinary organs

Source: World Health Organization, Manual of the International Statistical Classification of Diseases, Injuries, and Causes of Death, 8th Revision, Vol. 1, Geneva, 1967.

2. Mortality

Mortality rates per 100,000 population by county for total cancer, respiratory cancer, digestive cancer, and urinary cancer, were calculated from published mortality (2) and population (3) data. Crude rates were computed because deaths by age are not published at the county level.

3. Non-industrial correlates

The following non-industrial correlates of mortality were used as control variables: population, median age, percent aged 65 and over, percent non-white, population per square mile, percent urban, median family income, median last school year completed, physicians per 100,000 and hospital beds per 100,000 (3,4).

A number of counties and cities in Virginia and the five boroughs comprising New York City were combined into single observations to compensate for inconsistencies between data sources. The methods used to produce aggregate values for these observations are described in a computer printout available from OTS.

(2) U.S. Public Health Service. Vital Statistics of the United States 1970, Vol. 2, Mortality, Part B, Tables 7-9. Rockville, Md., 1974.

(3) U.S. Bureau of the Census. Census of Population 1970, Vol. 1, Characteristics of the Population, Part 1, U.S. Summary, Section 1, Chapter A, Number of Inhabitants, Table 24. Washington, D.C., 1973.

(4) American Medical Association. Distribution of Physicians in the United States 1970, Regional State, County, Metropolitan Areas, Table 12, Medical Practice Data by State and County. Chicago, Illinois, 1971.

B. Stepwise regression

Stepwise regression was used to remove the effects of non-industrial correlates from the mortality data and to select industries significantly related to the resulting residual mortality.

Regression analysis is a procedure for determining the mathematical relationship between a dependent variable and a set of independent variables when that relationship is not exact (i.e., when there is random error). There are many possible techniques for determining that relationship; the standard technique used in the pilot study is known as linear least squares. Linear means that the relationship between the independent and dependent variables is assumed to be linear. If there is only one independent variable, the relationship is depicted by a straight line on a graph. For three or more variables the relationship is still conceptually a straight line, although it can no longer be depicted on a single graph. Least squares is a method for selecting the most suitable straight line.

Stepwise regression is a method for selecting the "best" regression equation by including only those independent variables which improve the degree to which the mathematical equation fits the observed data by more than some predetermined amount. It is called stepwise because it selects variables one at a time in a series of steps. The first variable selected is that which has the highest significant correlation with the dependent variable. Subsequent variables are selected on the basis of a partial correlation between an independent variable and the dependent variable after the effects of all previously selected independent variables have been removed. The stepwise regression used in the pilot study retained only those variables whose correlations were significant at probability = .10 level.

The set of independent variables selected by this procedure, and the estimated slopes which describe their mathematical relationship to the dependent variable, are known collectively as the independent variable model. This language is used whether or not model building is the primary purpose of the analysis. Thus this report will refer to the "demographic model" or the "industrial model".

An important characteristic of stepwise regression is that it selects only those independent variables which have a significant relationship to the dependent variable over and above the effect of all other independent variables in the model. Thus a variable may fail to be selected even though, taken alone, it has a high correlation with the dependent variable. For example, both median age and percent aged 65 and over are highly correlated with cancer mortality but one or the other alone may sufficiently account for the relationship of both in the demographic model.

C. Removal of the effects of demographic and health care variables

Purpose and procedure: in order to isolate and identify the role of industry, it was first necessary to control for the non-industrial factors correlated with mortality. Because it is capable of handling a large number of variables simultaneously, regression analysis was used to remove the effects of the ten demographic and health care variables from the mortality data before the industrial analysis was performed. It is worth noting that some recent studies of the relationship between industry and cancer lack such safeguards. The stepwise procedure was used to make the computation more efficient by eliminating non-essential variables, not because of any interest as to which non-industrial variables fell into the model. The following procedure was used to remove the effects of the non-industrial variables from each of the four cancer mortality variables:

a. Determination of the non-industrial model by stepwise regression

b. Computation of the predicted mortality for each county based upon the model determined in (a).

c. Computation of the residual mortality for each county by subtracting the predicted from the observed mortality.

Thus the residual mortality for a particular county may be positive or negative depending upon whether that county's observed mortality is higher or lower than the non-industrial model would predict. A residual of 12, for example, indicates that there were 12 more deaths per 100,000 population than would have been predicted on the basis of median age, population per square mile, etc. After a statistician joined the study, another step was included:

d. Examination of scatter plots of the residuals to insure elimination of all observable effects of non-industrial variables.

This step turned out to be important; it was discovered that a potentially biasing curvilinear relationship remained between residual mortality and several non-industrial correlates. To remove these effects, additional non-industrial variables were created from the cross-products of the old ones (e.g., median age squared, population per square mile multiplied by income). These variables have little meaning in themselves but provide a way of depicting a curvilinear relationship with an essentially linear model. (True non-linear regression is a much more complex process.)

Steps (a) through (d) were repeated with the newly created variables added to the set. This process removed most of the curvilinear effect, but some irregularities remained in the median age and percent aged 65 and over variables. They appeared to be due to a single observation (Pinellas, Florida) with a very high median age. Since these irregularities could not be removed without the undesirable effect of excluding nearly five percent of the 200 counties, the Pinellas data remained.

D. Development of the industrial model

The relationships between the residual mortalities computed in Section C and the 30 specific industries were determined by stepwise regression. At this stage the model itself is of primary interest, particularly the set of industries falling into it.

As shown in Table 3, there were 12 instances of significant relationships between an industry and cancer mortality. Two industries were associated with both total and digestive cancer. Two relationships were negative, indicating lower than expected cancer mortality; eight industries had significant positive relationships to one or more of the four categories of cancer.

These results will not be detailed for reasons which will become apparent in the next section. However, two points need to be made:

1. Several of the significant relationships were due entirely to one or two counties because of the extremely skewed distribution of the industrial variables. Each industry was contained in only a small minority of the counties (ranging from a low of one percent for sulfur mining to a high of 35 percent for rubber products). However, even among these few counties the distribution was skewed, most having fairly low index values while one or two were extremely high. If the few counties with high index values also had high residual mortalities, there would be a significant positive relationship regardless of mortality rates in the other counties. Similarly, if they had low residual mortalities, there would be a significant negative relationship. One or two counties were responsible for both negative relationships and two of the positive relationships shown in Table 3.

Table 3

Industries Selected by Stepwise Regression as Significantly Related to Residual Cancer Mortality

Type of Cancer	Positively Related	Negatively Related
Total cancer	Iron ore mining	
	Dye and pigment manufacturing	
	Organic chemical manufacturing	
Respiratory	Sulfur mining	Dress manufacturing
	Steel works & mills	
	Stevedoring	
Digestive	Iron ore mining	Fertilizer manufacturing
	Dye and pigment manufacturing	
	Gypsum products	
	Wire drawing	

2. As pointed out in Section B, an industry could be significantly related to residual mortality but fail to be selected by the model because of its correlation (i.e., concentration in the same counties) with another industry. Comparison of Table 3 and Table 4 shows that fewer than half the industries significantly correlated with residual respiratory cancer were selected by the model.

E. Testing of the industrial model

In order to avoid spurious significance and instability produced by intercorrelations among industries, the relationship between the industries and residual cancer mortality was more stringently tested. The 200 counties were split into two groups of 100 counties each by random assignment and an industrial model was developed independently for each. The test was whether an industry was selected into both models. The traditional method for testing in independent groups is to develop the model with one group and apply the resulting mathematical equation to the other to determine how well it predicts the dependent variable values of the second group. But because this pilot study was more concerned with the industries chosen than with the estimated slopes of the equation, the procedure of developing two independent models was used.

The two independent analyses produced entirely different sets of industries (Table 5). Three factors were responsible for this result.

Table 4

Correlation Coefficients for All Industries Significantly Correlated with Residual Respiratory Cancer

Industry (abbreviation)	Correlation Coefficient
SULFMIN	.205
STLMILLS	.117
STEVEDOR	.159
DRESSES	-.210
PETROREF	.117
ORGANICS	.148
ANTHMIN	-.201
DOCKS	.146
OILSERV	.130

Table 5

Industries Selected by Stepwise Regression as Significantly Related to Residual Mortality in Each of Two Random 100-County Groups

Cancer Type (Residual)	"Group 0" Positively correlated industries (abbreviation)	"Group 0" Negatively correlated industries (abbreviation)	"Group 1" Positively correlated industries (abbreviation)	"Group 1" Negatively correlated industries (abbreviation)
Total	IRONMIN ORGANICS PLASTICS DYES STLMILLS	PETROREF	STEVEDOR GYPSUM WIRE	----------
Respiratory	DOCKS	----------	MEDCHEM STEVEDOR	CIGARS
Digestive	IRONMIN PLASTICS COALTAR DYES	----------	GYPSUM	FRTLIZER
Urinary	ANTHMIN BEER	BLOUSES	WIRE DRESSES ELASTICS	CIGARS MEDCHEM

1. This was intentionally a stringent test equivalent to using the .01 level of significance. Reducing the number of observations from 200 to 100 further decreased the probability of detecting a true difference. This problem will be mitigated somewhat in the full-scale study when there will be over 1500 observations in each group.

2. As previously observed, there were very few counties with high concentrations of each industry; there were fewer still when the number of counties was cut in half. Estimates of residual mortality at the high industrial concentrations, being based on few observations, were unstable, and therefore produced unstable estimates of the slope.

3. Results of additional analyses of the industrial variables indicated that multicolinearity (correlation of industries with each other) was a serious problem in this data set, with some correlations greater than .80. The high multicolinearity of a set of independent variables can produce estimates so unstable that a small difference in the data set can make a very large difference in the estimates and hence in the variables selected by stepwise regression. This is almost certain to be a problem in the full-scale study because of the many independent variables, even though the correlations are expected to be much lower for industries not in this set.

F. Combining industries to reduce multicolinearity

Three general approaches to dealing with the problem of multicolinearity were considered:

1. Although use of recently developed sophisticated statistical techniques (e.g. ridge regressions) would produce better regression estimates for mathematical relationships they would be less useful for the variable selection in this study. Because these techniques are new, reliable computer programs are not widely available to compute calculations and the techniques themselves require more explanation than the more familiar tests used here.

2. Individual regressions for each industry would avoid computational problems and would be easy to understand. However, with the number of variables involved in the full-scale study, it would be very expensive. Also, there would be the possibility that the significant correlation of one industry with another would be ignored. An analysis should underline this effect rather than ignore it.

3. The industrial data set could be reduced by combining correlated industries. Industries too highly correlated for their effects to be separated would be treated as an entity. This procedure was used in the pilot study and will probably be used in the full-scale study.

Although it is easy to describe and does not require new computer programs, there are several disadvantages. One disadvantage is the necessity of combining industries with nothing in common except location, thus producing a new variable which is difficult to refer to except by a listing of the component industries. Another is the possibility that some relationships between individual industries and mortality would be obscured. Additional analyses are under way to resolve the problem.

The procedure used in this study was to combine highly correlated industries while also keeping similar segments together as much as possible. The Standard Industrial Code (5) was used as a guide in combining industries as follows:

1. Correlated industries within the same three-digit code were combined by addition of index values; correlations were computed for the new set of industry groups.

2. Correlated industries within the same two-digit code were combined; correlations were computed for the new set.

(5) U.S. Bureau of the Budget. Standard Industrial Classification Manual. Prepared by the Technical Committee on Industrial Classification, Office of Statistical Standards, Washington, D.C., 1957.

3. Industries were combined on the basis of size of correlations until a set of essentially uncorrelated industries was produced. (A mathematical characteristic of the correlation matrix determined when the set was essentially uncorrelated.) This procedure produced a set of six uncorrelated industry groups, defined in Table 6. The reduction in the number of industries is not expected to be as drastic in the full-scale study where correlations will be lower on the average. Note that the sums were weighted to protect against domination of the combined variable by large industries. This followed naturally from the way the original index was computed.

G. <u>Relationships between combined industries and residual mortality</u>

The relationship between the residual mortalities computed in Section C and the newly defined set of six industry groups was determined independently by stepwise regression for each of the randomly selected subgroups of 100 counties. Results are shown in Table 6.

The "Combo 1" industry group was selected by both of the independently determined models as significantly related to residual respiratory cancer mortality. This was an indication that the failure to find any overlap in the 30-industry test was largely due to multicolinearity; the probability of finding overlap by chance alone would be much less with six independent variables than with 30.

Table 6

The 30 Selected Industries Combined into Six Groups

Group (abbreviation)	Industry (abbreviation)
ANCICLTH	ANTHMIN
	BLOUSES
	CIGARS
	DRESSES
BEER	BEER
COMBO 1	DOCKS
	ELASTICS
	FRTLIZER
	GYPSUM
	OILSERV
	ORGANICS
	PETROREF
	PETROMIN
	STEVEDOR
	SULFMIN
COMBO 2	ASPHALT
	COALTAR
	DYES
	STLFOUND
	STLMILLS
COMBO 3	BROADFAB
	COPRLNG
	COPSMELT
	*DRUGS
	NAROFAB
	PLASTICS
	RUBRPROD
	WIRE
IRONMIN	IRONMIN

*Since MEDCHEM (S.I.C. 2833) is a subdivision of DRUGS (S.I.C. 283), it is not listed separately.

The regression line is shown in Figure 1 for each of the 200 counties according to their values on combined index and residual respiratory mortality. The five counties with highest values on the combined variable are given along with their ranking among the 200 counties for each industry contained in Combo 1. Figure 1 illustrates how the correlation among industries necessitated combination. The two counties which contributed most to this regression, Jefferson, Texas, and Jefferson, Louisiana, contained 7 and 8, respectively, of the industries included in Combo 1; together they contained all 10. It would be misleading to report a relationship between respiratory cancer and the Combo 1 industries without also indicating that the same few counties were responsible for all of these correlations. There is no way to determine which industries if any were responsible for increased respiratory cancer mortality without followup studies.

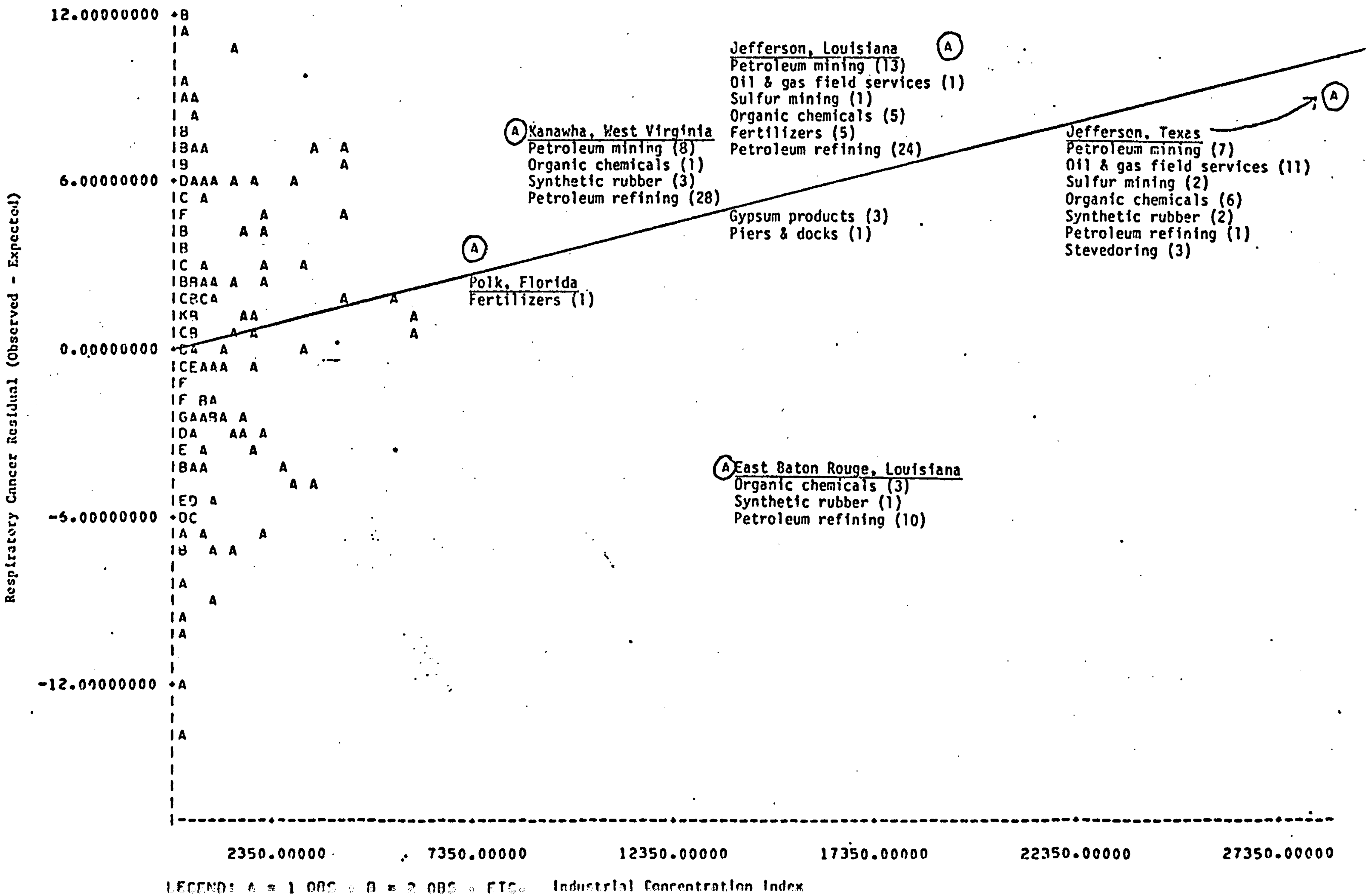
Studies relating industry to cancer mortality by county which deal with only one, or very few, industries, selected a priori, can be misleading. For example, a recent study by Thomas Mason (6) reported an increase of lung cancer among males and females in four counties with synthetic rubber manufacturing due largely to the high concentration of synthetic rubber manufacturing in Jefferson, Texas. However, as Figure 1 shows, although Jefferson, Texas, ranked high in concentration of synthetic rubber manufacturing it ranked even higher in petroleum refining and equally high in sulfur

(6) Mason, Thomas J. Cancer Mortality in U.S. Counties with Plastics and Related Industries. Environmental Health Perspectives, 2:79-84, June 1975.

Figure 1

PLOT OF RESPIRATORY CANCER RESIDUAL BY CONCENTRATION

OF GROUP OF CORRELATED INDUSTRIES*



mining. In the full-scale study all sectors of a wide range of industries will be screened and correlated industries will be combined to reduce the likelihood of drawing misleading inferences.

H. Summary

As an integral part of the analysis of the pilot study data, several methodological issues were identified, addressed and resolved:

1. The control procedure for non-industrial correlates was expanded to include non-linear relationships.
2. A method was developed and used to reduce spurious results by independent testing.
3. The problem of multicolinearity was studied and a viable solution adopted.
4. Excess respiratory cancer mortality rates were found in counties containing a group of correlated industries. This finding brings into question the results of some other recent studies.

METHODOLOGICAL PROBLEMS

The major purpose of the pilot study was to anticipate and resolve problems prior to implementing the full-scale project. In this respect, it was a success; many methodological problems and deficiencies were discovered. In terms of resolution they fall into three categories:

A. Those which could be and were resolved in both studies;

B. Those which were resolved for the full-scale project but could not be resolved in the pilot study;

C. Those for which we have not yet been able to find a solution and which may be insoluble.

The problems in each of the three categories are discussed below in greater detail in other sections of this report and in the Proposed Methodology Description, available from OTS, for the full-scale study.

A. Problems resolved in both studies

1. Intercorrelations among industries

Some of the industries in the pilot study were highly correlated with one another, i.e., concentrated in the same counties. These correlations created difficulties in the analysis by multiple regression techniques because they reduced the efficiency of estimates and significance tests. They also increased the possibility of wrongly attributing increased mortality to a particular industry. A similar problem is expected in the full-scale study. In the previous section a method for resolving this problem is described.

The intercorrelation of industries is a problem in any geographic study of mortality and industrial concentration whether or not this is recognized or acknowledged by the authors. As noted earlier, some recent country mortality studies attributed increased cancer mortality to a particular industry when it could have been equally attributed to any one of several other industries.

2. Oversimplification

Statistical summaries of numerous complex relationships give an incomplete picture and sometimes distort very dissimilar relation ships, causing them to appear similar. Plots of significant regressions were used to flush out the summary statistics in the pilot study and this same procedure will be used for the fullscale study.

3. Nonlinear relationships between control variables and mortality rates

The regression techniques used in the pilot study (and also to be used in the full scale study) to control for nonindustrial variables with a known or suspected relationship to mortality are convenient for handling a large number of variables but present difficulties with outliers and nonlinear relationships.

During the course of this study it was discovered that some of the control variables, most notably the two age variables, were curvilinearly related to cancer mortality. The problem was largely corrected, as previously described, although a small margin of error could not be corrected in the pilot study without arbitrarily excluding many counties. However, two major improvements were made for the fullscale project as a result of this discovery: The regression analyses of the control variables will allow for curvilinear relationships, and age-adjusted rates will be used.

B. Problems resolved in the fullscale study but not in the pilot study

1. Uncontrolled variables

   Many variables frequently shown to be correlated with cancer mortality were not considered in this pilot study although they will be included in the fullscale project. Some uncontrolled variables include elevation, latitude, mean precipitation and hours of sunlight, water hardness, alcohol consumption and gasoline sales.

   Other important variables were not considered because the data were unavailable. Cigarette smoking is one such important variable for which data are unavailable at the county level.

   Many studies which relate certain industries to cancer mortality, including some recently reported, have no adequate control over related variables. In this respect, this pilot study is superior.

2. Migration of industries

   The migration of industries between counties could lead to distortion of mortality resulting from industrial exposure. Industrial concentration data for several points in time, beginning in 1959, will therefore be examined for the full-scale study.

3. Latency period

   The latency period between exposure and mortality was assumed in the pilot study to be 11 years (1959 to 1970). This was the longest latency period allowable with the available data. It is usually assumed that the latency period for cancer is longer than this and it doubtless varies according to the type of disease.

The full-scale study will use industrial distributions for 1959, 1967, and 1973, thus permitting analysis of three latency periods. The best latency period can then be empirically determined for each disease. However, the longest latency period studied will still be a few years short of that generally assumed for cancer because employment data prior to 1959 are not available in the level of detail needed for this analysis.

4. Commuting

Intercounty commuting could obscure the effects of industries on mortality rates since deaths are assigned to the county of residence. Thus deaths resulting from daytime exposures, whether occupational or ambient, could be attributed to the wrong county when the worker commutes across the county line. Since these deaths are not related to industries within the county, the relationships between mortality and industrial concentration would be obscured. There are two types of commuters for any given county: the residents who commute outside the county to work, and the non-residents who commute into the county to work. There was a much stronger relationship between industry and cancer in those counties with a small proportion of workers commuting outside the county than in all 200 counties considered together. It is anticipated that information on commuting will also be used in the full-scale study to reveal such relationships and to help differentiate between occupational and industrial exposure.

5. Age bias

The mortality data were neither age-specific nor age-adjusted. An attempt was made to correct this deficiency by using median age as a control variable, but it was not adequate. The full-scale project will use age-adjusted rates for the age band 35-74.

6. Race/sex aggregation

The mortality data were presented for the total population and were not broken down by race or sex. Since mortality rates for specific cancer sites differ sharply by race and sex, this seriously clouded many relationships. These problems are discussed further in the next section. the full-scale project will use race and sex specific rates.

7. Small numbers

There is a possibility that the mortality rates in the pilot study were unstable since they were based on data for a single year in small geographic units, and therefore sometimes based on only a small number of deaths. The full-scale project will use five years of data and will delineate causes of death in greater detail.

8. Limited disease categories

Cancer mortalities were the only health measures used. For reasons discussed in a report of mortality trends in preparation by OTS, cancer is probably not the most promising indicator of health effects due to industry. The full-scale project will cover all mortalities, divided into 56 specific etiologic categories. The need for expansion of cause-of-death categories is further discussed in the next section.

9. Migration of county residents

The relationship between mortality and industrial concentration may be obscured by migration between counties when workers are exposed to the industry in one county but die in another county not containing the industry. High rates of migration into retirement areas or out of Appalachian coal fields represent obvious examples of the regression inaccurately reflecting mortality as a result of exposure to locally generated industrial pollution. At least two measures of migration will be used for the full-scale study to partially correct for this factor: net migration rate from 1960 to 1970 and percent of units occupied during 1965-1970.

10. Small population counties

The statistical methods used in the pilot study weighed each county equally, regardless of its population. However, because counties with larger populations have more stable mortality rates, they should be given more weight to produce a more accurate estimate of the relationship between mortality rates and other variables. The full-scale study will use regression analysis to weigh each county according to the population used in computing the mortality rate.

11. Selected counties

Since only the 200 most populous counties were studied, the results cannot be generalized to the entire U.S. The full-scale project will deal with all 3,069 U.S. counties.

12. Selected industries

The pilot study included 30 industries which were selected by a method which produced a highly intercorrelated set. Of these, only about six industry groups were sufficiently independent so that the effects of one could be separated from the effects of the others.

The full-scale study will screen all manufacturing, mining, transportation, agricultural, forestry, and fishery industries. They will not be pre-selected and are expected to have lower correlations than the industries in the pilot study.

13. Community vs. occupational exposure

Although the pilot study was most concerned with community rather than occupational exposure, in practice it was difficult to distinguish the results of one from the other.

Counties with occupationally related effects can be isolated by the use of race and sex specific rates, the addition of other variables such as percent commuting outside the county to work, and knowledge from other sources about the effects of exposure in certain industries. While these approaches were beyond the scope of the pilot study, most have been built into the full-scale project, and all will be pursued in that study.

C. Problems not resolved in either study

1. Biased industrial index

The industrial index value for a specific industry in a given county is derived by dividing the number of workers in that industry employed in that county by the county's total workforce. The purpose of the ratio is to permit valid county-by-county comparisons of the relative significance of the size of an industry workforce, after controlling for county size.

The denominator in the ratio is designed to avoid exaggerating the significance of the size of an industry workforce in large, densely populated or high employment counties or underestimating its significance in small, sparsely populated or low employment counties. Thus, the index values stress those industries that are located in counties with small workforces or are heavily concentrated in a few areas. This bias is purposely built into the methodology to avoid the large county bias noted above, but the present index may overcorrect for county size.

Alternative indices were considered for the full-scale study. The one selected, suggested by Christoper Gordon of System Science, Inc. (SSI), divides the square of the industry workforce by the total county workforce, thus giving more weight to counties with large workforces than the index used in the pilot study.

It will be recognized that the problem is inherent in the methodological approach. In particular, the adverse health effects of those industries that are distributed in general acccordance with the population distribution (e.g., gasoline service stations) cannot be detected by any index.

2. Surrogate measures

The industrial index is obviously not a direct measure of an industry's environmental pollution. The requirement for an index with interindustry comparability, and our general ignorance regarding the true adverse health effect potential of specific pollutants, both make such a surrogate index essential.

3. Chance findings

The probability levels reported in the pilot study applied to individual regressions, or to multiple regressions after individual regressions were scanned and selected to provide a strong multiple relationship. Since so many regressions were run or scanned, one would expect to find some "significant" at the .05 or .10 level by chance alone. Therefore the probability levels should not be taken literally. This is true of all large-scale hypothesis-formulating studies. The probability of spurious results will be reduced in the full-scale study but cannot be eliminated.

4. Community exposure across county lines

People living and working in one county may be exposed to industrial emissions from another county if, for example, the industry is near county line or if the water supply is polluted.

Such exposures would tend to obscure the effects of industrial exposure in the approach used in the pilot study and in the full-scale study. This is an unavoidable limitation of a large-scale correlational analysis.

METHODOLOGICAL ANALYSES:

Many small studies have been done by OTS to elucidate methodological problems. Only a few are reported here. Others are too technical for general interest or are still in progress. One (mortality trends analysis) is the subject of a special report.

A. Importance of race and sex specific rates

Data for race and sex specific mortalities by county were not in the published tables upon which the pilot study was based. These data will be available for the full-scale study.

There is a great difference in rankings of county mortality by sex and by color. Tables 7 and 8 show the 200 county ranks (200 were ranked, although only 15 appear in each table) on the age-adjusted respiratory cancer mortality rates in 1950-1969 for the specific race-sex group indicated. Table 7 shows the top 15 counties ranked on rates for white males. Counties similarly ranked for each of six cancer types were used to select the 30 industries examined in the pilot study. Table 8 contains the top 15 counties ranked on rates among non-white females. Only one county, Hudson, N.J., occurred in both ratings. Had industries been selected from counties with high rates among non-white females instead of white males, an entirely different set of counties would have been included, which may have led to a completely different set of industries. This problem will not occur in the full-scale project, of course, since all industries are being screened.

The average rankings in Table 7 are low for white males, of course, (indicating high rates) and higher for the other three race-sex groups (indicating lower rates). They differ much more by race than by sex. Similarly the average rankings in Table 8 are low for non-white females and also differ more by race than by sex. This has significance for the full-scale project because it suggests an approach for dealing with the occupational versus community exposure problem. Counties with high rates for both sexes of the one or the other race may reflect differential housing patterns and hence community exposure. They may also reflect some other living pattern aspect such as diet. High rates for males of both races may be due to occupational exposure.

Charleston, South Carolina (Table 7) had the highest mortality rates from respiratory cancer for white males and white females but close to the lowest for non-white groups; Northampton, Pennsylvania (Table 8) had just the opposite situation. Because of the racial difference in rates, if an industry associated with respiratory cancer mortality were present in one of these counties it would suggest that residential, rather than occupational, patterns may be involved. This information is lost when only one race-sex group is used or when all are combined in a total rate, as was the case in the pilot study. These data illustrate the need for race-sex specific rates.

Table 7

200-County Ranking of Age Adjusted Mortality Rates
for Respiratory Cancer, 1950-1969
Top 15, Ranked by White Male

County and State	RANKING WHITE Male	RANKING WHITE Female	RANKING NONWHITE Male	RANKING NONWHITE Female
Charleston, SC	1	1	181	151
Orleans, LA	2	30.5	23.5	51
Jefferson, LA	3	26	11	143
Baltimore City, MD	4	43.5	21	87
Norfolk City, VA	5	6	108	95.5
Mobile, AL	6	26	128	155
Hudson, NJ	7	32	9	7
Duval, FL	8	20.5	99.5	151
Middlesex, NJ	9	107	5	25
Chatham, GA	10	20.5	160.5	118
Caddo, LA	11	35.5	164.5	168.5
Harris, TX	12	4.5	85	99.5
Jefferson, TX	13	16.5	77.5	38
Prince Georges, MD	14	16.5	155	102
Hillsborough, FL	15	26	124	107
AVERAGE:	8	27.5	90.2	99.9

Table 8

200-County Ranking of Age-Adjusted Mortality Rates for Respiratory Cancer, For Four Race-Sex Categories 1950-1969

Top 15, Ordered by Nonwhite Female Rankings

	RANKING			
	WHITE		NONWHITE	
County and State	Male	Female	Male	Female
Northampton, PA	153.5	157.5	2	1
New London, CT	113	146.5	116	2
Winnebago, IL	190.5	136.5	36	3
Lancaster, PA	194	141	83.5	4
Cumberland, ME	61.5	107	191	5
York, PA	192.5	191.5	87	6
Hudson, NJ	7	32	9	7
Polk, IA	76	168	12	8
Washington, PA	135	168	110	9
Will, IL	150	182	168	10
Albany, NY	23.5	89.5	23.5	11.5
Westmoreland, PA	181.5	182	106	11.5
Sonoma, CA	155.5	68.5	167	14
Lucas, OH	47.5	107	35	14
Nueces, TX	66.5	12	70.5	14
AVERAGE:	116.5	125.9	81.1	8.0

B. Effects of commuting

Occupational exposure of workers who commute outside the county of residence could obscure the effect of any industrial exposure inside the county. To determine the extent to which this factor affected relationships in the pilot study analysis, the percent of county workforce commuting outside the county was added to our data set.

Table 9 shows results of the stepwise analysis for two groups of counties, divided according to the percent working outside the county. A ten percent level was used since it divided the 200 counties nearly down the middle. As the table shows, more industries were significantly related to cancer mortality residuals and the relationships for all four cancer types were generally stronger among those counties with fewer than 10 percent of the workers commuting. The individual plots show that almost all the industries related to a type of cancer for the counties with more than 10 percent commuting were also related for the counties with less than 10 percent. They sometimes failed to be selected by stepwise regression, evidently because of their high correlation with an industry which was selected. In most cases the counties with more than 10 percent commuting had fewer and weaker significant regressions than the 200 counties combined. An exception was respiratory cancer, where both the counties above and below 10 percent showed more positive regressions on industry than did all 200 counties. Further study is required to determine the kind of relationship

Table 9

Comparison of Stepwise Regressions on Industry
for Counties by Percent of Workforce Commuting

Type of Cancer	Industry (Abbreviation)	Less than 10% Commuting (Base=91 Counties) R	10% or More Commuting (Base=109 Counties) R	All 200 Counties R
Total Cancer	IRONMIN	.33		.19
	DYES		.18	.14
	ORGANICS	.29		.16
	ASPHALT	-.24		
	DRUGS	.21		
	STLFOUND	.17		
	All significant industries	.53	.18	.29
Respiratory Cancer	SULFMIN		.28	.17
	STLMILLS		.15	.13
	STEVEDOR		.22	.12
	DOCKS	.24		
	DRUGS	.22		
	PETROREF	.20		
	PLASTICS	.17		
	DRESSES	-.23		-.19
	COPSMELT		.21	
	CIGARS		-.30	
	FRTLIZER		-.17	
	All significant industries	.48	.49	.33
	Positive R only	.43	.32	.27
Digestive Cancer	IRONMIN	.23		.25
	GYPSUM			.17
	DYES		.17	.13
	WIRE			.12
	ASPHALT	.40		
	STLMILLS	.23		
	COALTAR	-.24	.16	
	FRTLIZER	-.23		-.14
	BROADFAB	.25		
	STEVEDOR	.15		
	COPSMELT		.21	
	All significant industries	.68	.31	.36
	Positive R only	.52	.31	.33
Urinary Cancer	ANTHMIN	.35	-.19	
	GYPSUM	.23		
	BLOUSES	-.22		
	BEER	.20		
	SULFMIN	.16		
	PETROMIN		-.17	
	All significant industries	.46	-.25	.00
	Positive R only	.35		.00

being summarized in these statistics. However, for the other cancer types, there was greater statistical association between industry concentration and cancer mortality on a county basis when more than 90 percent of the workers were employed in their county of residence. It is anticipated that this variable will be used in the full-scale project to separate industrial from community exposure.

C. Problem of negative relationships

The initial 200 county stepwise regression of residual cancer mortality on the 30 industries produced two significant negative regressions: respiratory cancer residuals on dress manufacturing, and digestive cancer residuals on fertilizer manufacturing. Scatter plots of these two relationships are shown in Figures 2 and 3.

In Figure 2, all counties appearing in the lowest 10 percent of the 200 counties on respiratory residuals and in the top 10 percent on dress manufacturing are indicated. Only two of these, Lackawanna and Luzerne, Pennsylvania, accounted for the significant negative relationship; without them, there would be no significant relationship for the two variables. In Figure 3, Polk, Florida, similarly accounts for the negative relationship.

It is interesting to note that the actual cancer mortality rates in Lackawanna and Luzerne, Pennsylvania, which are coal mining as well

Figure 2

PLOT OF RESPIRATORY CANCER RESIDUAL BY CONCENTRATION OF WOMEN'S DRESS MANUFACTURE

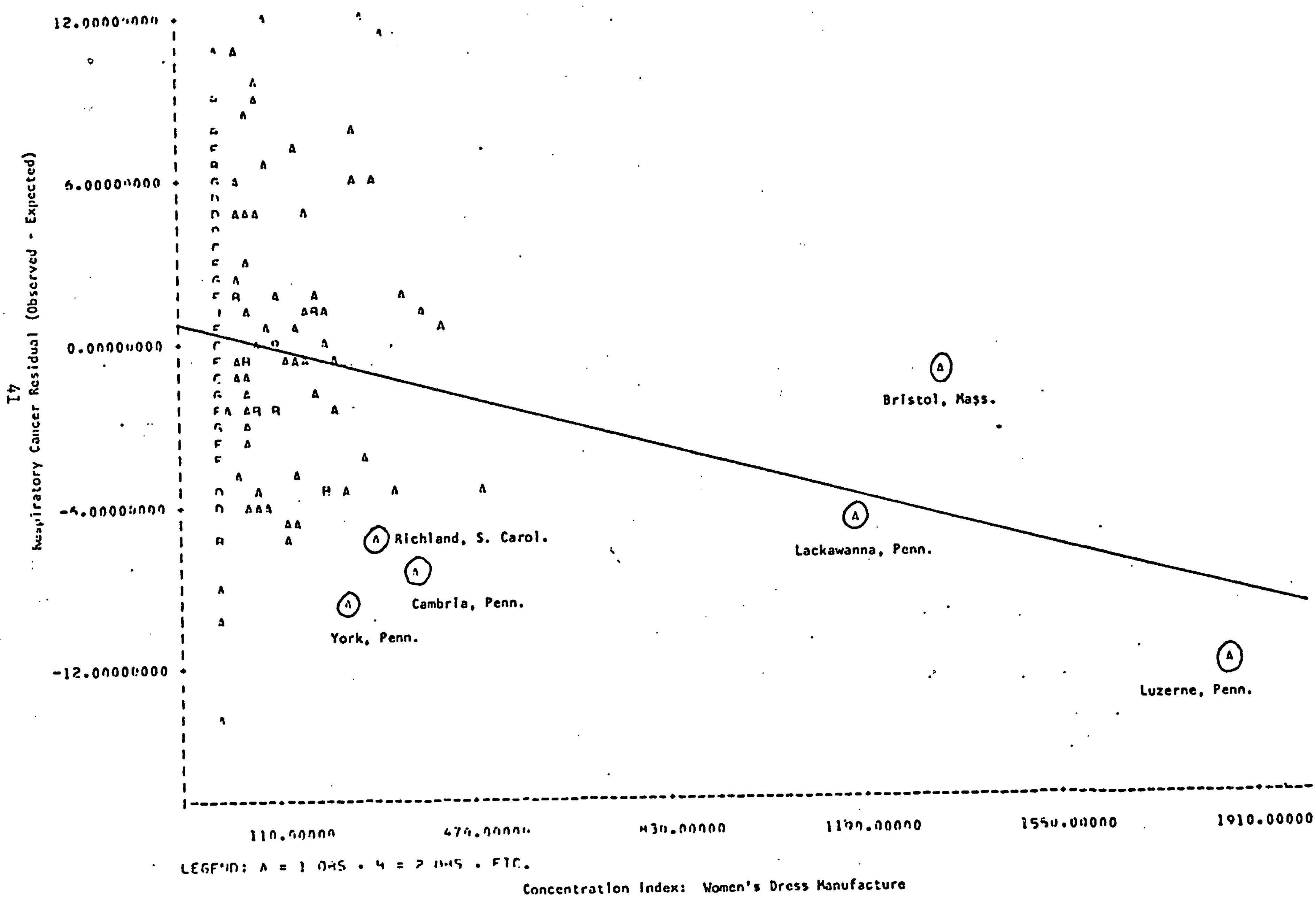
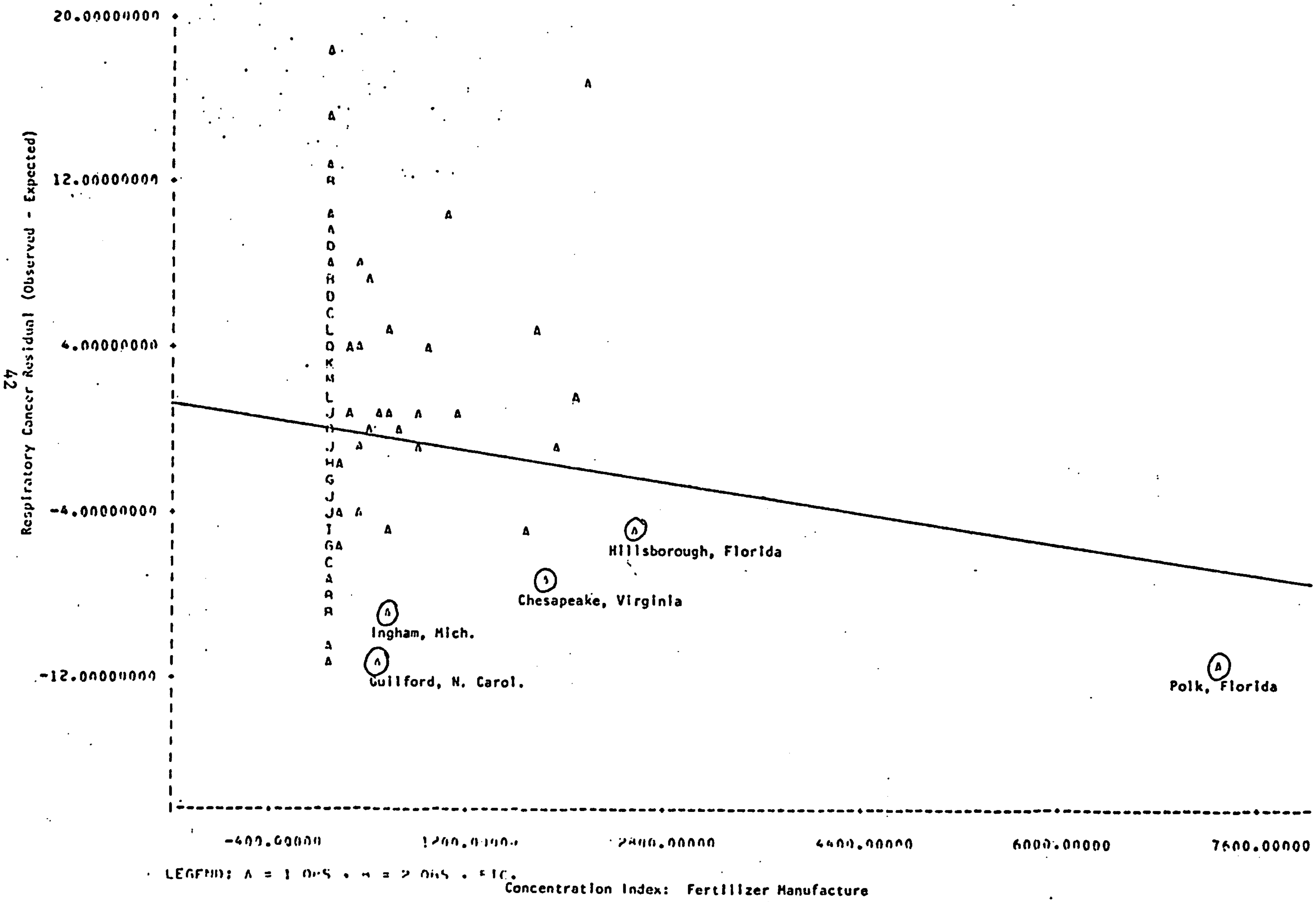


Figure 3

PLOT OF DIGESTIVE CANCER RESIDUAL BY CONCENTRATION OF FERTILIZER MANUFACTURE



as dress manufacturing counties, were so far below the predicted level. Coal mining also had a negative relationship with respiratory cancer, although it was not selected by the stepwise procedure; its negative relationship was also accounted for by these two counties. Competing causes could account for this.

Table 10 shows the percent of all mortalities attributed to certain causes for those counties in Figures 2 and 3; for the States in which they are located; for all metropolitan counties in the U.S.; and for the U.S. as a whole.

A high percentage of deaths in Lackawanna and Luzerne were in the residual disease category. This heterogeneous category contains, among other causes, anthracosilicosis (ICD 515.1) and anthracosis (a subcategory of ICD 516); many of the deaths in this category were probably due to these two diseases. Unfortunately a more specific category is not available by county. The residual respiratory category is close, but it is available only by state. Pennsylvania had a rate nearly three times that of the U.S. for this category although its rate for the broader disease category was only slightly higher than for the U.S.

These data support the hypothesis that the low respiratory cancer mortality rates in the counties which account for the negative regression between respiratory cancer and dress manufacturing, were due in part to a competing respiratory disease.

Table 10

Percent Distribution of Deaths by Cause, 1970:
U.S., Counties Contributing to Negative Stepwise Regression, and Their States

Area	Malignant Neoplasms 140-209	Diabetes Mellitus 250	Major Cardiovascular Diseases 390-448	Selected Respiratory Diseases 470-474 480-486 490-493	Selected Digestive and Urinary Diseases 571 580-584	Residual Diseases of 34 Causes	
						Respiratory 501-508 512 514-516 519	All
United States	17.2	2.0	52.5	4.9	2.1	0.6	7.2
Metropolitan Counties	18.0	2.0	51.7	4.8	2.4	---	7.3
Lackawanna, PA	16.2	2.6	56.8	3.2	2.0	---	10.8
Luzerne, PA	15.8	2.8	52.2	2.8	2.0	---	17.2
Cambria, PA	15.8	2.1	57.2	4.3	1.5	---	10.2
York, PA	17.8	2.2	54.5	4.7	1.3	---	8.7
Pennsylvania	17.3	2.2	54.9	4.0	1.9	1.4	8.0
Richland, SC	14.7	3.2	50.6	5.8	1.9	---	6.9
South Carolina	13.7	2.3	50.7	4.8	1.8	0.5	6.9
Polk, FL	17.1	1.5	46.7	6.1	1.4	---	7.5
Florida	18.3	1.7	50.7	5.2	2.0	0.5	7.2
Ingham, MI	17.5	3.4	48.5	5.4	1.4	---	7.3
Michigan	17.6	2.9	51.5	4.6	2.4	0.3	6.7
Chesapeake, VA	16.1	1.3	49.5	5.0	2.5	---	8.0
Virginia	16.2	1.5	51.0	5.2	1.9	0.6	7.8
Guilford, NC	16.5	2.1	48.4	5.1	2.3	---	8.5
North Carolina	14.6	1.9	50.9	5.2	1.9	0.5	7.6

Source: Vital Statistics of the United States, 1970. Respiratory Residual column: Table 1-27 (U.S. and States only); all other columns: Table 7-9

No such pattern was found for Polk, Florida. However, Polk had an unusual age distribution. It ranked 43rd from the highest median age of the 200 counties but 15th in percent aged 65 and over. As noted earlier, controlling for differences in age distribution by regression analysis is much less effective than using age-adjusted rates. Both negative relationships may disappear in the full-scale study.

These data illustrate the need to use age-adjusted rates and to include diseases other than cancer in the full-scale study.