RELATING SPATIAL DISTRIBUTIONS OF POLLUTANTS
TO HEALTH EFFECTS

Thomas W. Sager

TECHNICAL REPORT NO. 1
July 15, 1976

PREPARED UNDER THE AUSPICES OF
SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

9503527
There is no objection from the patent
point of view to the publication or
distribution of the document(s)
listed in this letter.

BROOKHAVEN PATENT GROUP

12/8/76 By

CHC

Prepared For
THE U.S. ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATION
UNDER CONTRACT NO. EY-76-S-02-2874.*000

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of 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. Reference herein to any specific commercial product, process, or service by 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. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

RELATING SPATIAL DISTRIBUTIONS OF POLLUTANTS TO HEALTH EFFECTS

BY

THOMAS W. SAGER
Stanford University

TECHNICAL REPORT NO. 1

JULY 15, 1976

STUDY ON STATISTICS AND ENVIRONMENTAL
FACTORS IN HEALTH

PREPARED UNDER SUPPORT TO SIMS FROM
ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATION (ERDA)
ROCKEFELLER FOUNDATION

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Energy Research and Development Administration, 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.

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

101

RELATING SPATIAL DISTRIBUTIONS OF POLLUTANTS TO HEALTH EFFECTS

Thomas W. Sager

Department of Statistics
Stanford University
Stanford, California

SUMMARY

A new and potentially useful statistical tool for epidemiology is introduced and some of its elementary properties are considered. The technique is a promising one for both data-analytic and inferential problems. Starting with the collection of isopleths of a spatially distributed explanatory variable (air pollution), the method produces a relationship between the explanatory variable(s) and the response variable (population-adjusted health effects) by accumulating the response within successively wider isopleths. Among other appealing features, the method has some of the flavor of regression, reduces the relationship between three three-dimensional distributions to a single easily interpreted two-dimensional graph, and effectively utilizes knowledge of the geographic location of data before discarding the location coordinates as nuisance parameters.

1. THE AIR POLLUTION PROBLEM

Air pollution has been a major national issue for several years. Public interest in the debate has been fueled by claims that air pollution causes adverse effects on human health, even at ambient atmospheric concentrations ordinarily found in our larger cities. Deleterious economic and esthetic consequences of air pollution are also alleged; but the health issue seems to be of primary public concern.

Research in community health effects of air pollution indicates that there may be adverse health effects (e.g., CHES (1974), Winkelstein (1972), Goldsmith and Hexter (1967), among an enormous literature). But interpretation of the evidence is fraught with difficulty because of incomplete understanding of physical processes involved and

statistical problems (missing data, inadequate data, spurious correlation, collinearity, spatial and temporal variability, confounding effects of all kinds, etc.). To a certain extent, insufficient statistical methodology contributes to some of the difficulty. Clinical evidence for adverse health effects is somewhat stronger, particularly under exposure concentrations considerably above those found in the air of our cities. Suspected polluters have sponsored their own studies which seek to minimize the seriousness of the health effects at ordinary pollution levels. The current state of knowledge could probably be summarized fairly by the statement that there seem to be adverse health effects - chronic and acute - at city-typical concentrations, but the seriousness of these effects remains uncertain.

The following diagram summarizes the suspected causal relationships:

In the anthropogenic category, mobile and stationary sources emit primary pollutants directly into the atmosphere. These pollutants are principally carbon monoxide, nitrogen oxides, hydrogen sulfide, sulfur oxides, hydrocarbons, particulate matter, and heavy metals such as lead. Photochemical reactions in the atmosphere produce secondary pollutants such as ozone and PAN (peroxyacetylnitrate). The concentrations of primary and secondary pollutants as fractions of the atmosphere (expressed by weight as micrograms per cubic centimeter, $\mu\text{g}/\text{cm}^3$, or by volume as parts per million, ppm) determine ambient air quality. But the effect of emissions on air quality is strongly mediated by weather (inversion level height, mixing, wind speed and direction, air pres-

sure, precipitation, etc.). Turning to the arrow on the right, we find human exposure to polluted air resulting in a complex, interrelated set of effects - chiefly on the lungs, eyes, and skin, but indirectly on the heart and other organs, and on general bodily health. It is thought that headache, eye irritation, asthma, emphysema, hay fever, angina, bronchitis, obstructive lung disease, mild and severe heart conditions, and lung cancer are among the conditions which can be caused or exacerbated by air pollution.

From the statistical point of view, determining the relationship between health effects and ambient air quality may be broadly construed as a regression problem with "health effects" as dependent variable(s) and "ambient air quality" as independent variable(s). Our analysis may be exploratory if we approach the data without a model and seek to discover one. Or it may be confirmatory if we have a model which we wish to test by means of the data. The approach which we will offer here may be either.

Since the nature of available data strongly affects the analytical methodology which may be applied, we comment on the data before considering specifics of our technique. In the simplest form of the problem, we wish to analyze the relationship between a health effect Y and a pollutant X, suitably defined, with all other variables held constant.¹ Measurements of the pollutant X are taken at air monitoring stations located at fixed points in an air pollution control district. Because of operation and instrumentation expense, the number

¹ In practice, other variables cannot be held constant and a useful expedient commonly adopted is to ignore caeteris paribus and impose a caveat on the interpretation of the results. Alternatively, variants on the idea of partial correlation may be employed.

of monitoring stations is usually small. For example, at the time of this writing, there were about 30 monitoring stations in the Bay Area Air Pollution Control District. Measurements of many pollutants are taken almost continuously in time. Thus the adequacy of any reconstruction of a region's temporal-spatial distribution of a pollutant is limited by the geographical representativeness of the region's air monitoring stations. Reconstructions of the spatial distributions of some pollutants (e.g., sulfur dioxide) may be relatively insensitive to the location of monitoring stations because of relatively stable spatial distributions and knowledge of physical processes involved. For such pollutants, diffusion models of the sort developed at I.B.M. (Shir and Shieh (1974)) offer a promising approach to estimating exposure levels. For more variable pollutants (e.g., carbon monoxide) the location of monitoring stations may be crucial (Ott (1971)). Although mobile monitoring could alleviate the problem of sparse spatial measurements, its use is curtailed by the relative immobility of most monitoring equipment, the possibility of contamination and nonrepresentative levels on the highway, and inconvenience. A more promising approach to spatial measurement seems to be remote detection by laser (e.g., the work of Byer (Kildal and Byer (1971)) at Stanford, which offers the possibility of real-time three-dimensional TV displays of pollution readings).

Health effects data are of an entirely different nature from pollution data. Whereas the values of pollution readings may occur on a relatively continuous scale, health effects data are often dichotomous: a person is classified either as asthmatic or as non-asthmatic. Whereas pollution data are collected at fixed points in space and virtually

continuously in time, health effects may occur wherever people live and at any time. Usually, however, the health data are collected at discrete points such as hospitals and doctors' offices and reported for discrete times such as the week of June 1 - 7. Some health effects may be measured on a continuous scale, e.g., ventilatory lung capacity or blood lead concentration. Although the techniques of this paper may be adapted for continuous measurements, it is dichotomous health variables which concern us here.

Regression methods are commonly used in epidemiology to analyze problems like the above. Characteristically, a large, geographical diversified population of subjects is partitioned geographically into subgroups. The subgroups may be very large. Air quality data - often gathered from a single air monitoring station chosen to represent the subgroup - supply estimated exposure levels. Cumulative health effects data for each subgroup are gathered. A linear regression of effects (or their logarithms) on exposure levels (or their logarithms) and other variables of interest yields the estimated relationship. The CHES (1974), Winkelstein (1972), Goldsmith and Hexter (1967) studies, for example, are of this flavor. The resulting regression equation removes the spatial and temporal coordinates of the variables,² but does so in a rather crude way. The exposure levels become average levels, in some sense, over time and over space, which are then assigned to the entire subgroup associated with the measurements. Satisfactorily inferring the health response for pollutant levels between any of the observed pollution estimates then depends not only on the adequacy of the data at observed measurements but also very strongly on the assumption,

² Properly so, for these coordinates are nuisance parameters.

of linear structure to the response.

The method to be presented in this paper, I think, makes better use of spatial coordinates, avoids linearity assumptions, and provides a good two-dimensional visual technique for comparing spatial distributions. Additionally, it retains the spirit and flavor of regression, without the special methods of the latter. It will certainly not replace regression as a tool, but it may prove a helpful alternative and/or companion tool to regression. Finally, it will not answer the question of whether air pollution causes poor health, but it can be a useful exploratory and confirmatory statistical tool in epidemiological work.

2. THE TECHNIQUE

We consider a geographical region A coordinatized arbitrarily by perpendicular axes z_1 and z_2 . For now, we ignore vertical distances and suppose that a plane is a reasonable representation of the region A . Locations in A may be specified by a pair (z_1, z_2) . We also suppose that air quality monitoring stations are sufficiently numerous in A to provide a reasonable estimate over the region A of the concentration distribution of a pollutant X of interest (possibly a pollutant index). Since air quality measurements occur in the temporal dimension, as well as in the spatial, we suppose that the measurements all occur in the same time interval and that the time parameter is eliminated by estimating the average concentration distribution of the pollutant, or the maximum concentration distribution of the pollutant (or some other functional) over A for the given time span.

Thus we suppose that we have an estimate $f(\underline{z}) = f(z_1, z_2)$ of the

(functional) ambient atmospheric concentration of the pollutant X for the given time span for each point $\underline{z}$ in A . Obtaining a reasonable estimate f is no small matter. We shall say more about this in section 4. For the moment, we shall accept f as given. Moreover, we shall assume that A and f are sufficiently regular that f may be integrated, finitely, over A .

Mathematically, f may be thought of as the graph of a function over a region in the plane. Pictorially, f is often displayed approximately as a number of isopleth curves drawn on a map. Figure 1 provides an example.

On a geographical terrain map, isopleths connect points of equal height above sea level. Here, by the p-quantile isopleth, we shall mean the set of all $\underline{z}$ in A where $f(\underline{z}) = h$ for some $h = h(p)$ satisfying

$$\int_{\{\underline{z} \in A; f(\underline{z}) \geq h\}} f(\underline{z}) \, d\underline{z} = p \int_A f(\underline{z}) \, d\underline{z}.$$

By p-modal region, we mean the set of all $\underline{z}$ in A such that $f(\underline{z}) \geq h(p)$. Thus the p-quantile isopleth is the boundary³ of the most polluted region of A containing 100p% of the total pollution of A ; and the p-modal region is simply that most polluted region, together with its boundary⁴.

³ We do not mean "boundary" in the topological sense. Cases in which an isopleth is not a topological boundary require the discontinuity of f or a "terrace" (see below) in the graph of f . In most practical examples, these cases will not arise.

⁴ Here we commit an inaccuracy: since f is merely the estimated concentration distribution, the p-quantile isopleth is more precisely the estimated boundary of the most polluted region of A containing 100p% of the total pollution of A . However, the awkwardness of this phrase induces me to a certain looseness of language in the sequel. My apologies.

Necessary and sufficient conditions for the existence of all p-modal regions, $0 < p < 1$.

It is clear that a p-modal region need not exist. For example, if the graph of f contains a "terrace" (i.e., a region of positive area over which f assumes a positive constant value) then the integral of f on the terrace is positive. And this limits the number of possible p-modal regions.

Let

$$F(h) = \int_{\{\underline{z} \in A; f(\underline{z}) \geq h\}} f(\underline{z}) \, d\underline{z} \quad \text{for all } h > 0.$$

A necessary and sufficient condition for the existence of all p-modal regions, $0 < p < 1$, is the continuity of $F(\cdot)$. For, observe that F is nonincreasing in h and $F(\infty) = 0$. If F is also continuous, then F assumes all values between $F(0)$ and $F(\infty)$. Hence, given p , $0 \leq p \leq 1$, there exists $h \geq 0$ such that $F(h) = p \cdot F(0)$. Conversely, if F is not continuous then there is an h such that $F(h^+) < F(h)$ or $F(h) < F(h^-)$. Let $q = F(h)/F(0)$ and $q^+ = F(h^+)/F(0)$, $q^- = F(h^-)/F(0)$. Hence $q^+ < q$ or $q < q^-$. By the monotonicity of F , if either of these conditions holds, then there are no p-modal regions for $q^+ < p < q$ (or $q < p < q^-$).

It follows additionally that a necessary and sufficient condition for F to be continuous is that the area of all isopleths be zero. This will be the case if isopleths are drawn (as they normally are) as a connected curved line.

Corresponding to each $h \geq 0$, we associate the value p , $0 \leq p \leq 1$, defined by $F(h) = p$. We assume now that F is continuous and we associate with each p , $0 \leq p \leq 1$, the value $h(p) = \inf\{h; F(h) = p\}$. Normally, for each p , $h(p)$ will be unique; but if f is discontinuous, this can not be guaranteed.

Consider the collection of p -modal regions $M(p) = \{z; f(z) \geq h(p)\}$, $0 \leq p \leq 1$. $M(0)$ is the mode of f . As p increases from 0 to 1, $M(p)$ spreads out from the mode, creating successively, continuously nested sets, each of which is bounded by an isopleth and each contains the most severe 100p% of the pollution of the region A.

If there are adverse health effects to pollution, then we would expect the effects (after adjusting for population density) to be most concentrated where the pollution is most concentrated. We would expect more than 100p% of the population-adjusted health effects to occur in $M(p)$ for most $p < c$ for some $0 < c < 1$. For p close to 1, we would expect complements of $M(p)$ to contain less than 100(1-p)% of the population-adjusted health effects. The occurrence of such phenomena, of course, would not tell us that pollution is responsible for the effects, but it would be suggestive⁵.

Since we are considering discrete community health effects (either present or absent in an individual), the number of occurrences of such effects may be tallied within each modal interval. We then define

$H(p)$ = relative proportion of all the health effects
observed in A which occur in $M(p)$.

⁵ We shall indicate later how other explanatory variables may be examined for confounding or contributory effects.

To adjust for population density, let

$$P(p) = \text{proportion of the total population} \\ \text{of A which lives in } M(p).$$

Ideally, available health data should include the patient's place of residence (or place of work, if the health effect is associated with an occupational hazard) but might be tallied only by hospital, doctor's office, borough, or administrative district. Similarly, population figures may not be available for smaller units than census tracts. It is assumed here that none of these geographic data collection units is of significant size in relation to the total area of A. In addition to small size with respect to A, it is also desirable that the geographic data units be small in relation to the spatial variability of the pollutant, so that the estimated pollutant concentration f is roughly constant on the unit. Too large geographic units will destroy the applicability of our technique through averaging intra-unit variability. But if these units are nevertheless rather large but not excessively so, it may be possible to obtain an interpolated estimate of the health effects density in the same way that the estimate $f(\underline{z})$ is obtained. Finally, when $M(p)$ only partially includes a geographic data unit, we suggest assigning the observed health effects in the unit to $M(p)$ and to the complement of $M(p)$ in proportion to the area of the unit contained in $M(p)$ and in its complement:

The graphs of $H(p)$ and $P(p)$ are nondecreasing step functions but are approximately continuous if the number of health cases and reporting units is large. We have drawn them as continuous in the examples (figures 2 and 3).

In figure 2, the 45° line is the graph of $F(M(p))$. For the hypothetical region A, we note that population rises more slowly than pollution at first because many of the worst polluting industries are located in industrial parks, away from heavily populated regions. Population, however, then rises rapidly because workers find it convenient to live not far from the industries where they work. Thus, about 40% of the population lives within the 25% most-polluted area of region A. Population thins out in the suburbs, away from the most polluted regions, and residual low levels of pollution decrease more slowly there than population.

In spite of the segregation of many industries in industrial parks, health effects for the hypothetical region A rise more rapidly than population or pollution, for many workers report their illnesses at work to company doctors. In the heavily populated regions near the industrial parks, health effects rise far above the levels expected in population of such density. Thus, 75% of the health effects occur in the 50% worst-polluted region of A where 60% of the people live. Fully $25\% \left(\frac{75-60}{60} \times 100\% \right)$ more health effects occur in this area than would be expected in the general population. Thereafter, $H(p)$ grows very slowly as there are few cases in the suburbs.

Figure 2 was drawn under the assumption that pollution does contribute to ill health. In figure 3, there is no such effect. In hypothetical region B, pollution concentrations are generally lower and more uniform than in A. Because the city is more compact than A and there are no industrial parks, population rises more rapidly than pollution. But health effects deviate little from the population curve, failing to support an alleged impact of pollution on health in community B.

3. DISCUSSION

The basic idea of our technique is the examination of the curve $H(p)$ with respect to $P(p)$ (e.g., figures 2 and 3) and of their derivatives $H'(p)$ and $P'(p)$ (e.g., figure 4). By accumulating health effects within modal regions of the air pollution distribution, we are able to relate severity of effects to intensity of the pollution insult. The derivative $H'(p)$ then expresses the average severity of the effect at the isopleth boundary of $M(p)$, where pollution concentration is constant. Then $H'(p)$ - divided by $P'(p)$ - expresses the basic "regression" relationship between health and air pollution. To adjust health effects for population density, we also accumulate population within the modal regions of the pollution distribution.

The method outlined above has a great deal of flexibility. For example, one need not study health effects for the entire community. Certain subgroups of people may be "susceptibles". It is thought, for example, that the elderly, the very young, asthmatics, and those with chronic heart and lung conditions suffer relatively more from the effects of air pollution than does the general populace. If it is decided to set air quality standards to protect such groups, then their responses to pollution insults must be determined. To adapt our method to study susceptibles, let $H_s(p)$ count the relative proportion of health effects occurring in $M(p)$ among a well-defined class of susceptibles living in region A. The population adjustment $P_s(p)$ is the proportion of susceptibles living in $M(p)$. Data on the spatial distribution of susceptibles might be rather difficult to develop. But if it is rea-

sonable to assume that susceptibles choose places of residence in approximately the same way as the general populace and if the total population of susceptibles is relatively large⁶, then the susceptible-population function $P_s(p)$ is easily obtained: it is the same as the total population function $P(p)$.

In practice, p will almost certainly be a monotone continuous function of h . Sufficient conditions for this property are zero area of all isopleths and continuity of $f(\cdot)$, for example. When this is the case, it is possible to transform figures 2,3,4 so that the horizontal axis becomes pollution concentration. Let $c = h(p)$ denote the common pollutant concentration value $f(\underline{z})$ for $\underline{z}$ on the p -quantile isopleth. Then $p = h^{-1}(c)$ and $F(M(p)) = F(M(h^{-1}(c)))$, $H(p) = H(h^{-1}(c))$, $P(p) = P(h^{-1}(c))$, $H'(p) = H'(h^{-1}(c)) \cdot [h'(h^{-1}(c))]^{-1}$, $P'(p) = P'(h^{-1}(c)) \cdot [h'(h^{-1}(c))]^{-1}$. This transformation directly displays the way population-adjusted health effects vary with pollution levels. Figure 2 is then transformed to something like figure 5. In this figure we see concentrations decreasing rapidly from high levels near emission sources to steadier levels in the nearby regions. Figure 5 supplies a more detailed picture of the phenomenon near highly polluted regions than does figure 2, but figure 2 provides a better picture for the overall community.

We can make a number of observations about graphs such as figure 2. First, each curve $H(p)$ and $P(p)$ originates at $(0,0)$ and terminates at $(1,1)$. But to support a claim that air pollution contributes to community health problems, the curve $H(p)$ must lie significantly above

⁶ Large enough for standard statistical tests of the equality of two distributions to have substantial power.

$P(p)$ for at least some p , particularly for relatively small p . The vertical difference $H(p) - P(p)$ is the excess proportion of health effects in $M(p)$ beyond expected. This difference should not be integrated on p to obtain a measure of average excess effects beyond expected because H and P are cumulative functions. An initial effect near $p = 0$ may result in a large excess which may persist even though the effect may not exist for most p . As cumulative functions, H and P may not be very stable near $p = 0$ and may be too stable near $p = 1$. Thus, large percentage differences between H and P for p near 0 must be interpreted cautiously and likewise for small differences near $p = 1$. However, analysis of the average (or cumulative) pollution-health problem within a particular p_0 -modal region may be accomplished by simply reading off the values of $H(p_0)$ and $P(p_0)$. An unacceptably high value for $H(p_0)$ in relation to $P(p_0)$ should not be interpreted as implying that the entire p_0 -modal region needs regulatory attention. Although high, H may be growing very slowly with respect to P at p_0 . This would indicate that in the immediate environs of the p_0 -quantile isopleth (the boundary of $M(p_0)$) the population-adjusted density of health effects is less than for the entire region A .

The discussion in the preceding paragraph suggests that a further very important relationship is that between the derivatives $H'(p)$ and $P'(p)$. It is the rate of change of H with respect to P at p_0 which indicates the severity of the pollution-health problem in the vicinity of the p_0 -quantile isopleth, $I(p_0)$. If $H'(p_0) < P'(p_0)$, the vicinity of $I(p_0)$ has lower population-adjusted health effects than average for the region A . If $H'(p_0) > P'(p_0)$, then $I(p_0)$ has a higher density of effects. And if $H'(p_0) = P'(p_0)$, then $I(p_0)$ has the same density of

effects as the region A. The subregion with the most severe problems is identified by the p which maximizes $H'(p)/P'(p)$. On the other hand, the subregion with the largest excess of $H(p)$ over $P(p)$ is one for which $H'(p) = P'(p)$. It is important not to overlook the implications of this difference for public policy.

If we believe that high pollution concentrations have etiological implications for health - the higher the concentration, caeteris paribus, the greater the effect - then we expect to find $H'(p)/P'(p)$ to be a monotone decreasing function of p . If the data, on the whole, seem to suggest monotonicity (plus noise), we may attempt to recover the monotonic trend by calculating the isotonic (actually, antitonic) regression of $H'(p)/P'(p)$. This is equivalent to evaluating the conditional expectation of $H'(p)/P'(p)$ given the sigma-lattice of intervals of the form $[0, p)$ or $[0, p]$, $0 \leq p \leq 1$. (Cf. Barlow, et al(1972))

Inspection of graphs such as figures 2,4,5 conveys at a glance an impression of the total health-pollution problem of a region A. Considerable data-compression is achieved in this way by eliminating the spatial dimensions of the three distributions of pollution, health effects, and population. But not much information is lost. It is not the relationship of $H(p)$ to $F(M(p))$ (the 45° line in figure 2) but the relationship of $H(p)$ to $P(p)$ which is important. And even more important is the relationship between $H'(p)$ and $P'(p)$. Each of these graphical quantities has a readily interpretable, intuitive physical meaning.

So far, we have considered only one explanatory variable. In practice, such a simple model is unlikely to provide an adequate representation of a complex phenomenon. In principle there is no reason why we may not treat other potential explanatory variables as we have air

pollution. We require an (estimated) concentration function $g(z_1, z_2)$ of the variable, from which we obtain isopleths, modal intervals, health effects curve H_g , and population adjustment P_g . If g-data are collected at fixed points in A, then $g(\underline{z})$ may be estimated like $f(\underline{z})$; but if g-data are dichotomous and may occur anywhere, then other methods are required (see section 4). We mention two examples among the host which have appeared in the literature: $g(\underline{z})$ = estimated mean daily maximum temperature at $\underline{z}$ for the period of study, or $g(\underline{z})$ = socioeconomic status of the family residing at $\underline{z}$ (e.g., Winkelstein (1972)).

To reveal the combined effect of the variables f and g (in a kind of "multivariate regression"), we may consider a function of f and g, say $\alpha(f, g)$, which maps the space of f and g into the real line. Through its dependence on f and g, α becomes a spatially distributed variable which has isopleths and modal regions of its own. If the graph of H_α rises proportionally more over P_α than does H over P, we say that α provides more explanatory power than f alone. High values of α are relatively more serious than high values of f. By considering functions α within a special class, e.g., $\alpha = \theta f + (1-\theta)g$, it should be possible to find optimal α for a criterion formalizing the idea of a rapid rise in H over P. This will not be done here. The extension of this idea to k explanatory variables is clear. We point out that if f and g are measurements for different pollutants, then the procedure just described provides a method for obtaining a health-based pollution index.

The effects of f and g will be difficult to separate if their spatial distributions are close, in a certain sense. Define

$$G_f(p) = \frac{\int_{M(p)} g(\underline{z}) d\underline{z}}{\int_A g(\underline{z}) d\underline{z}}$$

If the graph of G_f is close to the 45° line, and more importantly, if G_f' is close to 1, then clearly f and g are highly associated. But not necessarily identical, even in the extreme case $G_f' \equiv 1$.

A necessary and sufficient condition for $G_f'(p) = 1$ for all p , $0 < p < 1$, is that $g^*(I(p)) = c \cdot f(I(p))$ for all $0 < p < 1$, for some constant c , where g^* is the average value of g on the f -isopleth $I(p)$.

For, suppose $G_f' \equiv 1$. Choose p . Then for small positive ϵ , and $c_1 = \int_A g(\underline{z}) d\underline{z}$, we have

$$\lim_{\epsilon \rightarrow 0} \frac{1}{c_1 \cdot \epsilon} \left[\int_{M(p+\epsilon)} g(\underline{z}) d\underline{z} - \int_{M(p)} g(\underline{z}) d\underline{z} \right] = 1$$

Evaluate the integral by integrating g first along the isopleth $I(q)$ and then over q , $p < q \leq p+\epsilon$. We obtain

$$1 = \lim_{\epsilon \rightarrow 0} \frac{1}{c_1 \cdot \epsilon} \int_p^{p+\epsilon} \int_{I(q)} g(\underline{z}) d\underline{z} dq =$$

$$\lim_{\epsilon \rightarrow 0} \frac{1}{c_1 \cdot \epsilon} \int_p^{p+\epsilon} g^*(I(q)) \cdot \text{length } I(q) dz = g^*(I(p)) \cdot \text{length } I(p) \cdot c_1^{-1}$$

Similarly, we find $1 = f(I(p)) \cdot \text{length } I(p) \cdot c_2^{-1}$ where $c_2 = \int_A f(\underline{z}) d\underline{z}$.

Hence the condition holds with $c = c_1/c_2$.

Conversely, if the condition holds, then

$$\int_{M(p)} g(\underline{z}) d\underline{z} = c \cdot \int_{M(p)} f(\underline{z}) d\underline{z} \text{ for all } p. \text{ Hence } c = c_1/c_2 \text{ and}$$

$$c_1^{-1} \int_{M(p)} g(\underline{z}) d\underline{z} = p.$$

In the above we have implicitly assumed the finiteness and continuity of length $I(q)$.

As a corollary, we obtain an explicit formula for

$$h(p) : h(p) = \int_A f(z) dz / \text{length } I(p).$$

We remark that the condition $G_f' \equiv 1$ is not sufficient to establish that $g(z) = a \cdot f(z) + b$ for all z .

We note that g^* is the conditional expectation of g with respect to the sigma-lattice of p -modal regions of f . If $G_f' \equiv 1$, then g^* is essentially f . Hence, from the standpoint of our method, g contains all of the information which f contains about other variables such as health. We shall say f is a reduction of g in this case. In the symmetric case, when $F_g' \equiv 1$, we say g is a reduction of f .

In general, if the graph of population-adjusted health effects H over the modal regions of f displays the same pattern as do health effects over the modal regions of g , we may say the effects of f and g are confounded. But if the graph of H over the modal (g) regions is more like the graph of P than the graph of H over the modal (f) regions is like the graph of P then we may say f is primary in explaining health effects and g is secondary. Appropriate measures of primacy may be developed.

In the spirit of residual analysis, we may further examine the curves $H(p)$ and $P(p)$. Suppose g is a second explanatory variable whose pure effect on health - with the effect of f removed - we wish to examine. We shall merely indicate the technique here; it will be developed further in a subsequent paper. Now, health effects have a spatial distribution of their own, say $d(z)$ (see section 4 for estimation procedures). Let d^* , g^* denote the conditional expectations of d and g

given the sigma-lattice of modal intervals of f . Then $d(\underline{z}) - d^*(\underline{z})$ and $g(\underline{z}) - g^*(\underline{z})$ are the residual distributions of d and g adjusted for f^7 . The residual $d - d^*$ and population may be accumulated over the modal regions of $g - g^*$ to yield new H and P functions to be checked for (residual) explanatory power.

A Test. One of the issues we have raised is whether $H(p) \equiv P(p)$. Since both H and P are legitimate distribution functions, this hypothesis may be tested by the one - or two-sided Kolmogorov-Smirnov statistic. In a sense, however, there is no test to be applied, for there is no sample. Both H and P are presumed to be based on complete data for the region A . Hence the sample is the statistical population. In particular, census data are likely to supply a very accurate estimate of P . Depending on the health effect studied, H may be much less accurate. If P is regarded as known and H based on a sample, then the one-sample Kolmogorov-Smirnov statistic may be employed. In any event, the known probability distribution of the Kolmogorov-Smirnov statistic can provide the probability that the difference between H and P arose by chance.

The test of $H(p_0) = P(p_0)$, for given p_0 , may be based on a test of the equality of two binomial proportions, or of the equality of a binomial proportion to a certain constant.

4. ESTIMATION OF SPATIAL DISTRIBUTIONS

In the preceding development, we assumed that the estimated spa-

⁷ They are not distributions in the usual sense, for they may be negative for certain $\underline{z}$. But our method should extend to this case, with careful definitions of isopleth and modal region.

tial distribution $f(\underline{z})$ was obtained for every $\underline{z}$ in A. In practice, this may not be necessary. A few well-chosen isopleths may suffice for most purposes.

Probably the most common technique for drawing isopleths of spatially distributed variables is the hand-eye technique. An experienced map-maker draws them by eye. (E.g., pollution maps published by the Bay Area Air Pollution Control District (figure 1), the Buffalo study (Winkelstein (1972)), among many). One failing of hand-drawn maps is a tendency to make the isopleths too smooth.

Much of the serious statistical effort to automate cartography is found in the earth sciences, geology and geography. A number of computer programs can now draw isopleths (e.g., Behar (1972), and others discussed in Davis and Cullagh (1975).) If hand-drawn maps sometimes leave out features which are present, some computer programs may produce features which are not really there. A truly satisfactory algorithm is still in the future. Theoretical aspects of spatial analysis are studied in widely scattered sources. Davis and Cullagh (1975) collect a number of recent papers with extensive bibliographies.

The problem of estimating the spatial distribution of f is quite different from estimating a probability density function, for the sample points are largely fixed. Most explanatory variables will be of this kind. Most health-effects variables, on the other hand, give rise to data more like a random sample of points from some distribution on the plane. However, with the procedure developed here, we are not required to estimate the distributions of such variables. We merely tally them within the isopleths of the pollution (or independent) variable(s). Nevertheless, there may be occasion when an

explanatory variable g is of this second type. Then the extensive literature on probability density estimation - particularly, multivariate - may be drawn upon. Cacoullos (1966) presents a multivariate kernel method inspired by spectral analysis. An early paper on spline estimation with examples is Boneva, Kendall, and Stefanov (1971). Sager (1975) uses an iterative method to estimate isopleths simultaneously. A judicious selection from such methods is likely to prove satisfactory.

REFERENCES

- Barlow, R.E.; Bartholomew, D.J.; Bremner, J.M.; Brunk, H.D. (1972). Statistical Inference under Order Restrictions, Wiley, New York.
- Behar, J.V. Application of Computer Simulation Techniques to Problems in Air Pollution (1972). Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume VI, University of California Press, Berkeley, pp. 29-69.
- Boneva, L.I.; Kendall, D.; Stefanov (1971). Spline transformations. Jour. Royal Stat. Soc. B, 33, p.1.
- Cacoullos, T. (1966). Estimation of a multivariate density. Annals of the Institute of Statistical Mathematics, 18, pp. 178-189.
- CHESS (1974): Health Consequences of Sulfur Oxides: A Report from CHESS (1974). U.S. Environmental Protection Agency publication EPA-650/1-74-004.
- Davis, J.C. and Cullagh, M.J. [Editors] (1975). Display and Analysis of Spatial Data, Wiley, New York.
- Goldsmith, J.R. and Hexter, A.C. (1967). Respiratory exposure to lead: epidemiological and experimental dose-response relationship. Science, 158, p. 138.
- Kildal, H. and Byer, R.L. (1971). Comparison of laser methods for the remote detection of atmosphere pollutants. Proceedings of the IEEE, 59, pp. 1644-1663.
- Ott, W.R. (1971). An urban survey technique for measuring the spatial variation of carbon monoxide concentrations in cities. Stanford University Department of Civil Engineering. (Ph.D. dissertation).
- Sager, T.W. (1975). An iterative method for estimating a multivariate mode and isopleth. Stanford University Department of Statistics, Technical Report No. 105. (To appear in Jour. Am. Stat. Assoc.)
- Shir, C.C. and Shieh, L.J. (1974). A generalized urban air pollution model and its application to the study of SO_2 distributions in the St. Louis metropolitan area. Journal of Applied Meteorology, 13, pp. 185-204.
- Winkelstein, W. (1972). Utility or futility of ordinary mortality statistics in the study of air pollution effects. Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume VI, University of California Press, Berkeley, pp. 539-554.

Figure 2

Graphs of $H(p)$, $P(p)$, $F(M(p))$
for a hypothetical large, heavily
populated and polluted region A.

Figure 3

Graphs of $H(p)$, $P(p)$, $F(M(p))$
for a hypothetical large,
moderately populated, lightly
polluted region B.

Figure 4

Graph of $H'(p)$ and $P'(p)$
(from figure 2)

Figure 5

Figure 2 with horizontal
axis transformed by $c = h^{-1}(p)$