A NEW APPROACH TO REGRESSION IN CERTAIN TIME/SPACE SERIES PROBLEMS

by

THOMAS W. SAGER

Stanford University

TECHNICAL REPORT NO. 11

October 14, 1977

STUDY ON STATISTICS AND ENVIRONMENTAL FACTORS IN HEALTH

PREPARED UNDER SUPPORT TO SIMS FROM

ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATION (ERDA)
ROCKEFELLER FOUNDATION
SLOAN FOUNDATION
ENVIRONMENTAL PROTECTION AGENCY (EPA)
NATIONAL SCIENCE FOUNDATION (NSF)

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

DISCLAIMER

A New Approach to Regression in Certain Time/Space Series Problems

Thomas W. Sager*

Abstract. This paper introduces a new method for estimating a dose-response relationship from spatially averaged time series of air pollution and health data. Because time is perceived as a nuisance parameter to be eliminated, least-squares regression and traditional time series methodology (e.g., spectral analysis, Box-Jenkins methods) are rejected in favor of a nonparametric estimation procedure based on observing health effects in times of nearly equal pollution. The method requires estimating the ratio of two density functions and avoids problems of aggregation, linearity, and normality. Refinements are suggested in section 3. In spite of the formal tests described in section 5, the procedure seems most useful, at present, as a data analytic and data display device rather than as an inferential tool.

* Statistics Department, Stanford University, Stanford, California 94305.
This research was prepared under support to SIMS from Energy Research and Development Administration (ERDA), Rockefeller Foundation, Sloan Foundation, Environmental Protection Agency (EPA), and National Science Foundation (NSF).

Draft of a paper given before SIMS Conference on Time Series and Ecology, Alta, Utah, June 26-July 1, 1977.

1. Introduction. Clinical work and epidemiology are the two major tools for assessing the impact of air pollution on human health. Although complementary in character, neither has provided entirely satisfactory answers to the questions that scientists and regulatory agencies have asked. Laboratory work offers the possibility of control over relevant variables but the sample sizes are small and, because laboratory conditions do not resemble those in our cities, extrapolation is a problem. Epidemiology has enormous sample sizes and real-life conditions but lacks control over exposure levels, and measurement of both dose and response is laden with error.

In this paper we focus on epidemiology. Inadequacies in the data base, the sheer complexity of interactions among relevant variables, and other essentially nonmethodological issues all contribute to the problem of inferring the dose-response relationship between pollution and health. But at least part of the difficulty must be laid on the doorstep of methodology: traditional statistical models for regression inadequately cope with the time/space series character of pollution and health variables.

To see how this difficulty arises, consider the nature of the data. Pollution data are a sample of a pollutant concentration function $f(x_1,x_2,t)$, where x_1,x_2 locate the sampling point in space [the third spatial dimension is omitted here, as there has been little vertical profiling of pollution at this time] and t fixes the time of the sample. To permit some flexibility, we may allow f to measure the maximum concentration at (x_1,x_2) over some time period ending at t, or perhaps an

accumulated exposure until t--rather than just an instantaneous measure. The coordinatization is arbitrary. Typically, the exposures f are measured at a few fixed air monitoring stations in the area of interest and may be formally represented as a matrix $\{(f_{ij}, u_{i1}, u_{i2}, t_{ij})$; $i=1, \ldots, m; j=1, \ldots, n_i\}$ where f_{ij} is the measured pollutant concentration (taken to be univariate--possibly an index--in this paper) at station i located at (u_{i1}, u_{i2}) at time t_{ij}. Unlike pollution readings, health effects may occur throughout the affected region. Health data often arise as a sample <u>from</u> a health effects density function $h(x_1, x_2, t)$ and may be represented as a vector $\{(x_{i1}, x_{i2}, t_i)$; $i=1, \ldots, n\}$ which locate unitary occurrences of health effects in space and time. Reconstruction of $f(\cdot,\cdot,\cdot)$ from the data requires spatial and temporal interpolation and smoothing between sampling points and times, whereas $h(\cdot,\cdot,\cdot)$ must be estimated from the data by means of a histogram or other density estimation technique.

Pollution and health data are thus qualitatively different kinds of information, and they do not occur naturally paired in either space or time. Thus, without considerable spatial and temporal aggregation, a classical regression of health-effects on air pollution makes no sense. But much information may be lost through a coarse aggregation. A little reflection suggests that time and space, as parameters, are inherently uninteresting. What is desired is the relationship between pollution and health independent of (or invariant to) their locations in time and space. But the information contained in the data about this relationship is diffused over time and space and its recovery requires finesse.

In a previous paper [9], the spatial aspects of this problem were addressed. In the current paper we focus on temporal aspects. The method proposed largely avoids aggregation problems, least-squares, linearity, and normality. It is nonparametric in nature and appears to generalize to similar problems in which time and/or space are nuisance parameters.

2. The method. To be specific, let $f(t) = \int_A f(x_1,x_2,t)\, dx_1\, dx_2$ and $h(t) = \int_A h(x_1,x_2,t)\, dx_1\, dx_2$ represent the marginal pollution and health effects distributions over time, where A is the geographic region of interest. [Dividing $f(t)$ and $h(t)$ by the area of A yields the average pollutant concentration and health intensity in A at time t, but in the ratios which follow, this constant of proportionality has no effect because of scale-invariance.] Now if the pollutant does impact health, then the times of highest pollution should be related, caeteris paribus, to the times of greatest concentrations of health cases, the lag depending on a number of factors. Define

$$F(z) = \int_{\{t;\ f(t) \le z\}} f(t)\, dt / \int_I f(t)\, dt$$

$$H_\Delta(z) = \int_{\{t;\ f(t) \le z\}} h(t+\Delta)\, dt / \int_I h(t)\, dt$$

$$T(z) = \int_{\{t;\ f(t) \le z\}} 1\, dt / \int_I 1\, dt$$

where I is the time period under study. The heart of the method involves estimating and comparing these and derivative functions. Note that each is a distribution function in the statistical sense with common value 0 at the minimum pollution value and common value 1* at the maximum pollution concentration. A simultaneous graph of the three functions economically

*H_Δ approximately so; exactly, if $\Delta = 0$.

displays an enormous amount of information. For example if the pollutant is oxidant (in ppm) and the health effect is asthma, then the hypothetical values $F(.08) = .60$, $H_0(.08) = .40$, $T(.08) = .90$ tell us, respectively, that 60% of total oxidant exposure occurred at times of less than .08 ppm concentrations, that those same times accounted for only 40% of asthma cases, although they covered 90% of the time period of the study. If it is supposed that air pollution and health are unassociated and health effects are otherwise randomly distributed over time [see 3a below for relaxation of this assumption] then with $T(.08) = .90$ one would have expected $H_0(.08) \approx .90$. The difference between T and H_0 then may measure the health impact of pollution. But Δ need not be set equal to zero. Choosing a nonzero Δ allows the examination of lagged health effects in a simple manner.

But T, H_Δ, and F are cumulative functions of dose level. In the dose-response problem, one wants to predict the response to a specific dose (whether the dose be measured instantaneously, as a maximum, or as a total). This suggests that a better way to measure the response to dose $z = .08$ is to collect times when z is approximately .08 and examine health responses peculiar to those (perhaps lagged) times. Therefore, given z and $\varepsilon > 0$, the ratio $[H_\Delta(z+\varepsilon) - H_\Delta(z)]/[T(z+\varepsilon) - T(z)] \approx H_\Delta'(z)/T'(z)$ seems a good measure of the response to dose z . Suppose pollution does adversely impact health. Then one expects that if two time intervals of equal duration suffer constant but different exposures $z_1 < z_2$, the second interval (dose z_2)

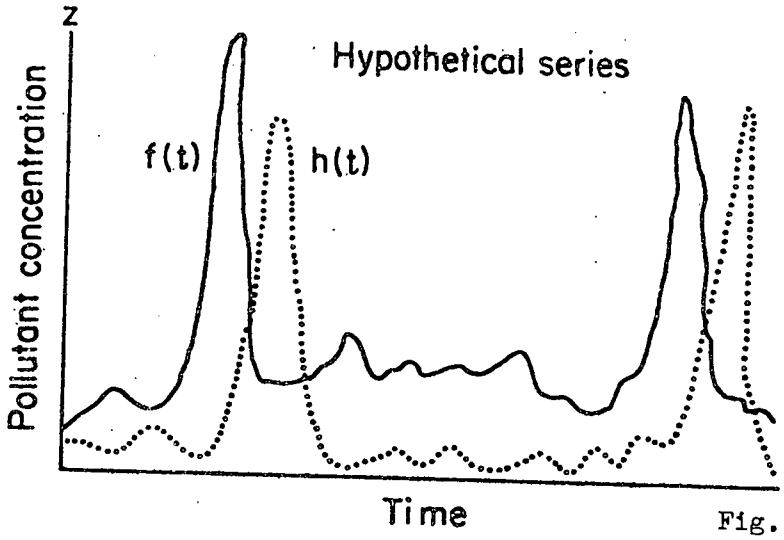

Fig. 1

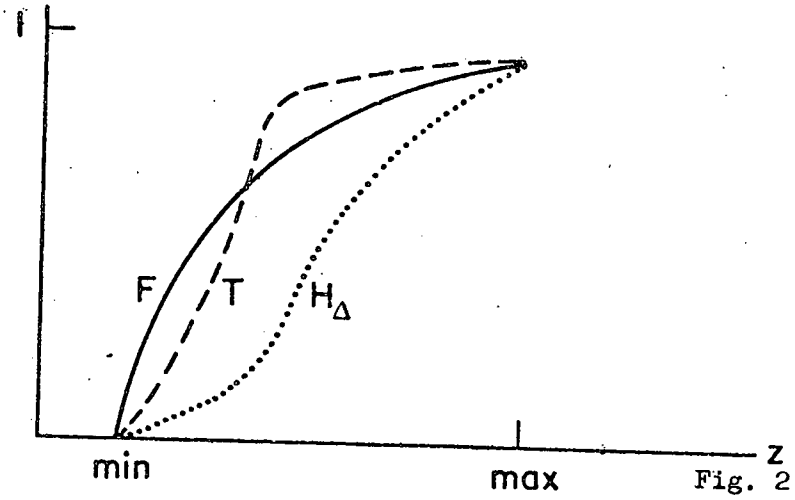

Fig. 2

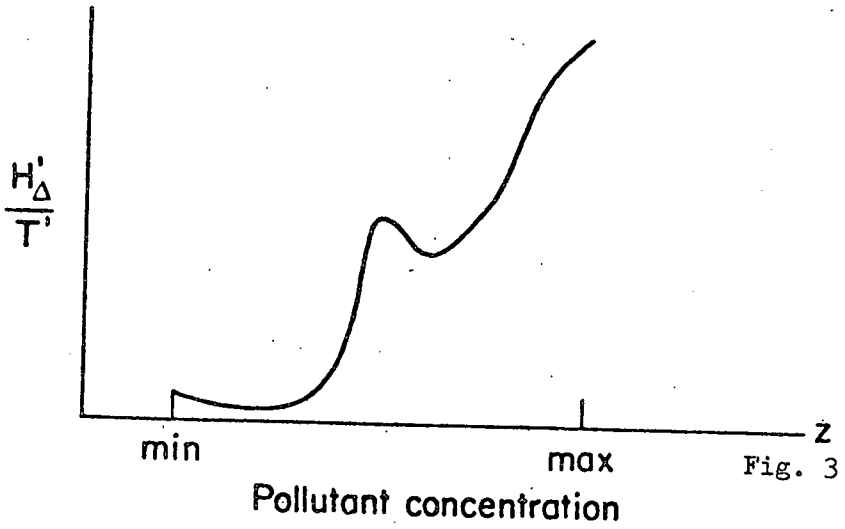

Fig. 3

will experience more health effects. Thus one anticipates that $H'_\Delta(z)/T'(z)$ will be monotonically nondecreasing in z if pollution is harmful. Figures 2 and 3 illustrate the method for the hypothetical series of figure 1.

3. Three refinements.

a. More than one explanatory variable. In the absence of air pollution, health effects are in fact not randomly distributed over time. The adverse conditions brought about by weather result in increased occurrence of health problems. But weather is also a major factor in pollution levels. Weather and other confounding effects render suspect the simple analysis proposed above. If we could remove these effects from the health distribution, we could analyze the remainder for an impact of pollution in the same spirit that the classicist examines residuals for partial correlations. The variable "temperature" will be used here to illustrate the process. Let $c(t) = \int_A c(x_1,x_2,t)\,dt$ denote the marginal distribution of temperature over time, and define

$$C(z) = \int_{\{t;\ f(t) \leq z\}} c(t)\,dt / \int_I c(t)\,dt \ .$$

Then any impact of pollution on health beyond the impact of temperature should show up in $H_\Delta(z) - C(z)$. It is tempting to think of this quantity as a cumulative residual. Then the ratio $[H'_\Delta(z) - C'(z)]/T'(z)$ seems a good measure of the "pure" effect of pollution, after adjusting for temperature. Again, one may expect this quantity to increase monotonically in z . To handle more than one confounding variable, a linear combination of their distributions $\sum_{i=1}^{k} p_i\, C_i$ could be subtracted from H_Δ .

b. Population adjustment. It has been tacitly assumed in the foregoing discussion that the population of the affected area A remains stable over time. If this is not the case, then $h(t)$ must be adjusted to reflect the number of health effects as a proportion of the population. To achieve this, simply replace the integrand 1 by $p(t)$ in the definition of $T(z)$, where $p(t) = \int_A p(x_1,x_2,t)\,dt$ is the population of A at time t. Note that if $p(t)$ = constant, then $T(z)$ is exactly as before. [If one is studying health effects among a subpopulation of "susceptibles," then $p(t)$ may be taken as referring to this group, the actual numbers of which may be unknown. However, if the group is thought to constitute a simple nonvarying proportion of the total population, then H_Δ may be calculated using the total population of the area for $p(t)$.] In all but the longest studies, population will probably remain fairly stable.

c. Spatial variation. A potentially serious defect in the time-analysis presented above is the spatial levelling of $f(x_1,x_2,t)$ and $h(x_1,x_2,t)$ by integrating out x_1 and x_2 to get $f(t)$ and $h(t)$. This was proposed in order to keep the analysis relatively simple. However, its effect may be to hide the impact of pollution, particularly if some regions of A show much more variability of pollution with time than others. An analogous problem in reverse arises with time averaging of $f(x_1,x_2,t)$ and $h(x_1,x_2,t)$ to study spatial variation of these series as was proposed in [9].

One way around this difficulty, which may be feasible in certain cases, is to refrain from averaging $f(x_1,x_2,t)$ and $h(x_1,x_2,t)$.

Then one could define

$$F(z) = \int_{\{(x_1,x_2,t);\ f(x_1,x_2,t) \le z\}} f(x_1,x_2,t)dx_1dx_2dt \Big/ \int_{A\times I} f(x_1,x_2,t)dx_1dx_2dt$$

$$H_\Delta(z) = \int_{\{(x_1,x_2,t);\ f(x_1,x_2,t) \le z\}} h(x_1,x_2,t+\Delta)dx_1dx_2dt \Big/ \int_{A\times I} h(x_1,x_2,t)dx_1dx_2dt$$

$$TP(z) = \int_{\{(x_1,x_2,t);\ f(x_1,x_2,t) \le z\}} p(x_1,x_2,t)dx_1dx_2dt \Big/ \int_{A\times I} p(x_1,x_2,t)dx_1dx_2dt$$

Here the "time-population" adjustment $TP(z)$ is essential, for although population may be stable over time, its density varies greatly over space. The analysis is carried out by comparing H_Δ to TP and examining H'_Δ/TP'. It is evident that the space-time regions $\{(x_1,x_2,t);\ f(x_1,x_2,t) \le z\}$ may not be simple. Their estimation could be achieved by preparing a time-sequential series of pollution maps of the region. On the other hand, estimation of H'_Δ need be no more difficult than outlined under section 4, in which we list the space-time coordinates of each health event and estimate corresponding pollution intensities $\hat{f}(x_1,x_2,t-\Delta)$.

4. Estimation. In general, the functions F, H_Δ, T and their derivatives will not be known and must be estimated. Since we identified H'_Δ/T' as the fundamental dose-response relationship in section 2, density estimation is required.

There are many procedures in the literature for estimating densities (e.g., kernel method [8], orthogonal series [4]; a somewhat dated review paper is [11]). For example, the kernel estimate $\hat{H}'_\Delta(z)$ may be constructed by the following procedure:

(a) list the times $t_1,\ldots, t_n$ of each health effect

(b) estimate $f(t_i-\Delta)$, the geographically-averaged pollution index at times $t_i-\Delta$

(c) from the estimates $\hat{z}_i = \hat{f}(t_i-\Delta)$ of $f(t_i-\Delta)$, construct the estimate $\hat{H}'_\Delta(z) = (nh)^{-1} \sum_{i=1}^{n} K\left((z-\hat{z}_i)/h\right)$ where h is a number depending on n and K is a kernel function (For more on h and K , see [8].)

Here we shall not address the problem of estimating $f(t)$, which may be obtained by cross-sectional, geographical averaging (or totalling) of estimates of $f(x_1,x_2,t)$. Methods now in use for estimating the latter at any given point in time include purely statistical interpolation between monitoring stations (e.g., Kriging [7] or gravity weighting [2]) and physical modelling (e.g., diffusion [10]).

If Δ is not specified but is to be estimated from the data, we suggest choosing Δ to maximize $\frac{1}{n} \sum_{i=1}^{n} \log \hat{H}'_\Delta\left(\hat{f}(t_i-\Delta)\right)$. Given $t_1, \ldots, t_n$, the likelihood is $\prod_{i=1}^{n} H'_\Delta\left(f(t_i-\Delta)\right)$. Since H'_Δ is not modelled parametrically, we replace it by our nonparametric estimate to obtain the above estimated likelihood function. The maximizing Δ is then an (estimated) maximum likelihood estimate.

T' may perhaps be estimated more simply than H_Δ since $T(z)$ is just the proportion of time that the concentration $f(t)$ spends below the level z . For, considerable empirical work (e.g., [3], [5]) suggests that the temporal distribution of pollutant concentrations may be lognormal. If so, then $T'(z)$ is just a lognormal density with parameters which can be easily estimated in the usual way from the data.

The regression ratio H'_Δ/T' may then be estimated by $\hat{H}'_\Delta/\hat{T}'$. If it is expected that this ratio may be nondecreasing (if pollution really matters), we may attempt to recover the trend by smoothing the irregularities through calculating the isotonic regression of H'_Δ/T' (see [1]).

5. Tests. A test for the effect of air pollution on health is a test for the monotonicity of H'_Δ/T'. Unfortunately, none of the tests for trend that have been developed in the literature is adaptable to the present circumstances. However, a test for the health impact against a broader class of alternatives may be performed by testing $H_0: H_\Delta = T$ against $H_1: H_\Delta \neq T$ (or $H_\Delta < T$, which favors monotonicity). Such a test will have less power than the desired test but the theory for testing equality of two distribution functions may be applied. Of course, with a large collection of data (as may be expected in this problem) results will tend to be significant, if only because of inevitable differences between model and reality. The significance level should therefore be interpreted as a relative measure of the degree of concordance.

If Δ is estimated as suggested in section 4, then the above significance tests can no longer be strictly applied. Instead, a likelihood ratio procedure may be employed. Letting the times of health cases be $t_1, \ldots, t_n$ with associated (estimated) pollution values $\hat{z}_1 = \hat{f}(t_1), \ldots, \hat{z}_n = \hat{f}(t_n)$, we calculate $\lambda(\hat{z}_1, \ldots, \hat{z}_n) = \sup_\Delta \prod_{i=1}^n \hat{H}'_\Delta(\hat{z}_i)/\hat{T}'(\hat{z}_i)$ and reject H_0 for large values of λ. If the t_i's are distributed

independently, $\hat{z}_i \approx z_i = f(t_i)$, and both H'_Δ and T' are estimated consistently, then $\lambda \approx \sup_\Delta \prod_{i=1}^n H'_\Delta(z_i)/T'(z'_i)$ and $-2 \log \Lambda = -2 \log \min\{\lambda, 1\}$ is approximately chi-square in distribution (section 7.13 of [6]).

6. Discussion. The method presented in this paper is largely data-analytic in nature. The unwieldy multidimensional time-space series in which pollution and health data are embedded is economically reduced to an informative graph of cumulatives and their derivatives. These graphs are potentially much more faithful to the original data than are the forced aggregations of least-squares regression. For the time being, inference with our method is limited to the nonoptimal tests of section 5. Further advances in statistical theory may be expected to provide (near) optimal tests, confidence intervals, and other desirable statistical properties.

REFERENCES

[1] R. BARLOW, D. BARTHOLOMEW, J. BREMNER, and H. D. BRUNK, Statistical Inference Under Order Restrictions, Wiley, New York, 1972.

[2] J. V. BEHAR, Application of computer simulation techniques to problems in air pollution, Sixth Berkeley Symposium on Math. Stat. and Prob., VI, University of California Press, Berkeley, 1972, pp. 29-69.

[3] J. B. KNOX and R. I. POLLACK, An investigation of the frequency distributions of surface air-pollutant concentrations, Proc. of Symp. on Statist. Aspects of Air Quality Data, U. S. Environmental Protection Agency EPA-650/4-74-038, 1974, chapter 9.

[4] R. KRONMAL and M. TARTAR, The estimation of probability densities and cumulatives by Fourier series methods, Jour. Am. Stat. Assoc. 63 (1968), pp. 925-952.

[5] R. I. LARSEN, A new mathematical model of air pollution concentration, averaging time, and frequency, J. Air Pollution Control Assoc., 19 (1969), pp. 24-30.

[6] E. LEHMANN, Testing Statistical Hypotheses, Wiley, New York, 1952.

[7] G. MATHÉRON, The intrinsic random functions and their applications, Adv. in Appl. Prob., 5 (1973), pp. 439-468.

[8] E. PARZEN, On estimation of a probability density function and mode, Ann. Math. Statist., 33 (1962), pp. 1065-1076.

[9] T. SAGER, Relating spatial distributions of pollutants to health effects, Proc. Ninth International Biometric Conference, Vol. II, The Biometric Society, Raleigh, N. C., 1976, pp. 35-58.

[10] C. SHIR and L. SHIEH, A generalized urban air pollution model and its application to the study of SO_2 distributions in the St. Louis metropolitan area, J. Appl. Meteorology, 13 (1974), pp. 185-204.

[11] E. WEGMAN, Nonparametric probability density estimation: I. a summary of available methods, Technometrics, 14 (1972), pp. 533-546.