THE INTERPOLATION OF AIR POLLUTION:
A STOCHASTIC APPROACH

by

RAY FAITH¹

Stanford University

MASTER

WORKING PAPER #5

This draft is intended for limited circulation.

November 1, 1976

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.

¹This research was supported by Grant from SIMS

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

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

427 *fig attached*

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.

The Interpolation of Air Pollution:

A Stochastic Approach

Ray Faith*

Abstract. In epidemiological studies of the effects of air pollution it is desirable to obtain estimates of the exposure to various pollutants of persons living throughout a geographical region under study. On the other hand, it is typically the case that pollution levels are measured at only a few locations. The problems of interpolating these values and of determining the accuracy of interpolation are amenable to analysis by way of spatial stochastic models. Such models can lead to an optimal interpolation technique, in a manner which is reviewed here. The method requires for its application knowledge of the spatial covariance structure of the pollutant to which it is applied. Some difficulties encountered in estimating this for oxidant levels in the San Francisco Bay Area are mentioned.

The Kriging Method. The method discussed here has been developed extensively by other authors, notably Mathéron (1973) and Delfiner (1976), and takes its name from Krige (1951). The following is an elementary special case of the more general theory.

*Department of Statistics, Stanford University, Stanford, California 94305.

Let Y_x , $x \in R$, be a real valued stochastic process, where R is a region in the plane. Let $x_1, \dots, x_K$ be points in R at which Y is known. Let $s = (x_1, \dots, x_K)$ and $Y_s = (Y_{x_1}, \dots, Y_{x_K})'$. A linear estimator of Y_x is an estimator of the form $\hat{Y}_x = \sum_{i=1}^K \lambda_i(x) Y_{x_i}$. Only the class of linear estimators is considered here. The error of estimation at x is $Z_x = \hat{Y}_x - Y_x$.

We introduce the following assumptions concerning the stochastic model.

(a) $EY_x \equiv \mu$ a constant, and

(b) $\frac{1}{2} E(Y_{x+h} - Y_x)^2 = \gamma(h)$ depends only on h .

The function $\gamma(h)$ is called the variogram of the process. If the process Y_x has finite variance, then these assumptions are equivalent to the assumption that Y_x is a weakly stationary stochastic process with covariance function $c(h) = \text{Cov}(Y_{x+h}, Y_x)$, and $\gamma(h) = c(0) - c(h)$. Assumption (b) is somewhat more general, however, and only requires Y to have stationary increments, thus allowing, for example, a model of random drift.

It is immediate from assumption (a) that a linear estimator $\hat{Y}_x$ is unbiased in the sense that the expected error is zero ($E Z_x = 0$) if and (except for the special case $\mu=0$) only if $\sum_{i=1}^K \lambda_i(x) \equiv 1$. An unbiased linear estimator, is therefore, a weighted average of the data values.

The Kriging method consists of taking $\hat{Y}_x$ to be that unbiased linear estimator of Y_x which minimizes the mean squared error $MSE_x = E(Z_x^2)$. An application of Lagrange multipliers shows MSE_x to be minimized by

$$(1) \quad \hat{Y}_x = \hat{\mu} + \gamma_{xs} \gamma_{ss}^{-1} (Y_s - \hat{\mu} \tilde{1})$$

where

$$\gamma_{xs} = (\gamma(x_1 - x), \dots, \gamma(x_K - x)) \quad 1 \times K,$$

$$\gamma_{ss} = (\gamma(x_i - x_j)) \quad K \times K,$$

$$\tilde{1} = (1, \dots, 1)',$$

and

$$\hat{\mu} = \frac{\tilde{1}' \gamma_{ss}^{-1} Y_s}{\tilde{1}' \gamma_{ss}^{-1} \tilde{1}}.$$

If Y_x has finite variance then $\gamma(h)$ can be replaced by $c(h)$ throughout the above expressions. It then follows that $\hat{\mu}$ is the usual minimum variance unbiased estimator of μ , and $\hat{Y}_x$ is the usual regression estimate of Y_x given Y_s with $\hat{\mu}$ substituted for μ .

In addition to producing the optimal estimator $\hat{Y}_x$, the above framework tells us its mean squared error, namely

$$(2) \quad MSE_x = E((\hat{Y}_x - Y_x)^2) = \gamma_{xs} \gamma_{ss}^{-1} \gamma_{sx} - \frac{(\tilde{1}' \gamma_{ss}^{-1} \gamma_{sx} - 1)^2}{\tilde{1}' \gamma_{ss}^{-1} \tilde{1}},$$

which, in the case of a finite variance, may be expressed as

$$MSE_x = \frac{(1' c_{ss}^{-1} c_{sx} - 1)^2}{1' c_{ss}^{-1} 1} - c_{xs} c_{ss}^{-1} c_{sx} + c(0) .$$

It is worthwhile to note that even in the case of a process Y_x with infinite variance, $\hat{Y}_x - Y_x = \sum_1^K \lambda_i(x) (Y_{x_i} - Y_x)$ has finite variance since, for each i , $Y_{x_i} - Y_x$ has variance $2\gamma(x_i - x) < \infty$.

Examination of (1) and (2) show that, at each data location $x=x_i$, the estimator $\hat{Y}_x$ equals the data value Y_{x_i} and the error variance is zero. If $c(h)$ tends to zero for large h , then the estimate tends to $\hat{\mu}$ at points far from the data and the error variance tends to

$$\frac{1}{1' c_{ss}^{-1} 1} + c(0) = \text{Var}(\hat{\mu}) + \text{Var}(Y_x) .$$

Estimating the Variogram. In order to use this method to interpolate between monitoring station values, it is necessary to either use an a priori variogram function $\gamma(h)$, or to estimate one from data. The former course of action is of questionable merit, since it is possible in this way to produce a wide variety of quite distinct interpolating functions $\hat{Y}_x$. (On a related question, however, Sheshinsky (1976) has shown that the application of this method to estimating a linear trend is relatively robust in $\gamma(h)$.)

The latter approach is a promising one and as a research area is by no means yet fully explored in the literature. In the past the

emphasis has been on data consisting of single measurements from each of a relatively large number of locations, as is found in geology, for example. With air monitoring data the more typical situation is that of relatively few locations, but with repeated measurements at each of them. This difference is likely to be reflected in the choice of methodology appropriate to estimate $\gamma(h)$.

With few locations available it is necessary to make some simplifying assumptions about $\gamma(h)$. For example, it may be assumed that it is an isotropic function ($\gamma(h) = \gamma(|h|)$), or is a simple function of directed distance (e.g. $\gamma(h) = \gamma_1(h_1) \cdot \gamma_2(h_2)$ where $h = (h_1, h_2)$.) If there are k stations and the pollutant is observed once at each of them, then there are $k(k-1)/2$ observations of $(Y_{x_i} - Y_{x_j})^2$, each one associated with a spatial difference $h_{ij} = x_i - x_j$. If independent measurements are made n times and, if the underlying model is true, then the average of $(Y_{x_i} - Y_{x_j})^2$ taken over the n observations should tend to $E[(Y_{x_i} - Y_{x_j})^2] = 2\gamma(x_i - x_j)$. If the n observations are obtained from a time series on the other hand, then the assumption of independence is not warranted and the convergence of these averages will be governed by not only the spatial, but also the time dependence of the process.

My personal experience with attempting to find a reasonably smooth function for $\gamma(x)$ using daily mean oxidant observations from 1971 to 1973 at 9 stations located in the San Francisco Bay Area (which is

meteorologically quite heterogeneous for a small area) has demonstrated the difficulty of this task. It occurs that there are stations which are located near each other which tend to differ greatly in oxidant levels, whereas there are other stations located far apart and which show relatively strong agreement. This phenomenon remains true after a yearly seasonal component is subtracted for each station's time series and even when a weekly seasonal effect is removed. It is possible that the high correlation observed over long distances is an artifact of the influence of weather on oxidant levels, the data being suggestive of this, but the issue has not been sufficiently studied to confirm this.

The state of the art is at this point very crude and it is impossible to say what approach is going to prove most effective in the determination of $\gamma(h)$ from the type of data available in air pollution work. In particular, it remains to be examined how best to incorporate seasonal and meteorological information into the estimation of $\gamma(h)$ and/or into the reformulation of the spatial model itself.

REFERENCES

- [1] P. DELFINER, Linear estimation of non-stationary spatial phenomena, Proceedings of NATO A.S.L. "Geostat 75", Rome, Italy. Reidel Publ. Corp., Dordrecht, Netherlands, 1976.
- [2] D. G. KRIGE, A statistical approach to some basic mine value problems in Witwatersrand, J. Chem. Met. Min. Soc. S. Afr., 52 (1951), pp. 119-139.
- [3] G. MATHÉRON, The intrinsic random functions and their applications, Adv. in Appl. Prob., 5 (1973), pp. 439-468.
- [4] R. SHESHINSKI, Robustness of linear estimators to misspecification of trend, SIMS Technical Report No. 4, Department of Statistics, Stanford University, 1976.