ESTIMATION OF SPATIAL DISTRIBUTIONS
FROM POINT SOURCES WITH APPLICATION
TO AIR POLLUTION MEASUREMENT

PAUL SWITZER

TECHNICAL REPORT NO. 9
JULY 1, 1977

PREPARED UNDER THE AUSPICES OF
SIAM INSTITUTE FOR MATHEMATICS AND SOCIETY

SIMS

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

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

BROOKLYN PATENT GROUP

1/12 1978 BY *CHC*

MASTER

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

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

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.

ESTIMATION OF SPATIAL DISTRIBUTIONS
FROM POINT SOURCES WITH APPLICATION
TO AIR POLLUTION MEASUREMENT

by

PAUL SWITZER

Stanford University

TECHNICAL REPORT NO. 9

JULY 1, 1977

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Department of Energy, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

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

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

CB

ESTIMATION OF SPATIAL DISTRIBUTIONS
FROM POINT SOURCES WITH APPLICATION
TO AIR POLLUTION MEASUREMENT

Paul Switzer

Stanford University

1. INTRODUCTION

Suppose we want to estimate the frequency distribution of an air pollution variable Z in a specified air basin from observations made at fixed monitoring stations over a period of time. For a possible mathematical formulation of the problem we suppose, for now, a stationary random space-time field for Z for which we want to estimate the stationary (marginal) distribution function F . The usual empirical distribution function, edf, of the data may be an unsuitable estimator because it does not take account of the correlated nature of the observations. The unsuitability of the usual edf estimator will appear most pronounced if there are "clusters" of highly correlated observations in the data set. The general problem of estimating distributions from correlated observations arises frequently with geographic data as, for example, also in the estimation of an ore-grade distribution from assay samples.

In Section 2 we introduce and explore a new class of "weighted" distribution function estimators which exploit correlation structure. In Section 3 we discuss the estimation of correlation structure from available data and illustrate with an

example using particulate concentration data from the San Francisco Bay Area air basin. In Section 4 we deal with the related problem of estimating the spatial distribution function of a fixed-time realization of the random field. That is, the objective is to estimate what proportion of the space domain has $Z \leq z$ (as a function of z) at a specified time point.

The advantage of the weighted estimators of distributions which will be presented lies not only in their increased precision relative to edf. For example, it may be that the data from some monitoring stations contribute little or nothing to these weighted estimates. Such a result could have implications for future data gathering.

2. WEIGHTED ESTIMATORS OF DISTRIBUTION FUNCTIONS

The N observations are denoted $Z_1, Z_2, \dots, Z_N$. Formally they are considered to be non-independent identically distributed random variables. To each Z_i we associate a weighting coefficient w_i which is positive or zero with $\sum w_i = 1$. A weighted estimator $\hat{F}$ of the stationary distribution function F is defined as

$$\hat{F}(z) = \sum_{i=1}^N w_i I(Z_i \leq z) \quad (1)$$

where I is the indicator function. Equivalently, the observation Z_j itself is an estimator of the p -quantile z_p of the distribution F where p is any number in the interval $(0,1)$ satisfying

$$\sum_{i=1}^N w_i I(Z_i < Z_j) < p \leq \sum_{i=1}^N w_i I(Z_i \leq Z_j)$$

Hence the quantiles estimated by Z_j depend not only on the rank order of Z_j but on the complete rank ordering of all the observations.

When all the observations are equally weighted, $w_i \equiv 1/N$, then $\hat{F}$ is the usual edf of the data. However, we will want to choose the weights w_i so as to exploit the non-independent structure of the data. To this end we specify a measure L of the discrepancy between $\hat{F}$ and F and then choose the w_i to minimize the mean value of this discrepancy measure taken over the joint distribution of the observations. A convenient choice for L is

$$L = \int_0^1 \frac{[\hat{F}(z_p) - F(z_p)]^2}{p(1-p)} dG(p) \quad (2)$$

where G is a specified distribution function on the interval $(0,1)$ which is used to weight the squared difference between $\hat{F}$ and F at different p -quantiles. The factor $1/p(1-p)$ in the above expression normalizes L to unity for any choice of G when the data consist of a single observation. If we are estimating one particular p -quantile of F , then G will have all its mass at that particular p value.

When the weighted estimator (1) is substituted into the expression for the criterion L , then the mean of L may be expressed as

$$E[L] = \sum_{i=1}^N \sum_{j=1}^N w_i w_j \bar{R}_{ij}$$

where

$$\bar{R}_{ij} = \int_0^1 R_{ij}(p) dG(p) \quad (3)$$

and

$$R_{ij}(p) = \{\text{Prob}[Z_i, Z_j \leq z_p] - p^2\} / p(1-p)$$

The derivation of the expression above is straightforward and uses the fact that $E[\hat{F}(z_p)] = F(z_p) = p$ for every p . The quantity $R_{ij}(p)$ is the ordinary correlation coefficient between two dichotomous random variables formed by splitting Z_i and Z_j at the common p -quantile z_p . In particular, $R_{ii}(p) = 1$ for all i and p , and when Z_i, Z_j are independent, then $R_{ij}(p) = 0$ for all p .

In general, the minimization of $E[L]$ over the vector of weights $\underline{w} = (w_1, \dots, w_N)$, subject to constraints $\sum w_i = 1$ and $w_i \geq 0$ all i , will require some convex quadratic programming algorithm and, in general, some of the minimizing weights will be zero. However, where the minimizing weights $\underline{w}^* = (w_1^*, \dots, w_N^*)$ are all strictly positive, then they could also have been obtained by elementary calculus methods. In this case it follows that

$$\begin{aligned} \min E[L] &= E[L^*] = 1/\sum \sum \bar{R}^{ij} , \\ w_i^* &= E[L^*] \cdot \sum_j \bar{R}^{ij} , \quad i = 1, \dots, N , \end{aligned} \quad (4)$$

where

$$\text{matrix}(\bar{R}^{ij}) = \text{matrix}(\bar{R}_{ij})^{-1} .$$

This representation of the solution involves inversion of an $N \times N$ matrix. In fact, we need only invert an $(N-1) \times (N-1)$ matrix in actual computations because of the constraint $\sum w_i = 1$.

The minimizing weights depend on the dichotomous correlations $R_{ij}(p)$ defined in (3). It would be interesting to see how these vary with p and how they are related to the ordinary correlation structure of the original observations themselves. If Φ denotes the standard normal distribution function, then the transformed random variables $T(Z_i) = \Phi^{-1} F(Z_i)$, $i = 1, \dots, N$, each have standard normal distributions provided the stationary F is continuous. In addition, suppose the transformed pairs $T(Z_i)$,

$T(Z_j)$ for all i, j are also jointly normal. Then the quantities $p^2 + p(1-p)R_{ij}(p) = \text{Prob}[T(Z_i), T(Z_j) \leq \Phi^{-1}(p)]$ may be found in tables of the standard bivariate normal distribution with correlation coefficient $\rho_{ij} = \text{corr}(T(Z_i), T(Z_j))$. Such bivariate tables, for example, may then be used to calculate the required dichotomous correlations $R_{ij}(p)$ as a function of the correlation ρ_{ij} of the T-transformed observations and of p . We note that $R_{ij}(1-p) = R_{ij}(p)$, $R_{ij}(p) = 0$ for all p if $\rho_{ij} = 0$, and $R_{ij}(p) = 1$ for all p if $\rho_{ij} = 1$.

Table I illustrates the insensitivity of these dichotomous correlations $R_{ij}(p)$ to values of p in the middle range.

TABLE I

Dichotomous Correlations $R_{ij}(p)$ as a Function of the Quantile p and the Usual Correlation ρ_{ij} between Normally Transformed Observation Pairs

$\rho_{ij} =$	0.0	0.1	0.3	0.5	0.7	0.9	1.0
$p = 0.1$ or 0.9	0	.06	.19	.33	.49	.71	1
$p = 0.3$ or 0.7	0	.06	.18	.32	.48	.70	1
$p = 0.5$	0	.04	.13	.25	.41	.65	1

Therefore, in the present context, the averaged dichotomous correlations $\bar{R}_{ij}$ used to calculate the observation weights (4) should depend very little on the specified averaging function G used in the loss criterion (2), provided G is concentrated in the middle range of p -values. In particular, if G is totally concentrated at $p = 0.5$, then it follows that

$$\bar{R}_{ij} = R_{ij}(0.5) = 2\pi^{-1} \arcsin \rho_{ij} \quad (5)$$

Therefore, in many practical situations we may dispense with bivariate normal tables and the integration of $R_{ij}(p)$ by using (5) as a general approximation for $\bar{R}_{ij}$. This approximation will be used for all the illustrations of this paper.

On the other hand it may be shown that the dichotomous correlation $R_{ij}(p)$ goes to zero as $p \rightarrow 0$ or $p \rightarrow 1$ whatever the correlation ρ_{ij} between the transformed observations may be ($|\rho_{ij}| \neq 1$). So if we emphasize estimation of extreme p-quantiles, then the weighting for a correlated set of observations will be more like that for independent observations, i.e., the usual edf.

It should be informative to examine a few simple illustrations of the relation between the correlation pattern of the observations and the corresponding distribution function estimates (1) determined by minimizing (3). Consider the case of three observations which, after a normalizing transformation, have correlations $\rho_{12} = \rho_{23}$ and $\rho_{13} > 2\rho_{12}^2 - 1$ to preserve positive definiteness. It may be useful to picture these data as being taken from monitoring stations forming an isosceles triangle. Suppose the dichotomous correlations are calculated as in (5). Let $\alpha = 1 - 2\pi^{-1} \arcsin \rho_{12}$ and let $\beta = 2\pi^{-1} \arcsin \rho_{13} - 2\pi^{-1} \arcsin \rho_{12}$. Then the observation weights which minimize the loss criterion (3) would be in the ratio $\alpha : (\alpha + \beta) : \alpha$ and the middle observation could then have any weight between 0.0 and 0.5. The value of the minimized loss $E[L^*]$ is $1 - 2\alpha^2/(3\alpha + \beta)$ which is 10%-20% less, for most correlation patterns, than the suboptimal value of $E[L] = 1 - 2(3\alpha - \beta)/9$ which we get from equal weighting of the observations. In the case where the correlation matrix has Markov structure $\rho_{13} = \rho_{12}^2 = \rho_{23}^2$, the minimum possible weight for the middle observation is $(3 - 2^{1/2})/7 = 0.227$ and the maximum is $1/3$.

As a second illustration, consider the case of four observations with $\rho_{12} = \rho_{13} = \rho_{23} = 0$ and $\rho_{14} = \rho_{24} = \rho_{34} = \rho$, say; for positive definiteness we need $\rho < 1/\sqrt{3} = 0.577$. It is useful to

regard the observations Z_1, Z_2, Z_3 at the corners of an equilateral triangle with Z_4 at its center. Applying (4) and (5) we find that the corresponding observation weights should be in the ratio $(1-R) : (1-R) : (1-R) : (1-3R)$ where $R = 2\pi^{-1} \arcsin \rho$, provided that this dichotomous correlation R is not greater than $1/3$. However, for all $R > 1/3$, equivalently $\rho > 1/2$, the central observation Z_4 would be given zero weight; and the resulting distribution function estimator is the same as the usual edf constructed from the exterior observations only.

For a third simple illustration consider a sequence of N dichotomized observations forming a stationary two-state first order Markov chain. The objective here will be to estimate the stationary state probabilities p_0 and $1 - p_0$ using a weighted frequency estimator of the form (1) and a loss criterion of the form (2) with $G(p)$ concentrated at p_0 and $1 - p_0$. The dichotomous correlation matrix is $R_{ij}(p_0) = R_{ij}(1-p_0) = c^{|i-j|}$ for some constant $c \in [0,1)$. The N observation weights determined according to (4) turn out to be equal for all interior points in the data sequence, but the two end observations are each given greater weight. Specifically, $w_i^*/w_1^* = w_i^*/w_N^* = 1 - c$ for all $i \neq 1$ or N , regardless of the length N of the sequence. Hence, the ratio of an inner weight to an outer weight can be arbitrarily small but for long sequences the inner weights will nevertheless be close to $1/N$. For the minimized mean loss criterion we get $E[L^*] = (1+c)w_1^*$ which is $N^{-1}(1+c)/(1-c)$ to order N^{-1} , the same value one would get from the suboptimal equal weighting of all the observations.

The rewards of differential weighting of observations would seem to be most pronounced when these observations come in correlated "clusters" of varying size. In a somewhat extreme form a formal notion of cluster may be given as follows--the data have cluster structure if there is a partition of the N observations into subsets (clusters) which are mutually statistically independent. A cluster is denoted generically by the letter K .

A cluster may consist of a single observation if it is independent of the rest of the data. The partition into clusters then implies a representation of the correlation matrix of the observations in a block-diagonal form. For data with such cluster structure we may also partition the mean loss criterion (3) thus:

$$E[L] = \sum_K \sum_{i,j \in K} w_i w_j \bar{R}_{ij} = \sum_K W_K^2 E[L_K]$$

where

$$E[L_K] = \sum_{i,j \in K} w'_i w'_j \bar{R}_{ij} ; \quad W_K = \sum_{i \in K} w_i ; \quad w'_i = w_i / W_K \text{ for } i \in K .$$

The expression above follows from the fact that $\bar{R}_{ij} = 0$ unless i and j are in the same cluster.

We will minimize $E[L]$ in two stages--first minimizing with a fixed set of total cluster weights $\{W_K\}$ and then minimizing over $\{W_K\}$. However, for each cluster K the minimum value $E[L_K^*]$ of $E[L_K]$ above does not depend on $\{W_K\}$ since it is equivalent to the minimization problem in which the data consist only of cluster K . Now minimizing $E[L]$ with respect to $\{W_K\}$, subject to $\sum W_K = 1$, can be done by elementary methods. The optimum total cluster weight W_K^* then turns out to be proportional to $1/E[L_K^*]$ for each cluster K .

To further illustrate, consider the "intra-class" correlation model wherein the observations within a cluster are exchangeable random variables. Then it follows from symmetry arguments or otherwise that all observations within a cluster will have equal weight. So if there are N_K observations in cluster K and R_K denotes their common averaged dichotomous pairwise correlation (3), then it follows by substitution into (4) that

$$E[L_K^*] = N_K^{-1} + (1 - N_K^{-1}) R_K , \tag{6}$$

$$W_K^* \propto 1/E[L_K^*] \text{ and } w_i^* = W_K^*/N_K \text{ for } i \in K .$$

These results are parallel to those in the theory of sample surveys for the problem of estimating a population mean from cluster samples, e.g., see Cochran (1963). As an example, suppose we have a cluster with three observations whose common dichotomous correlation is 0.5, and in addition two separate "isolated" observations. Here there are three clusters in all with $N_1 = 3$, $N_2 = 1$, $N_3 = 1$ and substitution into (6) gives weights $1/7$ for each of the three clustered observations Z_1, Z_2, Z_3 and weight $2/7$ for each of the two isolated observations Z_4, Z_5 . If the ordering of the observed values turns out to be $Z_1 < Z_2 < Z_3 < Z_4 < Z_5$, then the median observation Z_3 will be an estimate of the p -quantiles of the stationary distribution F for $p \in (2/7, 3/7]$ using the weighted estimator (1). However, if the order of the observed values turns out to be $Z_4 < Z_5 < Z_3 < Z_2 < Z_1$, then the median observation is still Z_3 ; but it will now be an estimate of the p -quantiles for $p \in (4/7, 5/7]$.

When the number of observations in a given cluster is allowed to become large, then the expression (6) shows that the total cluster weight W_K^* will approach an asymptote proportional to $1/R_K$ when $R_K > 0$. Now consider the case of a data set consisting of one large cluster and one isolated observation in the context of the aforementioned intra-class correlation model. The weight to be attached to the single isolated observation is very nearly $R_K/(1+R_K)$ however large the large cluster may be. The value of the minimized total mean loss criterion $E[L^*]$ is also nearly $R_K/(1+R_K)$ for $R_K > 0$ and large N_K , so that the equivalent number of independent observations is $1 + R_K^{-1}$. For example, when the dichotomous correlation $R_K = 0.5$, then the single isolated observation will receive weight at least $1/3$ no matter how large is the other cluster. If R_K is misspecified, then the assignment of weights will be suboptimal; and the resulting mean loss, say $E[L^S]$, is greater than $E[L^*]$; ratios of these two quantities are tabled below for the situation described in this paragraph. It can be seen from Table II that the degree of suboptimality is

minor in the middle range and is most serious when a moderate or strong intra-class correlation is assumed in place of an actual weak correlation.

TABLE II

Ratio of Suboptimal to Optimal Mean Loss $E[L^S]/E[L^*]$

		Actual intra-class dichotomous correlation				
		0.1	0.3	0.5	0.7	0.9
Assumed intra-class dichotomous correlation	0.1	1.00	1.11	1.26	1.43	1.59
	0.3	1.24	1.00	1.05	1.14	1.24
	0.5	1.71	1.06	1.00	1.03	1.08
	0.7	2.25	1.18	1.03	1.00	1.02
	0.9	2.77	1.33	1.09	1.02	1.00

3. ESTIMATION OF CORRELATION STRUCTURE

The dichotomous correlation $R_{ij}(p)$ between observation pairs Z_i and Z_j was defined in Section 2 as the ordinary correlation between these observations when they have been dichotomized at the p -quantile of the stationary distribution F . These correlations are what are required for the determination of observation weights appropriate to the minimization of the criterion (3) for estimation of a distribution function. It was noted that if a transformation to normality of Z_i and Z_j also induced bivariate normality, then $R_{ij}(p)$ could be expressed in terms of a standard bivariate normal integral depending only on p and the ordinary correlation ρ_{ij} between the transformed observations. Hence, the problem is now focused on estimating the $\rho_{ij} = \text{corr}(\Phi^{-1}F(Z_i), \Phi^{-1}F(Z_j))$.

It seems that the simplest idea to exploit is to replace the transformation function $T = \Phi^{-1}F$ by $\Phi^{-1}\hat{F}_0$ where $\hat{F}_0$ is a preliminary estimate of the stationary distribution function.

For example, if $\hat{F}_0$ is the usual edf, then the estimated transformation is approximately equivalent to replacing the original observations by normal scores according to their rank order. If we can then estimate the correlations of the normalized observations, we will get an improved weighted estimator of F which may be used to obtain an iterated estimate of the normalizing transformation.

The statistical properties of such an estimation procedure will be complicated if only because we are dealing with non-independent data. However, there is a situation of possible relevance to air pollution data for which available statistical theory may be invoked. Specifically, suppose the N observations may be doubly subscripted, Z_{hk} , for $h = 1, \dots, m$ and $k = 1, \dots, n$ with $N = mn$, such that Z_{hk} and $Z_{h'k'}$ are taken to be independent when $h \neq h'$. For example, h indexes time periods and k indexes monitoring stations. Then the preceding assumption allows for correlation between observations from different stations at the same time period (induced by common weather conditions) but requires that observations made at different time periods be independent (by using a sufficiently large time spacing in the data). Using the definitions of Section 2, the observations are in the form of m clusters each of size n . If we further allow that the correlation pattern between station observations is constant over time, then the weight to be assigned to station k is also constant over time. Therefore, it is sufficient to consider the problem at a single time point although time replication will naturally be used in the estimation of station to station correlations.

In particular, suppose we rank the observations Z_{hk} separately at each station and replace them by corresponding normal scores T_{hk} . Then, barring tied observations, we have the same collection of transformed observations at each station. Finally, the ordinary correlation coefficient $\hat{\rho}_{kk'}$, computed from the pairs $(T_{hk}, T_{hk'})$, $h = 1, \dots, m$, will be an estimate of the required

transformed correlation $\rho_{kk'}$, between stations k and k' , viz., we may deduce from Puri and Sen (1971, p. 375) that the limiting distribution of $m^{1/2}(\hat{\rho}_{kk'} - \rho_{kk'})$ is standard normal for $m \rightarrow \infty$ under certain continuity conditions on the joint distributions. Essentially, we are using a separate estimate $\Phi^{-1}\hat{F}_0$ of the normalizing transformation at each station by calculating a separate edf $\hat{F}_0$ at each station. Provided there is sufficient independent time replication, this procedure seems reasonable and fits an established statistical limit theory.

Chiefly as an exercise, we took a data base of $N = 294$ observations of 24-hour total suspended particulates consisting of measurements taken at $n = 14$ San Francisco Bay Area stations at three-day intervals (the usual frequency) in the months June, July, 1974, for a total $m = 21$ time points. Transformed correlations $\rho_{kk'}$, between stations were estimated according to the method of the preceding paragraph and are shown in Table III. Although there was also some time correlation present in these data, it was ignored for purposes of this illustration. Following the discussion surrounding (5) we took $\bar{R}_{kk'} = 2\pi^{-1} \arcsin \rho_{kk'}$, and derived the corresponding set of 14 station weights by minimizing (3). The weights, as shown at the bottom of Table III range from 0.01/m for the Sunnyvale (SU) station to 0.15/m for the San Francisco (SF) station. In an example such as this the loss criterion is very flat near its minimum; hence, we should expect that small changes in the correlations could induce rather noticeable changes in the minimizing set of observation weights. Therefore, no special importance should be attached to the particular weights shown in Table III. An interesting problem, not explored in this paper, is to find a set of "nearly" optimum weights in which as many as possible of the weights are zero.

However, the estimated distribution function itself (1) is not similarly sensitive. Using the weights shown in Table III it has been sketched in Figure 1 in the form of a histogram with solid lines; the broken lines in Figure 1 indicate the usual

TABLE III

Estimated Correlations of Normally Transformed TSP
Measurements for a 14-Station Monitoring Network
In SF Bay Area and Derived Station Weights
For Estimating the Distribution Function

	SF	SR	RI	PT	CC	FR	LI	SJ	SU	RC	BU	ST	NP	VA
SF	1													
SR	.39	1												
RI	.48	.51	1											
PT	.51	.61	.48	1										
CC	.37	.61	.69	.64	1									
FR	.45	.77	.40	.43	.48	1								
LI	.59	.68	.32	.61	.44	.71	1							
SJ	.49	.40	.66	.59	.58	.50	.52	1						
SU	.49	.70	.43	.57	.54	.88	.65	.62	1					
RC	.48	.54	.78	.44	.74	.51	.52	.63	.50	1				
BU	.42	.77	.51	.47	.69	.70	.56	.46	.60	.64	1			
ST	.42	.67	.47	.48	.39	.70	.54	.40	.62	.45	.62	1		
NP	.47	.84	.55	.57	.57	.83	.61	.38	.75	.53	.54	.72	1	
VA	.50	.59	.65	.54	.56	.43	.56	.46	.55	.67	.52	.43	.51	1
Wts.	.15	.01	.07	.09	.08	.06	.07	.10	.01	.03	.07	.12	.04	.10

unweighted histogram. The closeness of the two histograms is attributable in part to the absence of cluster structure in the correlation pattern of the monitoring stations. The estimated value of the minimized average loss criterion $E[L^*]$ is 0.41/m compared with 0.43/m for the unweighted procedure. Hence, we do not have the equivalent of even three mutually independent monitoring stations for which the criterion value would be 0.33/m.

As an alternative to using separate data transformations $\Phi^{-1} \hat{F}_0$ for each station we might have used the overall edf of all 294 observations as a preliminary $\hat{F}_0$. Effectively this means replacing the observations by normal scores according to their respective ranks in the full data set. However, we did not explore this alternative. There remains the question whether or

FIG. 1

Estimated Stationary TSP ($\mu\text{g}/\text{m}^3$) Distribution Derived
From 21 Observations At Each Of 14 Monitoring Stations

not it is useful to postulate a single distribution function F , i.e., a stationary random field, as a model for the variation of particulate concentrations at all 14 monitoring stations. It is indeed the case that some stations appear to have noticeably higher mean concentrations than others. This raises the question of the interpretation to be given to the estimated stationary distribution function, a topic not dealt with here.

When the data structure is not that of repeated independent observations of a multivariate random vector, as for example in a collection of sample ore grades from a mine, then additional structure needs to be exploited to obtain estimates of the correlations between observations. It has been usual in such situations to suppose that correlations may be calculated from a simple parametric function of the geographic separation between

observation pairs, thereby relying more explicitly on spatial second-order stationarity. The data are then used to estimate the parameters of the simple correlation function. For examples of such parametric correlation estimation see David (1976). Of course, we will still want the correlations of normally transformed data; so we first would apply an estimated transformation of the form $\Phi^{-1} \hat{F}_0$ as discussed earlier in this section. A different approach to the estimation of the required data transformation $\Phi^{-1} F$ has been discussed by Mathéron (1976). His approach does not use a preliminary $\hat{F}_0$ but rather approximates $\Phi^{-1} F$ by a polynomial whose coefficients are estimated by the data. Mathéron's approach is also useful for the more general problem of estimating the distribution function of an integral of a random field, a problem which arises in the distribution of average grades taken over blocks of ore.

Finally, it may be possible to bypass the notion of a transformation to bivariate normality as a means of estimating the required dichotomous correlations when the data have the form of repeated observations at a fixed network of stations. For example, the value of $R_{ij}(0.5)$ given in (3) may be estimated directly by dichotomizing the observations for each station at their respective sample medians and calculating the correlation coefficients between station pairs using these dichotomized observations. Such estimated correlations are sometimes called "corner" correlations. They were computed for the same 14-station air pollution data and were found to agree reasonably well with the values of the estimated dichotomous correlations $2\pi^{-1} \arcsin \rho_{ij}$ obtained via the normal transformation model and the correlation matrix of Table III. However, with as few as 21 observations per station the directly estimated corner correlations are rather choppy. In principle, we may estimate the $R_{ij}(p)$ for any p -quantile by a similar direct procedure in which we dichotomize at the sample p -quantiles at each station. These direct estimates will not be smooth functions of p and may have

poor statistical properties for p not near 0.5 unless the number m of repetitions is large.

Instead of using a separate estimated p -quantile at each station as a dichotomization point, we may consider dichotomizing all observations at the same point using a single overall estimated p -quantile derived from a preliminary overall distribution function estimate $\hat{F}_0$. However, the resulting correlation estimates may become very sensitive to the stationarity and time independence assumptions; for it could well happen that all or nearly all the data from a given station lie on one side of the overall preliminary p -quantile even when p is near 0.5. If the data are not in the form of repeated observations at a station network, then we may wish to use overall preliminary estimates of p -quantiles in a framework in which dichotomous correlations are described as a simple parametric function of geographic separation between observations.

4. ESTIMATION OF SPATIAL DISTRIBUTION FUNCTIONS AT A FIXED TIME

The problem posed in Section 2, that of estimating the stationary distribution function F presumed to have generated a finite set of correlated observations, is replaced in this section by a related problem more explicitly tied to the continuous spatial random field in which the observations are embedded. In the air pollution context, suppose we wish to estimate the areal proportion of the basin B over which the pollutant concentration on a given day did not exceed level z . This proportion, written as a function of the level z , is denoted $F_B(z)$ with the time notation suppressed. In the context of a random field model, $F_B(z)$ is a random cumulative distribution function which may be expressed as

$$F_B(z) = \int_{x \in B} I(Z(x) \leq z) d\mu(x) , \quad (7)$$

where, as before, $Z(x)$ is the stationary random field, I is the

indicator function, μ is Lebesgue or area measure, and $\mu(B)$, the total area of the basin, is taken to be unity.

Now we wish to estimate the realization of F_B rather than the stationary F . The available data are still the same N point observations $Z_1, Z_2, \dots, Z_N$, and we still are looking for a weighted estimator of the form (1) which is here denoted by $\hat{F}_B$. The loss criterion we now use, analogous to L defined in (2), is

$$L_B = \int_0^1 \frac{[\hat{F}_B(z_p) - F_B(z_p)]^2}{p(1-p)} dG(p),$$

where z_p still denotes the p -quantile of F . By substituting expressions (1) and (7) into the loss criterion and taking the mean value over the random field it follows that

$$E[L_B] = E[L] - 2 \sum_{i=1}^N w_i \bar{R}_{iB} + \bar{R}_{BB}, \quad (8)$$

where $E[L]$ is given in (3), $\bar{R}_{iB}$ is the averaged dichotomous correlation between the observation Z_i and the pollution concentration value at every (unobserved) point in B , and $\bar{R}_{BB}$ is the averaged dichotomous correlation between the values at every pair of (unobserved) points in B . More explicitly

$$\begin{aligned} \bar{R}_{iB} &= \int_0^1 \int_{x \in B} R_{ix}(p) d\mu(x) dG(p), \\ \bar{R}_{BB} &= \int_0^1 \int_{x \in B} \int_{y \in B} R_{xy}(p) d\mu(y) d\mu(x) dG(p), \end{aligned} \quad (9)$$

$$R_{ix}(p) = \{P[Z_i, Z(x) \leq z_p] - p^2\} / p(1-p),$$

$$R_{xy}(p) = \{P[Z(x), Z(y) \leq z_p] - p^2\} / p(1-p).$$

The set of observation weights $w_1^B, \dots, w_N^B$ which minimize the criterion $E[L_B]$ for this new problem will now depend on the $\bar{R}_{iB}$. However, the problem remains one of minimizing a quadratic form in the w 's subject to $\sum w_i = 1$ and $w_i \geq 0$ all i .

If the data are in the form of m independent time replications of observations at a fixed network of n monitoring stations and if B represents the whole air basin on a current day, then only the current day observations can have positive weight, although all the data may be used for the required estimates of correlation structure. Suppose that $Z_1, \dots, Z_n$ are the current data and that the stations are fairly evenly spaced throughout the basin, then we may approximate the integrals

$$\int_{x \in B} R_{ix}(p) d\mu(x) \text{ by } \sum_{j=1}^n R_{ij}(p)/n \text{ for } i = 1, \dots, n \quad (10)$$

when this approximation is substituted into the expression (8) for the concave loss criterion $E[L_B]$, it can be shown that

$$\frac{\partial E[L_B]}{\partial w_i} = 0 \text{ when } w_i = 1/n \text{ for } i = 1, \dots, n$$

Hence, equal weighting yields the unconstrained minimum of the loss criterion and, a fortiori, also the constrained minimum subject to $\sum w_i = 1$ and $w_i \geq 0$ all i . This leaves us with the following mildly surprising conclusion: for the problem of estimating the random spatial distribution function F_B over a domain B , the usual edf of the current data is approximately optimal in the class of weighted estimators whenever the approximation (10) is reasonable, e.g., for evenly spaced stations whatever may be the correlation pattern of the data.

There is another interesting special case which arises when dichotomous correlations are modeled as a function v of geographic separation in B and are found to decay fairly rapidly to zero.

In particular, suppose that correlations vanish at distances greater than d . Let A_d denote a circular region of radius d centered at the origin. Then

$$\int_{x \in B} R_{ix}(p) d\mu(x) = \int_{x \in A_d} v(\|x\|) d\mu(x) \quad \text{for all } i = 1, \dots, n,$$

provided that no observation station is too close to the boundary of B . Then the quantities $\bar{R}_{iB}$ defined in (9) are also all equal for $i = 1, \dots, n$, and the minimization of $E[L_B]$ in (8) is exactly equivalent to the minimization of $E[L]$ in (4) of Section 2. So for the situation described in this paragraph the appropriate station weights for estimating the random distribution function F_B are the same as those derived for estimating the stationary distribution F .

In the case of the 14-station air monitoring network, it appears that correlations are not a simple function of distance and that substantial correlations persist even between the most distant stations. Hence, the estimate of F_B could be quite different from the estimate of F . However, the network distributes the stations fairly evenly throughout the air basin; so by our earlier remarks the appropriate estimate of F_B is approximately the usual edf of the current observations. Finally, we mention that the random distribution function F_B , as defined in (7), is well-defined even in the case where $Z(x)$ is a non-stationary random field. Therefore, the problem of estimating F_B in a possibly non-stationary case is well-posed, even though the estimation of a stationary F may not be meaningful.

ACKNOWLEDGMENTS

This work was supported by a grant from SIAM Institute for Mathematics and Society (SIMS) to the Stanford University Statistics Department funded by the Rockefeller Foundation, the Sloane Foundation, the Energy Research and Development

Administration (US), the Environmental Protection Agency (US), and the National Science Foundation (US).

BIBLIOGRAPHY

- (1) Bay Area Air Pollution Control District (1974). Contaminant and Weather Summary. San Francisco.
- (2) Cochran, W.G. (1963). Sampling Techniques. New York: John Wiley & Sons, Inc.
- (3) David, M. (1976). Geostatistical Ore Reserve Estimation. Montréal: Ecole Polytechnique de Montréal.
- (4) Mathéron, G. (1976). Forecasting block grade distribution. Advanced Geostatistics in the Mining Industry. Dordrecht: Reidel.
- (5) Puri, M.L., and Sen, P.K. (1971). Nonparametric Methods in Multivariate Analysis. New York: John Wiley & Sons, Inc.

(Key words: random fields, distribution functions, air quality data)

ABSTRACT

A class of weighted estimators of a stationary distribution function is introduced which exploits correlation structure in the data as frequently found in air quality measurements. The nature of these weighted estimates is examined for several simple cases, in particular for spatially clustered measurements where the advantage over the usual empirical distribution function seems most pronounced. An illustration is presented using total suspended particulate concentrations from a 14-station monitoring network. The class of weighted estimators is also proposed for the related problem of estimating the actual realization of a spatial distribution function separately at each fixed time period.