MASTER

LA-8508-PR
Progress Report
UC-51
Issued: August 1980

Geostatistics Project of the National Uranium Resource Evaluation Program

October 1979—March 1980

K. Campbell
T. R. Bement
J. A. Howell
R. J. Beckman
K. Jackson
P. Buslee

GEOSTATISTICS PROJECT OF THE NATIONAL URANIUM RESOURCE EVALUATION PROGRAM

OCTOBER 1979 - MARCH 1980

by

K. Campbell, T. R. Bement, J. A. Howell, R. J. Beckman, K. Jackson and P. Buslee

ABSTRACT

During the period covered by this report, the authors investigated the serial properties of aerial radiometric data. Results were applied to the choice of minimum segment width in the maximum variance segments algorithm and to the use of aerial radiometric data in the design of ground sampling experiments.

The report also presents the results of a comparison of normal and lognormal percentile estimation techniques.

Twenty-two quadrangles are being analyzed in the search for a uranium favorability index. Computer codes developed during this investigation have been provided to the Bendix Field Engineering Corporation in Grand Junction, Colorado.

I. INTRODUCTION

This report outlines the activities and progress of the Los Alamos Scientific Laboratory (LASL) on the Geostatistics project during the first half of FY1980. The Geostatistics project is part of the National Uranium Resource Evaluation (NURE) program sponsored by the US Department of Energy (DOE), Grand Junction, Colorado, office. The NURE program is designed to assess the potential uranium resources throughout the conterminous United States and Alaska. In close cooperation with the Grand Junction Office of DOE, the Geostatistics project at LASL applies statistical methods to the analysis of data collected by airborne instrumentation. To handle a broad range of problems

related to the NURE, LASL maintains a close statistical consulting relationship with the DOE Grand Junction Office and the Bendix Field Engineering Corporation (BFEC) in Grand Junction.

We developed a statistical model for the correlation between observations taken along a flight line and considered two applications. The first involves the choice of a minimum segment width for the maximum variance segments algorithm being tested by BFEC. The second application involves the use of aerial radiometric data to aid in the design of ground-based sampling experiments. This last application developed as a result of a joint BFEC-LASL effort to design a statistically valid ground-based experiment in East Texas.

Standard deviation maps that are produced as part of the aerial program are closely related to the estimation of percentiles in the normal and lognormal probability distributions. This, plus the fact that anomalies are often defined explicitly or implicitly in terms of percentiles, motivated a study to determine the effect of using the wrong distribution (normal instead of lognormal or vice versa) when estimating percentiles.

Twenty-two quadrangles with assigned uranium favorabilities are being analyzed in the search for a uranium favorability index. Computer codes developed during this investigation have been transferred to BFEC.

II. NOTES ON THE TWO "MAXIMUM VARIANCE SEGMENTS" ALGORITHMS

In Ref. 1 Bement and Waterman give (1) an algorithm for determining the largest sum $S_k(N)$ of the variances of up to k disjoint segments from the data $x_1, x_2, \ldots, x_N$, and (2) a much simpler algorithm for determining $S(N) = \sup S_k(N)$.

The second algorithm is incorrectly stated; it should read

$$S(0) = 0$$

$$S(j) = \max_{1 \le \ell \le j} \{ S(\ell-1) + v(\ell,j) \},$$

where $v(j,j) = 0$ by definition. When modified to allow only segments of length between w and W, this algorithm becomes

$$S(j) = 0, \quad 0 \le j < w$$

$$S(j) = \max\{S(j-1), \max_{\ell_1(j) \le \ell \le \ell_2(j),} [S(\ell-1) + v(\ell,j)]\}, \quad w \le j \le N$$

where $\ell_1(j) = \max(1, j - W + 1)$ and $\ell_2(j) = j - w + 1$.

Given $S(j)$, $1 \le j \le N$, plus, for each j, a flag $L(j)$ defined by

$$L(j) = \begin{cases} 0 \text{ if } S(j) = S(j-1) \\ \ell \text{ if } S(j) = S(\ell-1) + v(\ell,j), \end{cases}$$

it is possible to find the segments which went into S(N), the maximum possible total sum of variances for segments of the data. This is the algorithm implemented by a program that BFEC is using.

Similarly, given $S_k(j)$, $1 \le j \le N$, as computed by the first algorithm, and a flag $L_k(j)$ defined by

$$L_k(j) = \begin{cases} 0 \text{ if } S_k(j) = S_{k-1}(j) \\ \ell \text{ if } S_k(j) = S_{k-1}(\ell-1) + v(\ell,j), \end{cases}$$

one can work backwards to find the best k segments. This algorithm is much harder to implement, however, and requires either much more computer time or a large random-access memory such as "extended core" on the Control Data Corporation machines.

These two algorithms select segments of data on different principles, and there is no reason to suppose that the k segments with largest variance produced by the second algorithm will be the k segments that go into $S_k(N)$ as computed by the first algorithm. In fact, unless k is very large, on the order of N/w, this will hardly ever be the case.

The first algorithm picks segments to maximize the total variance achieved by exactly k segments. The second algorithm, on the other hand, maximizes the total variance achievable by any number of segments. For noisy data, the variance is relatively independent of the length of the segment, and therefore the way to maximize

$$\sum_{j=1}^{k} v[\ell(j), j]$$

is to pick as many segments as possible in $x_1, \ldots, x_N$, which means picking them as short as possible. For such data, therefore, the second algorithm almost invariably selects segments of the minimum length, w, and never gets beyond w+2 or w+3.

Interestingly enough, although the first algorithm has no such built-in bias toward short segments provided $k << N/w$, it also appears to choose segments of the minimum length more frequently than expected, given that the expected variance of the segment increases with the length of the segment. (This is for correlated data and an assumption that we have something besides noise, although the noise is large). However, it chooses somewhat different segments (although it identifies, in a general way, the same regions pulled out by the second algorithm). As the minimum segment width, w, increases, the first algorithm appears to have a greater tendency to choose segments of length greater than w, and the sum of the variances of the top k segments found by the second algorithm becomes a smaller fraction of $S_k(N)$.

Table I shows the top ten segments in a string of $n = 560$ observations from one geological type selected by the first algorithm, where the length of the segment was restricted to lie between 5 and 15 seconds of data. Table II shows the results from the second algorithm, similarly restricted. Roughly

similar areas are identified, but the chosen segments are not identical, and the second algorithm has clearly made a suboptimal choice, with total variance of 1103.1 in ten segments, compared with $S_{10}(N)$ = 1177.0 computed by the first algorithm. In particular, the second algorithm appears to have missed a fairly significant segment, 467-471, chosen sixth by the first algorithm. It also selected two segments, 253-257 and 260-264, near the first algorithm's fourth choice, 256-260.

Tables III and IV represent a similar comparison for segments between 10 and 25 seconds in length. Here it is interesting to note that the first algorithm occasionally selects a very long segment, something which is almost impossible for the second algorithm. The second algorithm achieves a total variance of 790.4 in ten segments, compared with $S_{10}(N)$ = 887.2.

When applied to the NURE data, these discrepancies are probably not important, as long as the two algorithms identify approximately the same areas as interesting. There is some advantage to having all of the segments of about the same length, because then their variances are directly comparable, and we can justify picking off the top k, without worrying about the fact that the expected value of the variance increases with length.

TABLE I

SEGMENTS SELECTED BY FIRST ALGORITHM

(w = 5, W = 15)

Endpoints	Variance	Length
548 - 553	199.1	6
69 - 73	125.9	5
19 - 23	122.4	5
256 - 260	121.6	5
554 - 560	111.8	7
467 - 471	108.5	5
232 - 236	100.5	5
168 - 172	96.9	5
119 - 123	96.7	5
1 - 5	93.5	5

Note: Total variance = $S_{10}(N)$ = 1177.0.

TABLE II

SEGMENTS SELECTED BY SECOND ALGORITHM

Endpoints	Variance	Length
549 - 553	188.1	5
72 - 76	119.3	5
554 - 560	111.8	7
253 - 257	105.8	5
231 - 235	100.2	5
260 - 264	97.0	5
168 - 172	96.9	5
119 - 123	96.7	5
16 - 20	93.7	5
1 - 5	93.5	5

Note: Total variance = 1103.1

TABLE III

SEGMENTS SELECTED BY FIRST ALGORITHM

(w = 10, W = 25)

Endpoints	Variance	Length
551 - 560	159.4	10
69 - 78	97.5	10
253 - 262	86.4	10
39 - 53	84.5	15
162 - 173	81.9	12
404 - 428	78.8	25
19 - 36	77.8	18
117 - 126	75.8	10
224 - 235	73.9	12
1 - 10	71.4	10

Note: Total variance $S_{10}(N)$ = 887.2.

TABLE IV

SEGMENTS SELECTED BY SECOND ALGORITHM

(w = 10, W = 25)

Endpoints	Variance	Length
551 - 560	159.4	10
256 - 265	86.4	10
162 - 173	81.9	12
1 - 10	71.4	10
11 - 20	70.8	10
115 - 124	67.9	10
71 - 80	66.9	10
278 - 288	65.6	11
409 - 418	61.8	10
41 - 50	58.4	10

Note: Total variance = 790.4.

III. SERIAL PROPERTIES OF THE RADIOMETRIC DATA

In this section we describe a statistical model for the correlation between observations taken along a flight line. Physically the model is over-simplified, failing to take into account much of what is known about the processes of decay, gamma emission, scattering and measurement. However, a basic statistical model of this type will help us to see how the aerial data is related to what is on the ground, and in particular, what kinds of information might be extracted from the data. We consider applications related to the choice of a *minimum* "bin" or segment width for the *maximum* variance segments algorithm and the prior estimation of the components of variance in a ground-based survey.

A. The Statistical Model

Suppose that the gamma emissions (at a given frequency or in a band of frequencies) at a point (x,y) on the ground at time t have intensity $X(x,y,t)$ counts per second per unit area. Let $EX(x,y,t) = \lambda(x,y)$, independent of time, where $\lambda(x,y)$ is presumably proportional to the concentration of the source element or elements at (x,y). In turn, $\lambda(x,y)$ is assumed to be a realization of a spatial random process $\Lambda(x,y)$, and the principal task of the following

development is to see how the properties of the Λ process (e.g., its covariance function) are reflected in the aerial data.

(The notion that the concentration $\lambda(x,y)$ at a given point is a "realization" of a stochastic process is a statistical fiction which is very useful in practice. For example, it underlies the work of Matheron, et al. (Ref. 2) and is similar to models used in other branches of statistics, for example, in sampling theory. There is nothing in the discussion which follows, however, that makes its use imperative, and the reader may prefer to think of the quantities which we denote using E (expected values) as simply the appropriate spatial averages.)

The integrals of X over a finite region and time interval,

$$\int_t^{t+\Delta t} \iint_S X(x,y,\tau)\, dx\, dy\, d\tau,$$

have Poisson distributions with mean

$$\Delta t \iint_S \lambda(x,y)\, dx\, dy.$$

Letting S shrink to zero and $\Delta t \to 0$ implies that

$$E[X(x,y,t)|\Lambda=\lambda] = \lambda(x,y) \tag{1}$$

$$Var[X(x,y,t)|\Lambda=\lambda] = \lambda(x,y), \tag{2}$$

although X itself is not Poisson (it does not have units of counts).

The number of photons reaching the airborne detector during the time interval $(t-\Delta t)$ (which are recorded at time t) is a random quantity

$$K(t) = \iint_{-\infty}^{\infty} \int_{t-\Delta t}^{t} p(x,y)\, X(vs-x,-y,s)\, ds\, dx\, dy + B(t)$$

$$= I(t) + B(t), \tag{3}$$

where $p(x,y)$ is the instrumental point spread function, usually modeled as the product of a frequency-dependent exponential term decreasing with altitude and a geometrical factor. The detector is assumed to be traveling with velocity, v, along the x-axis ($y=0$). $B(t)$ is radiation from cosmic sources, from the aircraft, and so forth.

Of the $K(t)$ photons, only

$$Z(t) = \varepsilon K(t) + \eta(t) \tag{4}$$

are actually recorded, where ε is the efficiency of the detector and η is a random component. This recording process might be modeled as a binomial, $Z(t) \sim \mathcal{B}[K(t),\varepsilon]$, or a Poisson process, $Z(t) \sim \mathcal{P}[\varepsilon K(t)]$.

We begin by assuming that $\Lambda(x,y)$ is a second-order stationary process, which means that

$$E\Lambda(x,y) = m = \text{constant for all } (x,y), \text{ and}$$

$$\begin{aligned} \mathrm{Cov}[\Lambda(x,y),\ \Lambda(x+\Delta x,\ y+\Delta y)] &= E\{[\Lambda(x,y) - m][\Lambda(x+\Delta x, y+\Delta y) - m]\} \\ &= C_{\Lambda}(\Delta x, \Delta y), \end{aligned}$$

a function which depends only on the separation $(\Delta x,\Delta y)$ of two points, not on the absolute position (x,y). If the background $B(t)$ is also a stationary process, then so is $Z(t)$. In particular,

$$\begin{aligned} EZ(t) &= E[I(t) + B(t)] \\ &= m\,P_v + EB(t), \end{aligned}$$

where

$$P_v = \int_{-\infty}^{\infty}\int_{0}^{\Delta t} p(x-vs,y)\, ds\, dx\, dy. \tag{5}$$

We next show how $C_{\Lambda}(\Delta x,\Delta y)$ is related to

$$C_Z(h) = \mathrm{Cov}[Z(t),\ Z(t+h)].$$

First we consider

$$C_X(\Delta x,\Delta y) = \mathrm{Cov}[X(x,y,t),\ X(x+\Delta x,y+\Delta y,t+\Delta t)].$$

(C_X will in fact turn out to be independent of Δt, as well as of x, y and t.) By definition, this is

$$E\{[X(x,y,t) - EX(x,y,t)]\,[X(x+\Delta x,y+\Delta y,t+\Delta t) - EX(x+\Delta x,y+\Delta y,t+\Delta t)]\}.$$

Conditioning on $\Lambda(x,y)$, this can be written as the sum of two terms,

$$E\Big(E\{[X(x,y,t) - \Lambda(x,y)][X(x+\Delta x,y+\Delta y,t+\Delta t) - \Lambda(x+\Delta x,\ y+\Delta y)]\,|\Lambda\}\Big)$$

$$+ \mathrm{Cov}\{E[X(x,y,t)|\Lambda],\ E[X(x+\Delta x,y+\Delta y,t+\Delta t)|\Lambda]\}. \qquad (6)$$

As $E[X(x,y,t)|\Lambda] = \Lambda(x,y)$, the second term in Eq. (6) is just $C_\Lambda(\Delta x,\Delta y)$. In the first term, we can assume that the random variates $X(x,y,t)$ and $X(x+\Delta x,\ y+\Delta y,\ t+\Delta t)$ are independent, given Λ, unless $\Delta x = \Delta y = \Delta t = 0$. (That is, the deviation of X from its conditional expected value Λ at one point in space and time is independent of the deviation at a different point in space and/or at a different time.) When $\Delta x = \Delta y = \Delta t = 0$, the first term in Eq. (6) is

$$E\{\mathrm{Var}[X(x,y,t)|\Lambda]\} = E\Lambda(x,y) = m,$$

using Eq. (2). Thus,

$$C_X(\Delta x,\Delta y) = \begin{cases} C_\Lambda(0,0) + m & \Delta x = \Delta y = 0 \\ C_\Lambda(\Delta x,\Delta y) & \text{otherwise.} \end{cases} \qquad (7)$$

Next we consider

$$C_I(h) = \mathrm{Cov}[I(t),\ I(t+h)]$$

$$= \int_{-\infty}^{\infty}\int dx\,dy \int_{-\infty}^{\infty}\int d\xi\,d\eta \int_{t-\Delta t}^{t} ds \int_{t+h-\Delta t}^{t+h} d\tau\ \ p(x,y)\ p(\xi,\eta)$$

$$\mathrm{Cov}[X(vs-x,-y,s),\ X(v\tau-\xi,-\eta,\ \tau)].$$

Make the following changes of variables:

$$\begin{aligned} Z &= t-s \\ r &= x+vz \\ \zeta &= t+h-\tau \\ \rho &= x-\xi+vh-v\zeta \\ \sigma &= y\eta. \end{aligned}$$

The above integral then becomes

$$\int_{-\infty}^{\infty}\int dr\,dy \int_{-\infty}^{\infty}\int d\rho\,d\sigma \int_{0}^{\Delta t} dz \int_{0}^{\Delta t} d\zeta\ \ p(r-vz,y)\ p(r-\rho+vh-v\zeta,y-\sigma)$$

$$\mathrm{Cov}[X(vt-r,-y,t-z),\ X(vt-r+\rho,-y+\sigma,t+h-\zeta)].$$

The covariance is just $C_X(\rho,\sigma)$, independent of all of the other integration variables. Define

$$p_V(\alpha,\beta) = \int_{0}^{\Delta t} p(\alpha-v\tau,\beta)\ d\tau. \qquad (8)$$

Then after performing the integrals over z and ζ, what remains is

$$\int_{-\infty}^{\infty}\int d\rho d\sigma\ C_X(\rho,\sigma) \int_{-\infty}^{\infty}\int dr dy\ p_V(r,y)\ p_V(r+vh-\rho,\ y-\sigma).$$

Define the inner integral as $C_p(vh-\rho,-\sigma)$, which leads finally to

$$C_I(h) = \int_{-\infty}^{\infty}\int C_X(\rho,\sigma)\, C_p(vh-\rho,-\sigma)\, d\rho d\sigma$$

$$= (C_X * C_p)(vh,0), \tag{9}$$

where * denotes the convolution of two functions,

$$(f*g)(x,y) = \iint f(\zeta,\eta)\, g(x-\zeta,y-\eta)\, d\zeta d\eta.$$

Thus, the covariance function of I is equal to the covariance function of X (which in turn was essentially the covariance function of Λ, differing from the latter only at h=0, and this makes no difference in the integral) convolved with a function which depends on the point spread of the instruments.

Finally, from Eqs. (3) and (4), assuming that the signal from the ground, I(t), the background signal B(t), and the measurement noise $\eta(t)$ are all uncorrelated, we get

$$C_Z(h) = \mathrm{Cov}[Z(t),\, Z(t+h)]$$

$$= \begin{cases} \varepsilon^2[(C_\Lambda * C_p)(vh,0) + C_B(h)], & h \neq 0 \\ \varepsilon^2[(C_\Lambda * C_p)(vh,0) + C_B(h)] + E(\mathrm{Var}\ \eta(t)), & h = 0. \end{cases} \tag{10}$$

Some examples will make the nature of the terms in Eq. (10) clearer. The function C_p arose as a convolution-like integral,

$$C_p(x,y) = \int_{-\infty}^{\infty}\int p_v(\zeta,\eta)\, p_v(\zeta+x,\eta+y)\, d\zeta d\eta,$$

where

$$p_v(x,y) = \int_0^{\Delta t} p(x-vt,y)\, dt.$$

Several models for the point-spread function p have been proposed. Two are based on simple geometrical considerations (Ref. 3). They are not too different from one which appears in Ref. 4. Figures 1 and 2 are based on the "elementary rod" model in Ref. 3, normalized as

$$p(x,y) = p'(r)$$

$$= \exp(-\mu\sqrt{h^2 + r^2}) \cdot \frac{h}{4\pi(h^2+r^2)^{3/2}}. \tag{11}$$

In Eq. (11), $r^2 = x^2 + y^2$, h is the altitude of the aircraft and μ is a frequency-dependent linear coefficient. Figure 1a is a plot of $p(x,0) = p'(x)$ for h = 125 m and $\mu = 0.005\ m^{-1}$. Figure 1b is a plot of $p_v(x,0)$ for Δt = 1 second, v = 120 mph. The second x-axis converts units of distance to units of time at the given velocity. Figure 1c shows the corresponding function $C_p(x,0)$. Figures 2a and 2b are for a velocity of 70 mph, more typical of helicopter speeds.

In general, we can expect any signal, such as Λ or B, to have what we will call both a "continuous" component and a noise component. By this we mean that the covariance function has a discontinuity at the origin, for example,

$$C_\Lambda(0,0) \neq \lim_{\Delta x, \Delta y \to 0} C_\Lambda(\Delta x, \Delta y). \tag{12}$$

$C_\Lambda(0,0)$ is the total variance of Λ. In the case of a geological process, such as Λ, it may be hard to visualize a sharp discontinuity in any realization λ, but it is often the case that estimates of $C_\Lambda(\Delta x,\Delta y)$ strongly suggest that Eq. (12) holds. The reason for this is probably that λ has not been sampled finely enough. The small-scale structure in λ, if it were studied, might lead to a smooth extrapolation of C_Λ to zero (apart from negligible measurement error), but when samples are available only at greater distances, the

estimated autocovariance function appears to have a large discontinuity or "nugget" at the origin. In the case of the background signal, this discontinuity is due to the random Poisson nature of the signal. In any case, their covariance functions have in general a shape such as shown in Fig. 3, where the central spike has zero width in the case of C_B or small but finite width in the case of C_Λ. The "continuous" component of Λ corresponds to the remaining positive portion of the function,

$$C_\Lambda(\Delta x,\Delta y) = C_\Lambda(0) \quad \mathrm{Corr}[\Lambda(x,y),\ \Lambda(x+\Delta x, y+\Delta y)],$$

and is positive at distances over which the Λ process is positively correlated. That is, nonzero correlation over short distances is taken to represent some degree of continuity in the underlying process.

The following data examples are taken from the single record reduced count rates for the Rawlins quadrangle, a segment which was flown by helicopter over the tertiary Browns Park formation (TBP). These reduced count rates are obtained after considerable manipulation of the raw window count rates, the quantities that might reasonably be modeled by Z(t) in Eq. (4). However, perhaps because the reduction procedures are linear operations, even these data show the structure anticipated in Eq. (10).

The thallium record (Fig. 4a) has little structure apart from some short trends over distances on the order of two kilometers. The total variance of this segment [$C_Z(0)$ in the notation of Eq. (11)] is 47.3, while the next points in the autocovariance estimate (Fig. 4b) drop down to something on the order of 13 to 16. Thus, the noise term [E(Var η) in Eq. (10)] has a magnitude of about 32, and the signal in Fig. 4a is more than two-thirds noise, which certainly agrees with the visual impression. Part of this is undoubtedly due to a pure noise component in $C_B(h)$; that is, we suspect that the background signal is partially uncorrelated from second to second, although in part it comes also from highly continuous sources, such as the "bismuth-air" signal. The continuous part of the covariance, the $C_\Lambda * C_p$ term (which is always continuous as $h \to 0$ even though C_Λ may not be continuous by itself) and the continuous part of $C_B(h)$, are reflected by the decline of the sample autocovariance function from 13-16 at $h \sim 0$ to zero at about 2.0 km. As 2.0 km is considerably greater than the width of C_p, the positive correlation out to this distance must be ascribed either to some underlying correlation in Λ

out to distances of 1.5 km or greater, or to the continuous (correlated) part of the background signal, or both. (The effect of convolving C_p with C_Λ is to broaden C_Λ by about the half-width of C_p, which was not more than 0.5 km; for this reason, if $C_\Lambda * C_p$ is positive out to 2.0 km, C_Λ must have been positive out to at least 1.5 km.) At the moment we have no way of estimating the relative contributions of the two terms to the continuous part of C_Z. For this purpose, it will be necessary to do some analyses of the "bismuth-air" and "cosmic" signals, as well as to investigate the details of GeoMetrics' instrument and aircraft correction procedures.

The reduced potassium record for the same segment is shown in Fig. 5a. The visual impression given by these data is that the signal-to-noise ratio is higher than in Fig. 4a, and this is confirmed by Fig. 5b, which shows that $\lim_{h \to 0} C_Z(h)$ is about 600, more than 80% of the total variance $C_Z(0)$. The potassium signal is also apparently positively correlated out to at least 3.0 km.

The reduced bismuth signal is interesting because of the sharp jump at about 9.0 km (Fig. 6a). This presumably reflects a point anomaly on the ground, and the effect of such a discontinuity is to add a significant contribution to the underlying discontinuous part of C_Λ. When this spike at the origin is convolved with C_p the result has the shape of C_p, and this is clearly visible as a hump in Fig. 6b. The noise component in this case is about 30.0 (about half of the total). From about 34.0, sample autocovariance function decreases to about 23.0, at 0.5 km, with a shape like that seen in Fig. 2b. The continuous part of $C_Z(h)$ extends farther, to perhaps 3.5 km. This extension, again, comes from the continuous part of C_Λ convolved with C_p and/or the continuous part of C_B.

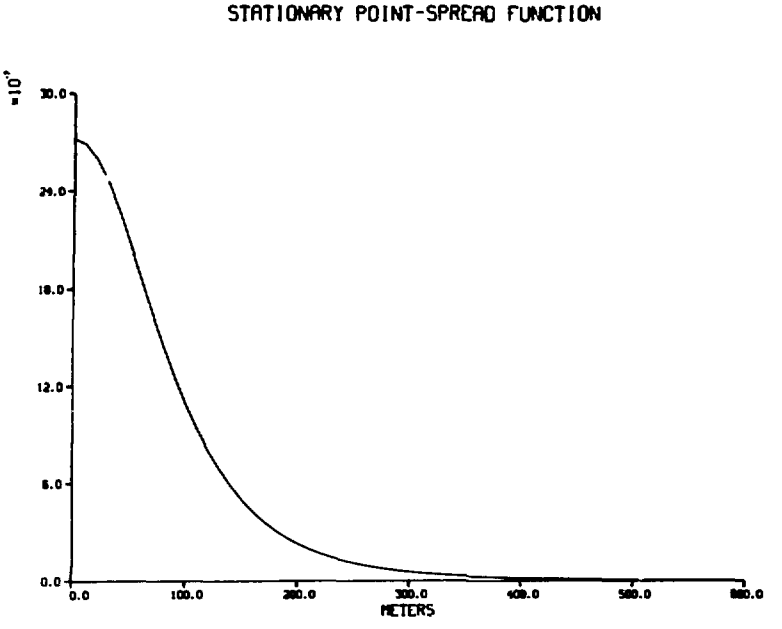

Fig. 1a. One model for an instrumental point-spread function.

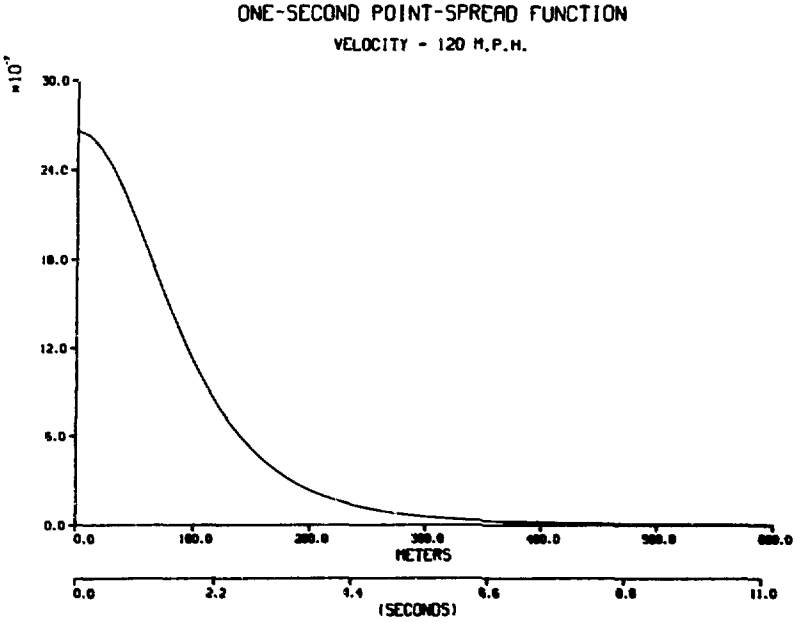

Fig. 1b. Widening of point-spread function due to integration over one second at 120 mph.

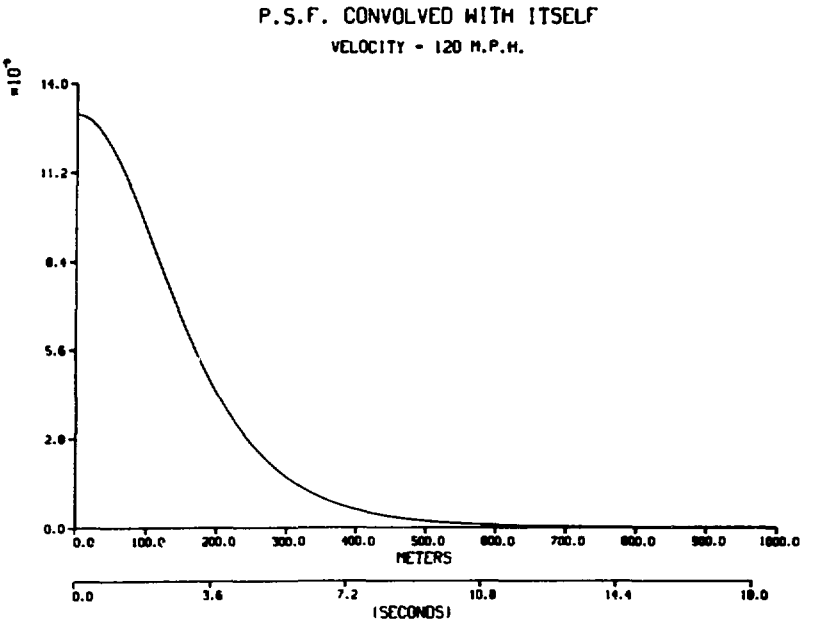

Fig. 1c. Convolution of the point-spread function with itself (120 mph).

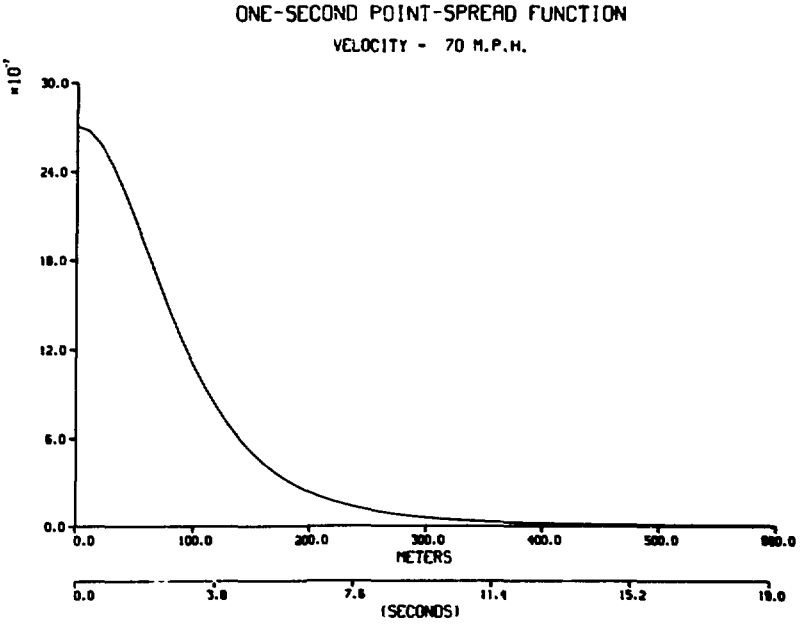

Fig. 2a. Widening of point-spread function at 70 mph.

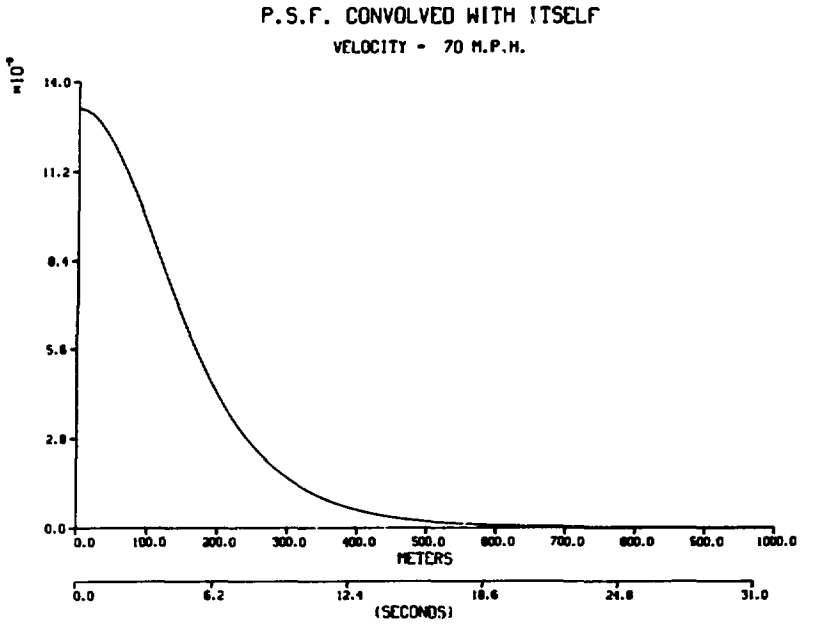

Fig. 2b. Convolution of point-spread function with itself (70 mph).

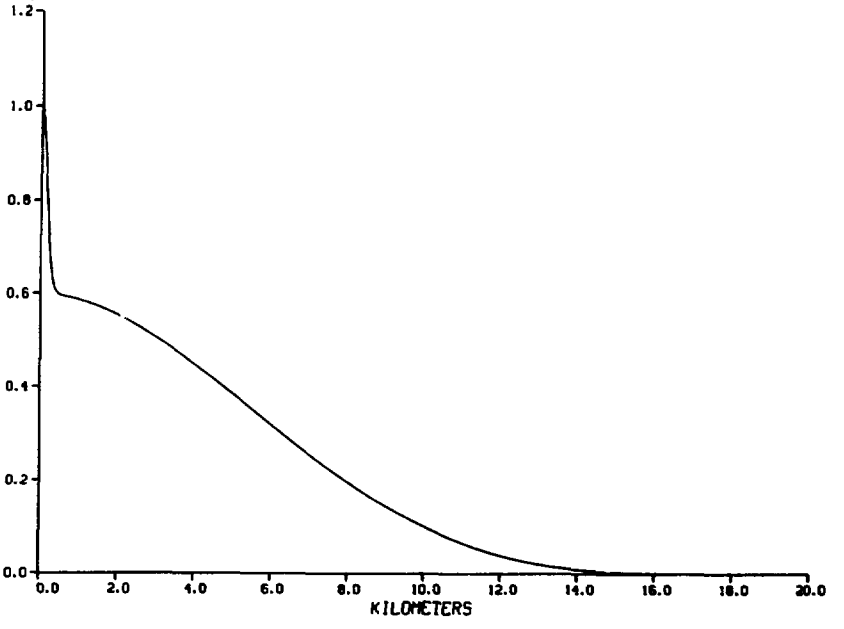

Fig. 3. Typical shape for autocovariance function of correlated,noisy observations.

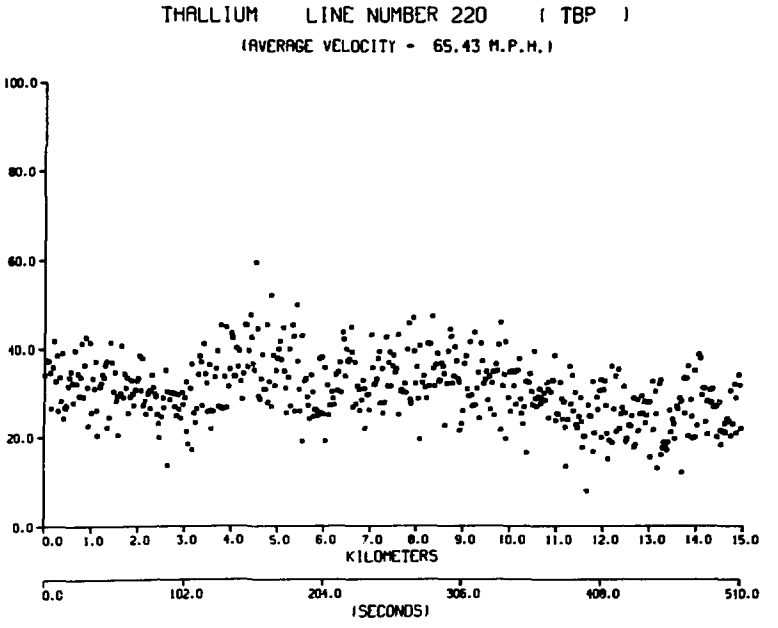

Fig. 4a. Single record reduced count data for ^{208}Tl.

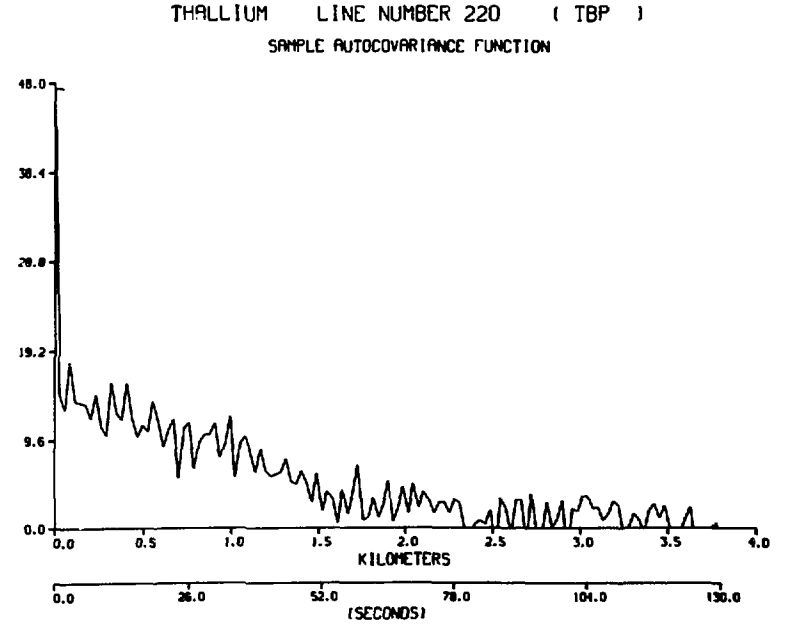

Fig. 4b. Sample autocovariance function for ^{208}Tl.

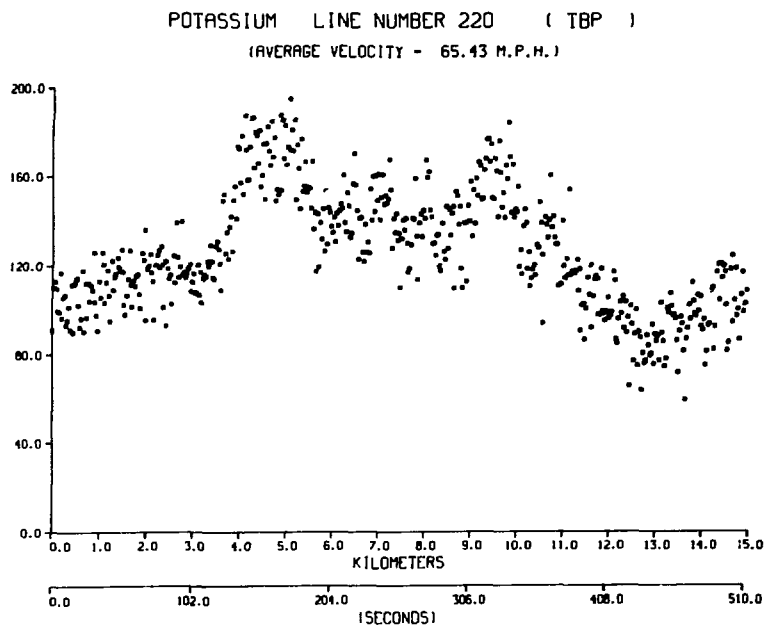

Fig. 5a. Single record reduced count data for ^{40}K.

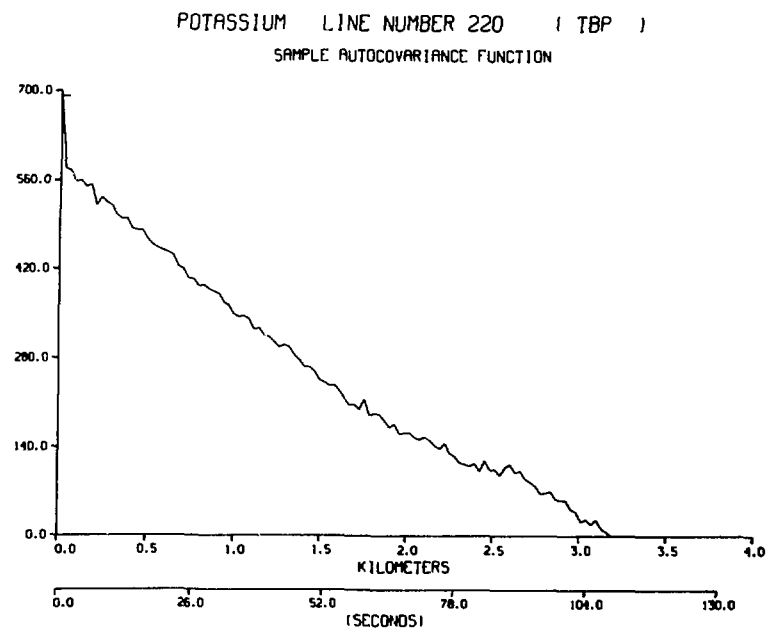

Fig. 5b. Sample autocovariance function for ^{40}K data.

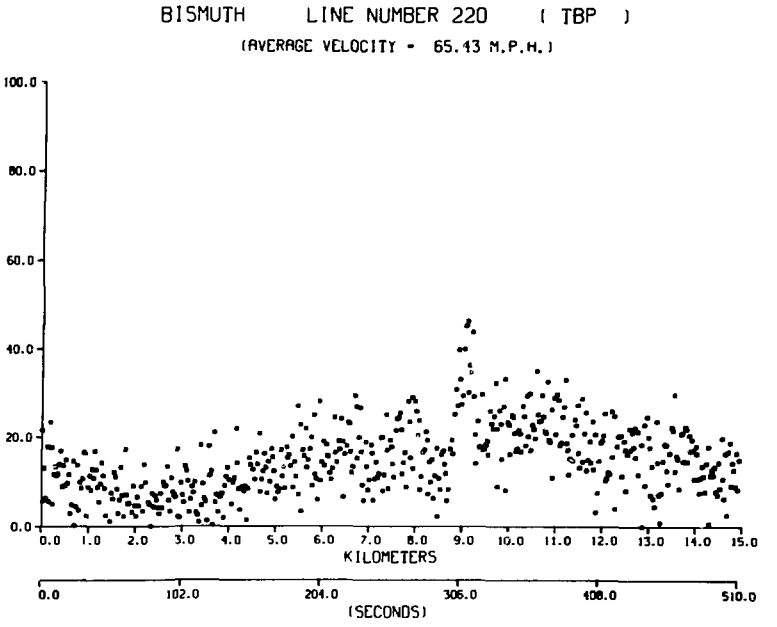

Fig. 6a. Single record reduced count data for ^{214}Bi.

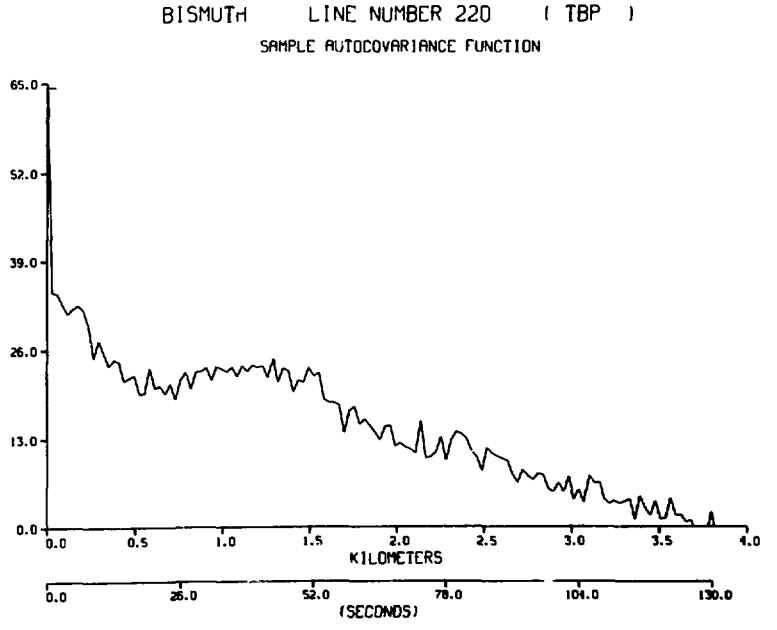

Fig. 6b. Sample autocovariance function for ^{214}Bi.

B. Other Descriptions of Serial Properties

There are a number of functions which are related to the autocovariance function $C_Z(h)$. One of the best known is the power spectrum,

$$S_Z(\omega) = \int_{-\infty}^{\infty} e^{-i2\pi\omega h} C_Z(h)\, dh \tag{13}$$

(also called the spectral density function). $S_Z(\omega)$ is estimated by the periodogram

$$I(\omega) = \frac{\Delta h}{N} \left| \sum_{n=1}^{N} Z(n\Delta h) e^{-i2\pi\omega n} \right|^2, \tag{14}$$

assuming that the process $Z(h)$ has been sampled at regular intervals Δh. It can be shown that

$$I(\omega) = \Delta h \sum_{k=-\infty}^{\infty} e^{-i2\pi\omega k} C_Z(k\Delta h),$$

in immediate analogy with Eq. (13), where $C_Z(k\Delta h)$ is the sample autocovariance function

$$C_Z(k\Delta h) = \begin{cases} \frac{1}{N} \sum_{n=1}^{N-|k|} [(Z(n\Delta h)-\bar{Z}][Z((n+|k|)\Delta h)-\bar{Z}] & 0 \leq |k| < N \\ 0 & |k| \geq N. \end{cases} \tag{15}$$

(These are the functions plotted in Figs. 4b, 5b and 6b, estimating $C_Z(h)$.) Under the Fourier transform of Eq. (13), Eq. (10) becomes

$$S_Z(\omega) = \varepsilon^2[S_\Lambda(\omega) \cdot S_p(\omega) + S_B(\omega)] + E(\text{Var } \eta).$$

$S_p(\omega)$ [the transform of $C_p(x,0)$] is shown in Fig. 7. Figure 8 is the periodogram of the data in Figure 4a. The transforms of C_Λ and C_B are much narrower than S_p, so the periodogram falls off rapidly to about 0.5 cycles

per km, where it is virtually down to the noise level (the constant E(Var n) times Δh; the product is slightly less than 1).

Another function, which will be useful in establishing a criterion for the minimum bias width in the maximum variance segments algorithm, is an integral of $C_Z(h)$ called the variance-length curve,

$$V_Z(L) = C_Z(0) - \frac{L}{L^2}\int_0^L (L-h)\, C_Z(h)\, dh. \tag{16}$$

It can be shown that $V_Z(L)$ is the expected variance of Z within a segment of length L,

$$V_Z(L) = E\{\frac{1}{L}\int_0^L [Z(t) - m_L]^2\, dt\}, \tag{17}$$

where

$$m_L = E\left[\frac{1}{L}\int_0^L Z(t)\, dt\right].$$

If Z(t) were a pure noise process, $V_Z(L)$ would be constant for all L. However, when Z is positively correlated over short distances, $V_Z(L)$ is an increasing function of L, where $\lim_{L\to 0} V_Z(L)$ is the noise level (the discontinuous part of Z) and $\lim_{L\to\infty} V_Z(L) = \text{Var } Z$.

The variogram

$$\gamma_Z(h) = \frac{1}{2}\text{Var}[Z(t+h) - Z(t)] \tag{18}$$

can be defined even if Z is not itself stationary, but does have stationary increments. When Z is stationary, then

$$\gamma_Z(h) = C_Z(0) - C_Z(h). \tag{19}$$

Figure 9 is the variogram for the data in Fig. 5, showing again the hump, now inverted, which arises from convolving C_p with the discontinuous part of C_A. Fig. 9 is not exactly Fig. 5b as inverted because the experimental variogram of Fig. 9 was computed as

$$\gamma_Z(n\Delta h) = \frac{1}{2(N-n)} \sum_{k=1}^{N-n} [Z(k\Delta h)-Z((k+n)\Delta h)]^2,$$

which is not quite $C_Z(0) - C_Z(n\Delta h)$.

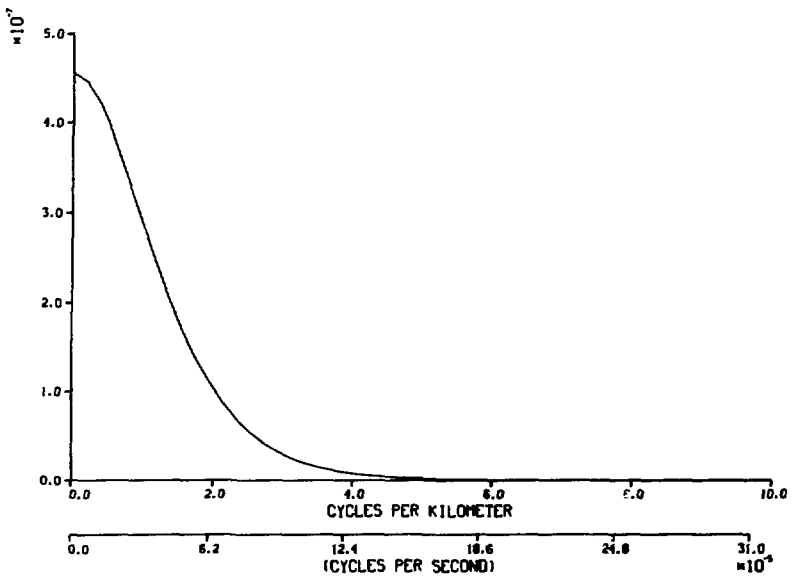

Fig. 7. Fourier transform of convolution of point-spread function with itself (70 mph).

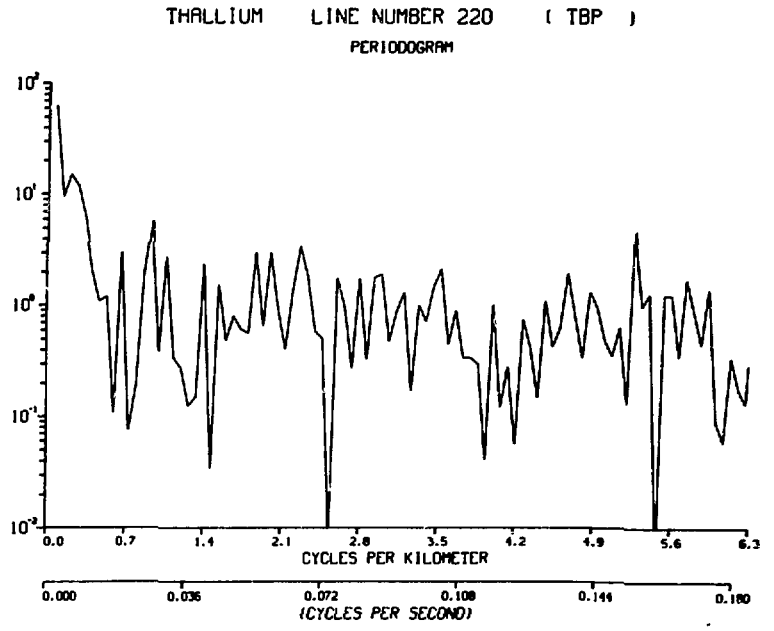

Fig. 8. Periodogram of ^{208}Tl data.

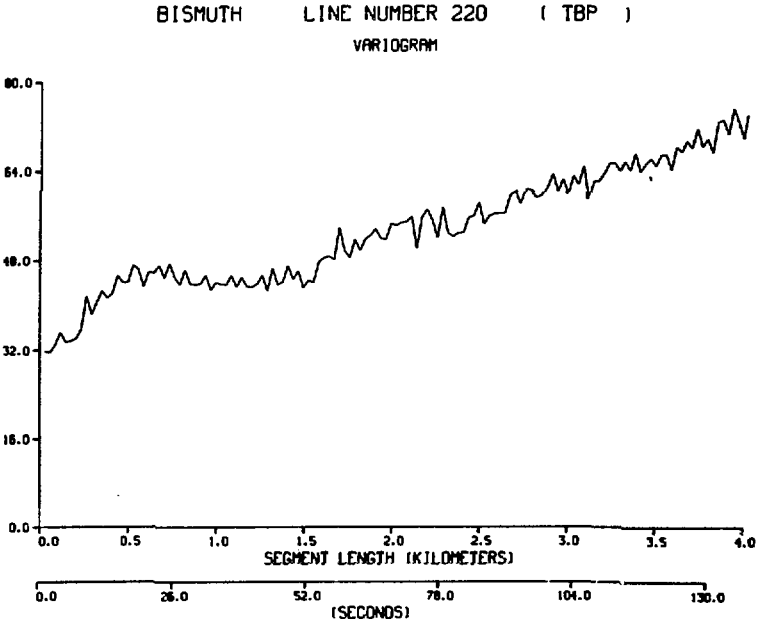

Fig. 9. Experimental variogram of ^{214}Bi data.

C. The Effect of a Moving Average

One method which is popular for increasing the signal-to-noise ratio in a segment such as shown in Fig. 4a is the use of a seven-point (sometimes five-point) moving average. This is equivalent to convolving Z(t) with a rectangular window

$$R(t) = \begin{cases} 1 & -a \leq t \leq a \\ 0 & \text{otherwise.} \end{cases}$$

For the data in Fig. 4a, a is approximately 0.1 km. The result of the convolution is shown in Fig. 10a. Indeed, the noise has been much reduced. Fig. 10b is the corresponding sample autocovariance function, which is given by

$$C_Z(h) * C_R(h),$$

where $C_R(h)$ is the convolution of R with itself, a triangular function of h whose half-width is about 0.2 km. The triangle is seen out to about 0.2 km in Fig. 10b, the result of convolution with the discontinuous part of $C_Z(h)$. The remainder of the autocovariance function is smoothed and broadened slightly by the convolution (for example, Fig. 4b).

However, the moving average introduces some distracting artifacts which are very noticeable in Fig. 10a, specifically, small oscillations on the order of 2 to 3 cycles per kilometer. The reason for this is clearer if we consider the "frequency" domain. The Fourier transform of the convolution of Z with R is the product of the Fourier transforms of Z and R. The transform of R is shown in Fig. 11. It is nearly zero for frequencies between about 4.5 and 5.25 cycles per kilometer, and thus multiplication with the transform of Z suppresses these frequencies in the product. As a result, the lower frequencies, relatively unmodified (compare Fig. 12 with Fig. 8) become distractingly apparent in the original data. For this reason a filter with a smoother "roll-off" in both the spatial and frequency domains would be preferable.

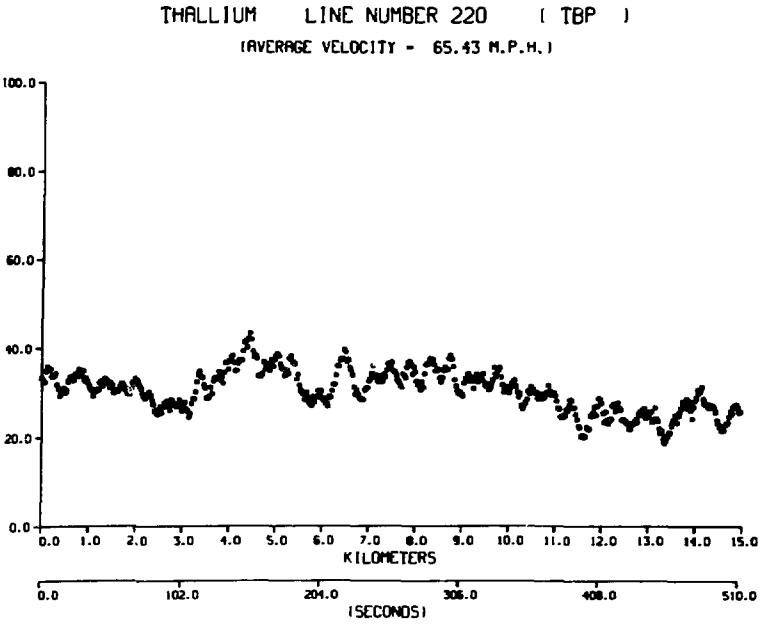

Fig. 10a. Result of applying a seven-point moving average to the $^{208}T\ell$ data.

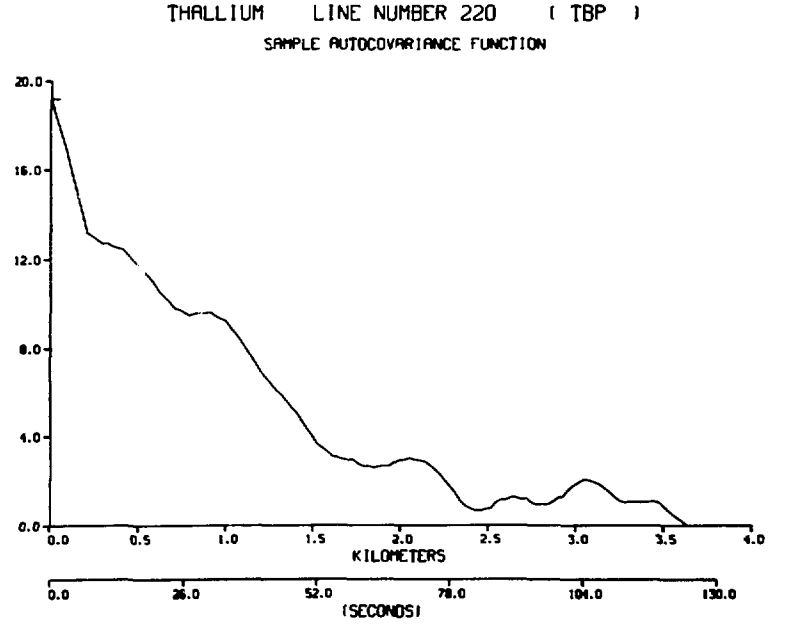

Fig. 10b. Sample autocovariance function of the averaged $^{208}T\ell$ data.

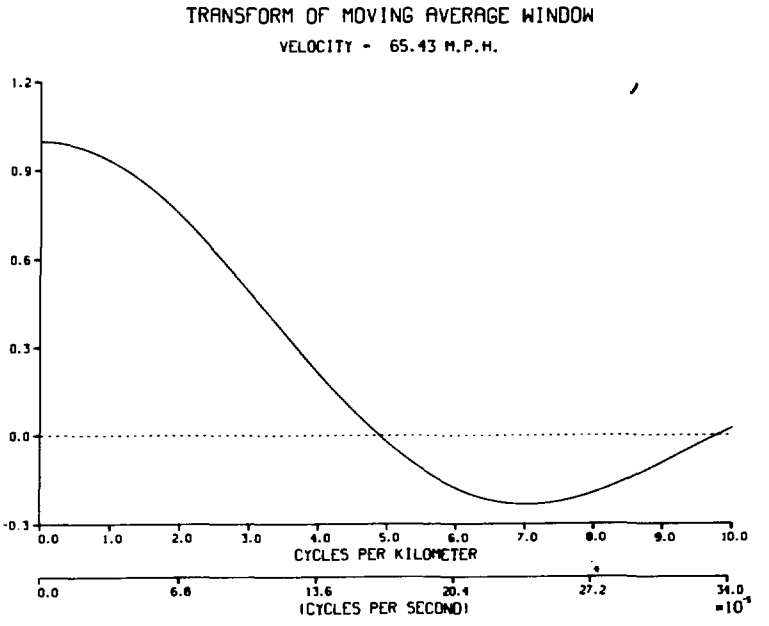

Fig. 11. Fourier transform of the seven-point moving average.

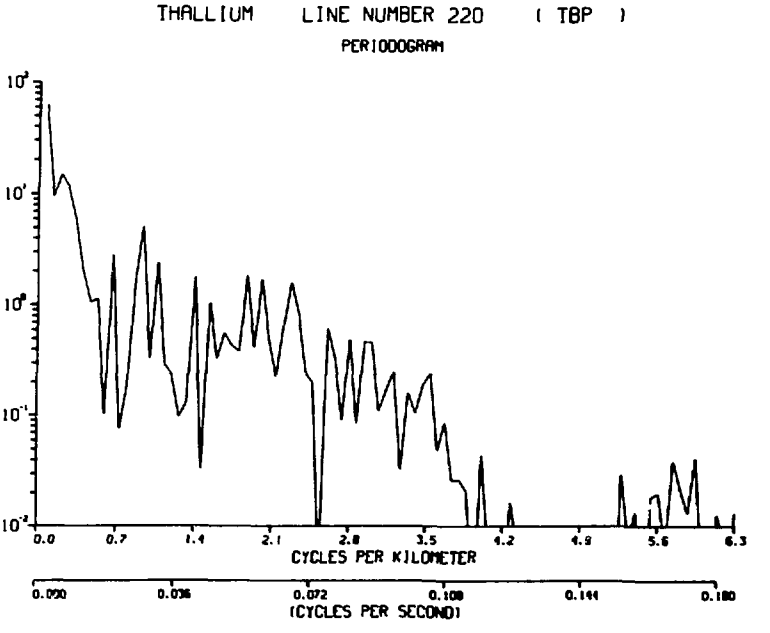

Fig. 12. Periodogram of the averaged ^{208}Tl data.

D. Selection of Minimum Length for "Maximum Variance Segments"

Given that the current maximum variance segments algorithm almost invariably produces segments of the present minimum length, w, the choice of w becomes a matter of some interest. Fortunately, some rational basis for this choice exists because of the correlation introduced into successive observations by the instruments. Each observation is a weighted average of the signal over a fairly large area on the ground beneath the plane. [The weighting function is the instrumental point spread function, one model of which was given in Eq. (11).] Even at 120 mph the plane does not move very far in one second, compared to the area over which it is integrating. As a result, even if the ground signal is completely uncorrelated from one point to another, *adjacent* records are highly correlated, so that significant variation over short segments is attributable almost entirely to noise added by the recording procedure, not to variation in the signal. Therefore, short segments should be excluded from consideration by the maximum variance segments algorithm.

For present purposes, the correlation introduced by the instruments is most conveniently described by the variance-length curve, $V_Z(L)$, defined in Eqs. (16) and (17). Let us rewrite $C_Z(0)$ as the sum of an instrumental noise variance component, V_N, and an incoming signal variance component, V_S, and set

$$C_Z(h) = V_S R(h), \quad h \neq 0,$$

where R(h) is the autocorrelation of two observations separated by a distance h (in the absence of additional measurement error). Then Eq. (16) becomes

$$V_Z(L) = V_N + V_S\left[1 - \frac{L}{L^2} \int_0^L (L-h)\, R(h)\, dh\right], \tag{20}$$

and the fraction of the variance over a segment of length L which can be ascribed to the signal (as opposed to the white noise added by the recording procedure) is

$$\frac{V_Z(L)}{V_N} = 1 + \frac{V_S}{V_N}\left[1 - \frac{2}{L^2} \int_0^L (L-h)\, R(h)\, dh\right]. \tag{21}$$

Now, referring back to Eq. (10) we see that even if $C_B(h)$ were zero for $h = 0$ (that is, the background signal were pure noise) and $C_A(\Delta x, \Delta y) = 0$ for nonzero Δx or Δy (that is, the concentration of the source on the ground is uncorrelated from one point to another), still the function $C_Z(h)$ would have nonzero values for $h \neq 0$, as it is proportional to $C_p(vh, 0)$, and $V_Z(L)$ would be an increasing function of L. This extreme case provides an upper bound on how fast $V_Z(L)$ can increase with L. Figure 13 shows plots of $V_Z(L)/V_N$ computed according to Eq. (21), where $R(h) = C_p(vh,0)/C_p(0,0)$ for three values of the overall signal-to-noise ratio V_S/V_N. Here C_p is based on Eq. (11) with a velocity of 75 mph.

A reasonable way to select the minimum segment width, w, might be to require that $V_Z(w)/V_N > 2$, because when $V_Z(L)/V_N < 2$, the variability of a segment of length L is more than 50% due to noise. Table V gives the value of w, in meters, for which $V(w)/V_N \sim 2$, read off the graphs of Fig. 13, for three values of the signal-to-noise ratio. In the third and fourth columns of Table V these values are converted to seconds for observations from fixed-wing aircraft and helicopters.

In particular, a minimum segment width of five seconds for the example discussed in the section comparing the two maximum variance algorithms (which is based on helicopter data, flown at about 75 mph) is certainly too small, as the signal-to-noise ratio is not greater than four.

If the signal on the ground enjoys some spatial autocorrelation, then the signal received by the instruments will be correlated out to larger distances and the variance-length curve, $V_Z(L)$, will rise more slowly than suggested by Fig. 13. Empirical variance-length curves suggest that this is indeed the case, and that it may take several kilometers for $V_Z(L)/V_N$ to rise above 2. However, this will be extremely variable from one string of data to another, and furthermore, it is precisely those cases where $V_Z(L)/V_N$ rises rapidly which are of interest. As computations based on the point-spread model give a lower bound on L such that $V_Z(L)/V_N = 2$, this seems to be a reasonable way to choose a lower bound on the length of segments to be considered.

TABLE V

APPROXIMATE MINIMUM BIN WIDTH, SELECTED BY REQUIRING

$V_Z(w)/V_N = 2$

Signal-to-Noise Ratio	Width (meters)	Width (seconds) At 125 mph	Width (seconds) At 70 mph
2	460	8-9	14-15
3	300	5-6	9-10
4	240	4-5	7-8

The total signal plus noise is traditionally estimated by the total variance of the string of observations $Z_1, Z_2, \ldots, Z_N$:

$$V_S + V_N = \frac{1}{N} \sum_{i=1}^{N} (Z_i - \bar{Z})^2. \tag{22}$$

Estimates of the signal variance, V_S, or the noise variance, V_N, are readily formed from plots such as Fig. 5b [use $\lim C_Z(h)$ for V_S] or Fig. 9 [use $\lim_{h \to 0} \hat{\gamma}_Z(h)$ for V_N]. As a single computed indicator one might use

$$V_S = \hat{C}_Z(\Delta h) = \frac{1}{N} \sum_{i=1}^{N-1} (Z_i - \bar{Z})(Z_{i+1} - \bar{Z}) \tag{23}$$

or

$$V_N = \hat{\gamma}_Z(\Delta h) = \frac{1}{2(N-1)} \sum_{i=1}^{N} (Z_i - Z_{i+1})^2. \tag{24}$$

An estimate of the signal-to-noise ratio V_S/V_N can be formed by any of the obvious combinations of Eqs. (22), (23) and (24).

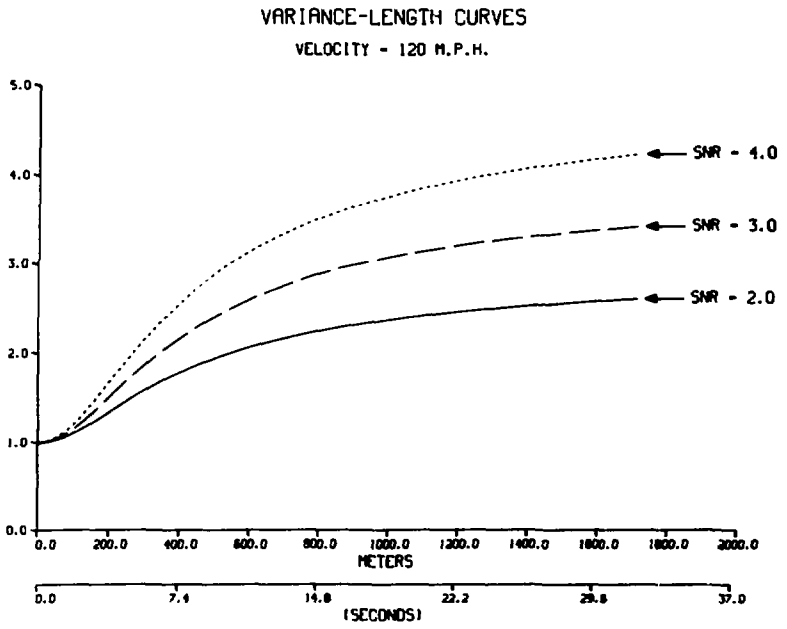

Fig. 13. Increase in the signal-to-noise ratio with segment length.

E. Relating the Components of Variance in Ground-Based and Aerial Measurements

An important problem in the design of a ground-based sampling experiment is the determination of an adequate sample size. Sample size determination requires information about the variance inherent in ground-based measurements. The suitability of aerial radiometric data as a source of this information will now be considered.

The ground-based measurements can be thought of as arising from a classical two-stage nested analysis of variance (ANOVA) design:

(1) several "samples" near one "position" and

(2) several "positions" along each "line."

Within this framework, the "samples" are assumed to vary independently with variance σ_S^2, and this effect is additive to the "position" effect with variance σ_P^2 and the "line" effect with variance σ_L^2.

Let $Q_k(x,y)$ denote the true concentration, in ppm, at a point (x,y) near the kth flight line. (x denotes the coordinate along the line, and y is small.) In the ANOVA model, Q is the sum of several components,

$$Q_k(x_{ij}, y_{ij}) = \mu + L_k + P_{ki} + S_{kij}, \tag{25}$$

where (x_{ij}, y_{ij}) is the location of the jth sample taken at the ith position, μ is the overall mean, taken to be constant, and L_k, P_{ki} and S_{kij} are the random effects associated with the line, position and sample, respectively. The random effects model assumes these are all independent, and

$$L_k \sim \eta(0, \sigma_L^2)$$

$$P_{ki} \sim \eta(0, \sigma_P^2)$$

$$S_{kij} \sim \eta(0, \sigma_S^2),$$

where $\eta(\mu, \sigma^2)$ denotes a normal probability distribution with mean μ and variance σ^2.

Let X_{kij} be the measured value of the sample taken at (x_{ij}, y_{ij}),

$$X_{kij} = Q_k(x_{ij}, y_{ij}) + \varepsilon_{kij}, \tag{26}$$

where the measurement error ε_{kij} has a $\eta(0, \sigma_m^2)$ distribution.

For a balanced design with s samples per position and p positions per line, the ANOVA table is given in Table VI. The F statistic for testing for a line effect will have expected value

$$1 + \frac{sp\sigma_L^2}{\sigma_S^2+\sigma_m^2+s\sigma_P^2}.$$

This expected value is bigger than F, say, if $\sigma_L^2 > 0$ and s and p are so large that

$$\frac{1}{p}\,\frac{\sigma_p^2}{\sigma_L^2} + \frac{1}{sp}\,\frac{\sigma_S^2+\sigma_m^2}{\sigma_L^2} < \frac{1}{F-1}. \tag{27}$$

TABLE VI

ANALYSIS OF VARIANCE FOR GROUND SAMPLES

Sum of Squares	Degrees of Freedom	Expected Mean Square
$\sum_{k,i}\sum_{j}(X_{kij} - \bar{X}_{ki.})^2$	$\ell p(s-1)$	$\sigma_S^2 + \sigma_m^2$
$\sum_{k}\sum_{i}(\bar{X}_{ki.} - \bar{X}_{k..})^2$	$\ell(p-1)$	$\sigma_S^2 + \sigma_m^2 + s\sigma_P^2$
$\sum_{k}(\bar{X}_{k..} - \bar{X})^2$	$\ell - 1$	$\sigma_S^2 + \sigma_m^2 + s\sigma_P^2 + sp\sigma_L^2$

The problem considered below is the possibility of learning something about the relative magnitudes of σ_S^2, σ_m, σ_P^2 and σ_L^2 from the aerial data. If the ratios on the left-hand side of Eq. (27) could be estimated, then the sample size required to demonstrate a "line effect" could be determined. Unfortunately, it turns out that only the ratio σ_P^2/σ_L^2 is available from the aerial data. Furthermore, preliminary work on uranium concentrations in the water and sediments of several Rocky Mountain quadrangles suggests that σ_S^2 is far from negligible (though possibly σ_m^2 is very small). Therefore, if s and p are chosen under the assumption that the second ratio is negligible, the result will be much too small.

To make the connection between ground-based and aerial measurements, we will now consider $Q_k(x,y)$ as a random variable defined along a flight line, and not only at the sampling positions. As in Sec. III.A., we will assume that $Q_k(x,y)$ is a second-order stationary process. (Stationarity is, perhaps, improbable, but the development which follows, based on these assumptions, gives results which are qualitatively similar to those obtained with more general assumptions, such as the existence of stationary differences only. Moreover, in the experimental situation being considered here, stationarity within "lines" may be quite a reasonable assumption.)

Parallel to the ANOVA analysis, we write

$$Q_k(x,y) = \mu + L_k + P_k(x,y) + S_k(x,y), \tag{28}$$

where

μ is constant over the whole region;

L_k is constant along a line, and the values for each line are drawn from a $\eta(0, \sigma_L^2)$ population;

$P_k(x,y)$ is a relatively slowly varying quantity, i.e., $\mathrm{Cov}[P_k(x), P_k(x+\Delta x)]$ is close to σ_P^2 for small Δx, and does not fall off to zero until Δx is several kilometers, or something on the order of the distance between "positions;"

$S_k(x,y)$ is a rapidly varying quantity, $\mathrm{Var}\, S_k(x,y) = \sigma_S^2$, but $\mathrm{Cov}[S_k(x,y), S_k(x+\Delta x, y+\Delta y)]$ becomes zero for $\sqrt{\Delta x^2 + \Delta y^2}$ on the order of a few meters.

As usual, define

$$C_Q(\Delta x, \Delta y) = \mathrm{Cov}[Q_k(x,y),\ Q_k(x+\Delta x, y+\Delta y)].$$

As a function of Δx (Δy held constant), C_Q may be quite smooth at the origin, as suggested by Figure 14a, or rather peaked, as in Figure 14b. The first case corresponds to little local sampling variation, that is, to σ_S^2 being very small. In this case if the "positions" are at least b kilometers apart, $\sigma_P^2 = C_Q(0,0)$ is the main component of variance of observations within a line.

Preliminary studies of hydrogeochemical data in the Albuquerque and Montrose quadrangles suggests that Figure 14b is more typical, however. Figures 15a and 15b are estimates of $\gamma_Q(|h|)$ for log uranium concentrations in sediments and in water in the Montrose quadrangle. [Recall that in the stationary case $\gamma(h)$ is related to the autocovariance function by Eq. (19).] It is clear that $\lim_{|h|\to 0} \gamma_Q(|h|)$ is not zero, and that considerable variability, at least a quarter of the total over the whole quadrangle (the quantity designated as "variance" in the subtitle) can be expected among samples separated by less than 200 meters. Thus, C_Q has to be modeled with a substantial discontinuous part, corresponding to σ_S^2.

From the aerial data for each flight line, or perhaps an average over all flight lines, an estimate of the autocovariance function $C_Z(h)$ can be formed, which is related to $C_Q(\Delta x, \Delta y)$ as in Eq. (10):

$$C_Z(h) = \begin{cases} \varepsilon^2\theta^2[(C_Q * C_p)(vh,0) + C_B(h)] & h \neq 0 \\ \varepsilon^2\theta^2[(C_Q * C_p)(vh,0) + C_B(h)] + E\ \mathrm{Var}\ \eta(t), & h = 0. \end{cases}$$

Here we have assumed that the emissions $\Lambda(x,y)$ are related to the concentration $Q(x,y)$ by a proportionality factor θ,

$$\Lambda(x,y) = \theta Q(x,y).$$

Assume that the background contribution is negligible. From Eq. (29), the contributions to Q which are not constant within a flight line (the only terms which contribute to C_Q) come from P_k and S_k, so

$$C_Z(h) \sim \varepsilon^2\theta^2[(C_P * C_p)(vh,0) + (C_S * C_p)(vh,0)], \quad h \neq 0. \tag{29}$$

The effect of convolution with C_p is to broaden the original convolution function (C_P or C_S) by the width of C_p (about 500 m) and to multiply it by the volume under C_p, which is P_v^2, in the notation of Eq. (5). Assuming C_P represents the continuous part of the autocovariance function C_Q, and $\lim_{\Delta x, \Delta y \to 0} C_P(\Delta x, \Delta y) = \sigma_P^2$, the first summand in Eq. (29) is approximately

$$\varepsilon^2\theta^2(C_P * C_p)(vh,0) \sim \varepsilon^2\theta^2 P_v \sigma_P^2$$

for small h. On the other hand, C_S is the discontinuous part of C_Q, and convolution with C_p results in a term

$$\varepsilon^2\theta^2\sigma_S^2 C_p(vh,0)$$

for small h. (See remarks in Sec.III.A in connection with Fig. 6b.) Now, $C_p(0,0) \ll P_v$, so unless σ_S^2 is extremely large, the contribution of the second term in Eq. (29) is negligible, and

$$\hat{C}_Z(\Delta h) \sim \varepsilon^2\theta^2 P_v \sigma_P^2. \tag{30}$$

(Compare Eq. (23); again we are using $C_Z(\Delta h)$ as an estimate of $\lim_{h \to 0} C_Z(h)$.) The overall variance of the aerial data over all flight lines includes an additional term, which is $\varepsilon^2\theta^2 P_v \sigma_L^2$,

$$\text{Var } Z \sim \varepsilon^2\theta^2 P_v^2(\sigma_L^2 + \sigma_P^2) + E(\text{Var } \eta). \tag{31}$$

The average variance within a line, on the other hand, is

$$\overline{\text{Var } Z_k} \sim \varepsilon^2 \vartheta^2 p_v^2 \sigma_P^2 + E(\text{Var } \eta). \tag{32}$$

Combining Eqs. (30), (31) and (32) we can arrive at an estimate of the ratio σ_P / σ_L^2. However, as we saw above, a term involving σ_S^2 is seldom available. Some experiments to determine σ_S^2 directly are therefore necessary.

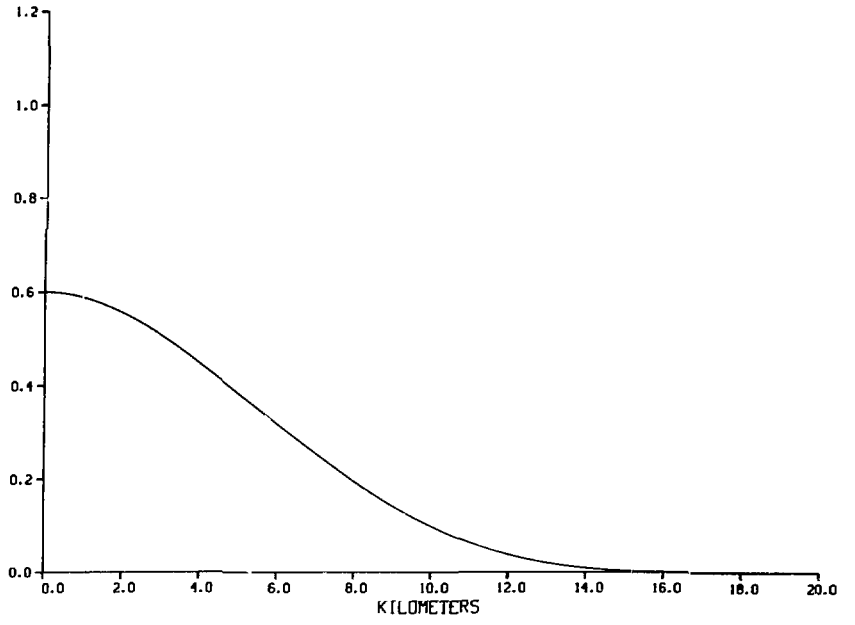

Fig. 14a. Possible shape for autocovariance function of a smooth process.

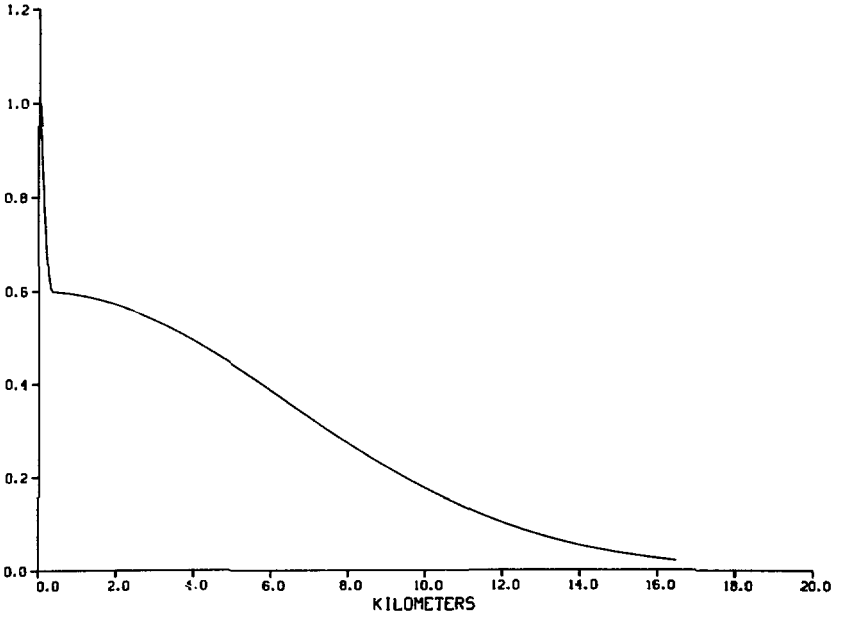

Fig. 14b. Possible shape for autocovariance function of a noisy process.

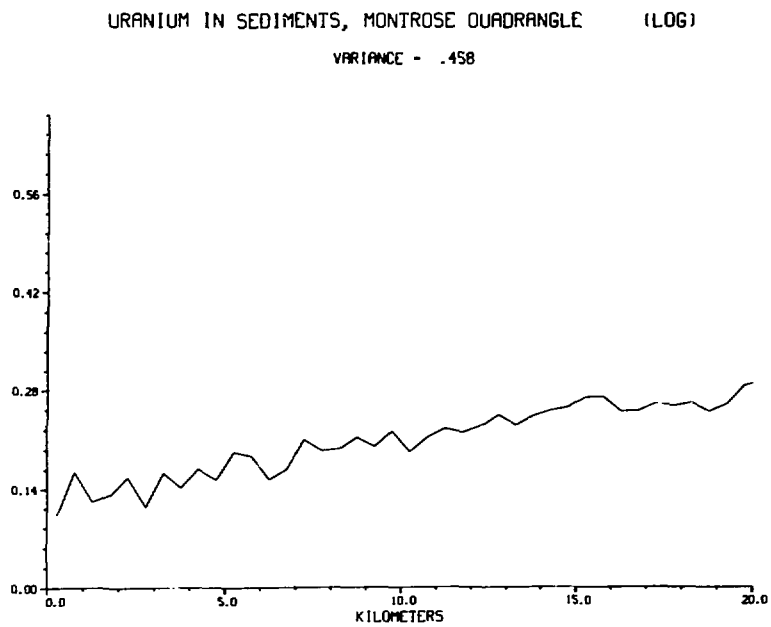

Fig. 15a. Experimental variogram of the logs of uranium concentration in sediments.

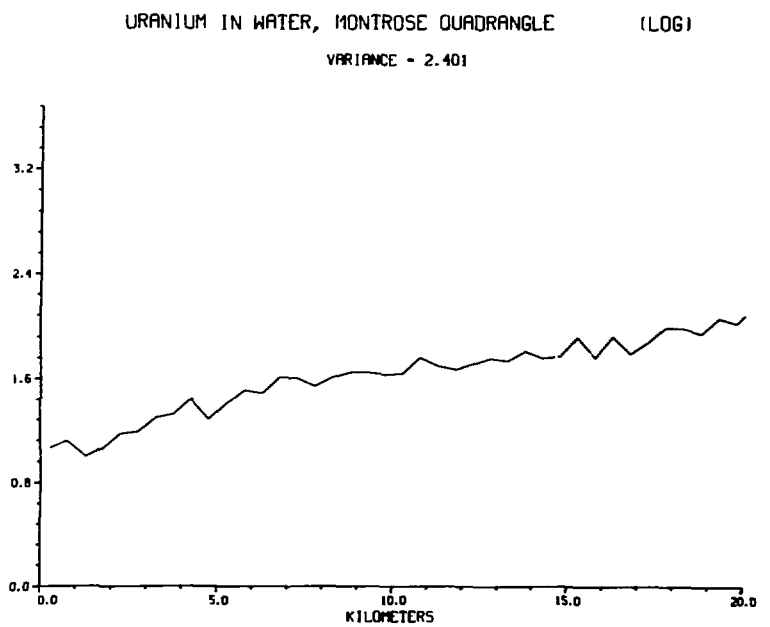

Fig. 15b. Experimental variogram of the logs of uranium concentration in water.

IV. COMPARISON OF NORMAL AND LOGNORMAL METHODS OF PERCENTILE ESTIMATION

Implicitly or explicitly, percentile estimation has always been an important feature of the treatment of National Uranium Resource Evaluation (NURE) aerial radiometric data. Standard deviation maps actually show the location of records which fall outside certain percentile ranges if one assumes the data follow either a normal or lognormal distribution. The definition and identification of anomalous points in regions is usually very closely related to the estimation of percentiles.

Because of the prominance of percentiles in the analysis of NURE aerial radiometric data, a study was conducted to compare two commonly used methods of estimating them. The main thrust of the study was to determine how estimation is influenced by specific types of departures from distributional assumptions and to provide a suggested method for estimating percentiles. A report presenting the results of this study will be released by DOE, Grand Junction, in May or June, 1980. Highlights are presented here.

The two methods of estimation considered are (1) the normal assumption method and (2) the lognormal assumption method. In general, if a random sample is known to have come from a particular type of distribution, the best way to estimate percentiles is to use the properties of that distribution. The two methods assume that the data are from either a normal or lognormal distribution and make use of the properties of these distributions.

Whenever a procedure is applied which, either implicitly or explicitly, is based on a particular probability distribution, the user should be aware of the consequences of failure to meet mathematical assumptions. For instance, under normal distribution theory, $\bar{X} + 1.645S$ provides an estimate of the 95th percentile, where $\bar{X}$ and S are the sample mean and standard deviation. (The quantity 1.645 is obtained from a normal table.) The meaning of $\bar{X} + 1.645S$ is uncertain, however, if the distribution from which the sample was taken is not normal.

There are many ways in which distributional assumptions can be violated and we consider two specific cases.

(1) assuming that a distribution is lognormal when it is really normal, and

(2) assuming that a distribution is normal when it is really lognormal.

Figures 16 and 17 summarize the results of a Monte Carlo study to determine the percentage error realized by using the wrong distribution. The error realized depends on two variables: the percentile being estimated and the coefficient of variation (CV) of the population being sampled. (The coefficient of variation is the ratio of the standard deviation to the mean.)

The curves in Fig. 16 show those CV-percentile combinations for which there is a 5% error when the wrong distribution is used. The regions where errors are greater than and less than 5% are also indicated. Figure 16 shows that making an error by assuming the data come from a normal population when they really come from a lognormal population causes problems primarily when estimating tail percentiles. When estimating upper percentiles, however, the error of making a log-transformation has more serious consequences than does the error of failing to transform. When estimating lower tail percentiles, the error of failing to perform a log-transformation has much more serious consequences than does the error of transforming. When one is estimating percentiles that are between the 10th and 80th, the consequence of either type of error is not great if the coefficient of variation is reasonably small, say less than 0.27.

Figure 17 provides some more detail for the upper percentiles. The 5% error line resulting from the failure to take logs is shown. The 10% error line for erroneously taking logs is also shown. For coefficients of variation between 0.32 and 0.27, the error of assuming normality causes an error of less than 5% when estimating the 90th and 95th percentiles while the error of assuming lognormality causes at least a 10% error.

Based on the above observations the following suggestions can be made. These are general rules of thumb and should not be interpreted to be precise directions for percentile estimation. As throughout the development, they assume that normal and lognormal populations are the only possibilities.

(1) The coefficient of variation should be computed. If the coefficient is large, say greater than 0.27, percentile results should be interpreted cautiously.

(2) If a choice between the two distributions cannot be made and if the lower percentiles (10th and below) are of interest, the safest course of action is to assume lognormality. If the central percentiles (10th to 80th) are of interest, it makes little difference which distribution is chosen. When the upper percentiles are of interest, the safest course of action to take when in doubt is to assume the normal distribution. All of these remarks assume that the coefficient of variation is in the favorable range defined above.

(3) Because of the large errors which can result from the choice of the wrong distribution, an effort should be made to determine which distribution best fits a set of data. A chi-square, Kolomogorov, Lilliefors, or other test for normality (Ref. 5) should be performed. This is particularly important in the case of a large coefficient of variation.

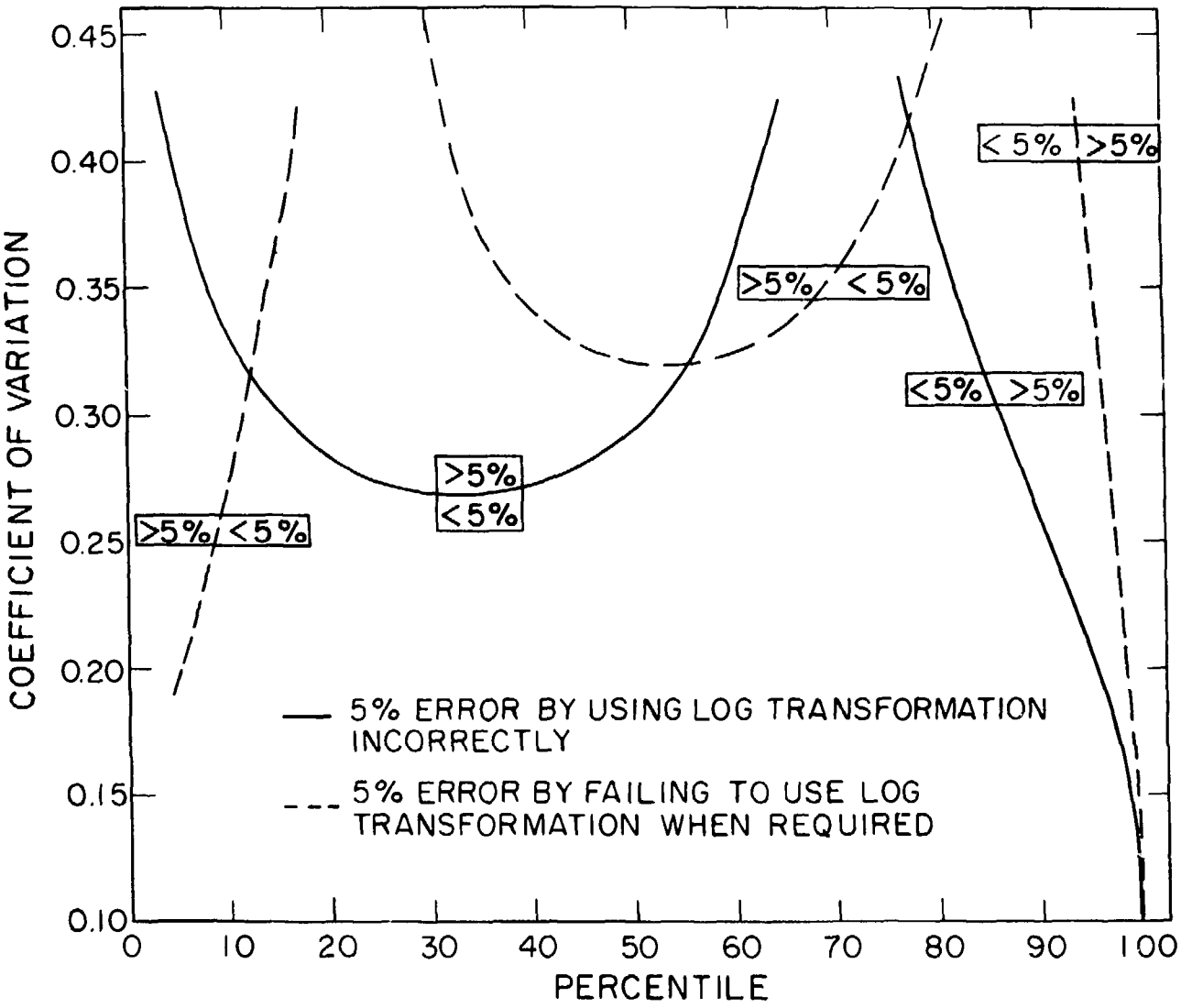

Fig. 16. Coefficient of variation/percentile combinations for which there is a 5% error when using the wrong distribution.

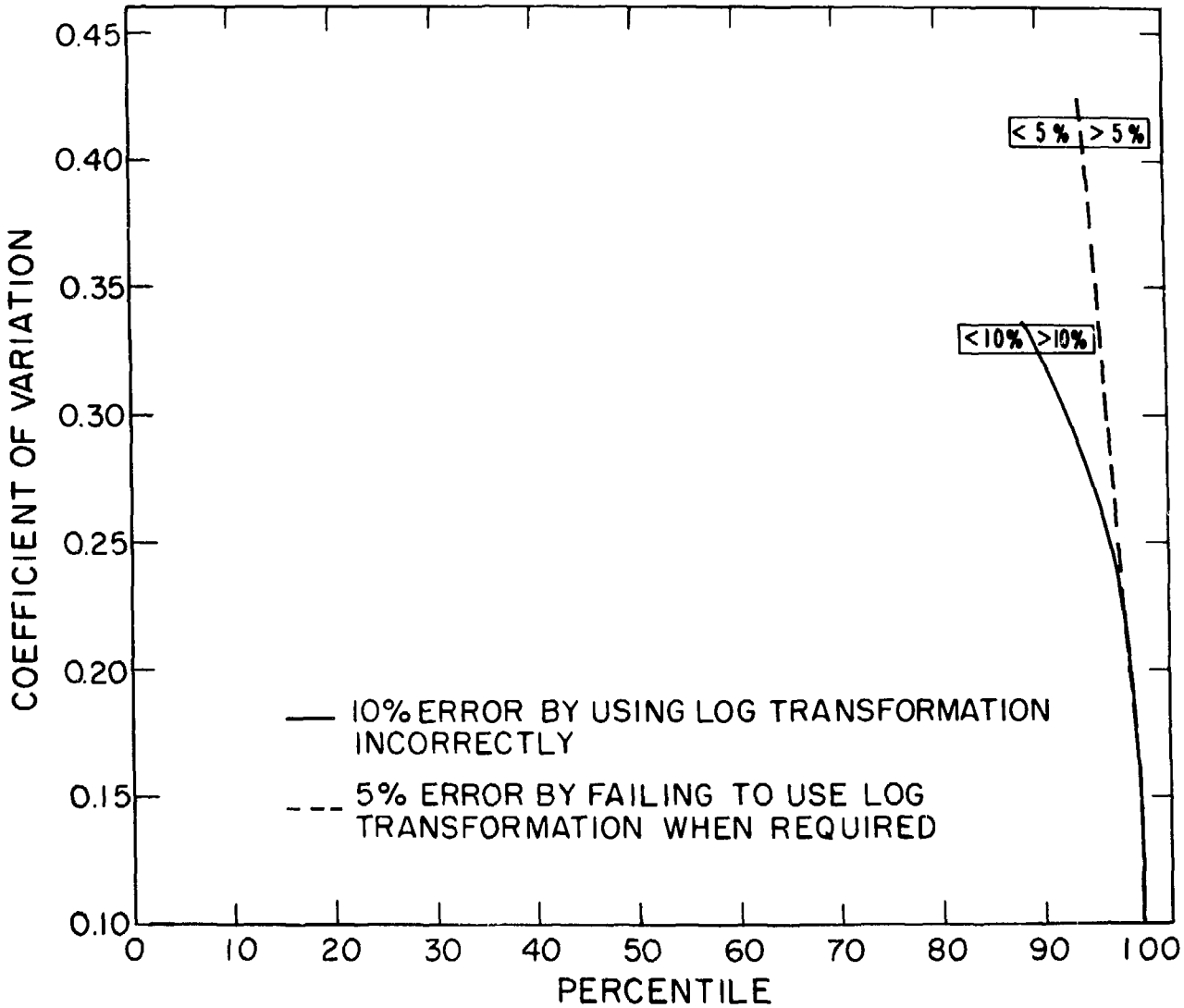

Fig. 17. Region where error of transforming is greater than error of not transforming.

V. URANIUM FAVORABILITY INDEX

During the reporting period computer programs were developed to efficiently and accurately process large sets of data. These codes were applied to the data collected on the 22 quadrangles listed below.

Rawlins	Sherman
Casper	Pratt
Greeley	Rice Lake
Mesa	Ashland
Rock Springs	Eau Claire
Enid	Green Bay
Harrisburg	Hutchinson
Scranton	Lamar
Tyonek	Manhattan
Mt. McKinley	Talkeetna
Lime Hills	Wichita

The data base containing output from these codes plus the original data now consists of approximately 30 million 60-bit words of information.

A. Code Development

We have developed a program which finds percentiles for data sets of up to 200,000 records. The algorithm, which is based on the Blum algorithm (Ref. 5), can be summarized as shown below. The algorithm returns the value of the Tth largest number from a set of size n.

(1) Select a sample of 1000 equally spaced records from the original data. Store the sample in an array, Y, in ascending order.

(2) From this sample, estimate the location of the Tth largest element in the original data. Cutoffs of 5% above and below this estimate are computed. Compute:

$$\text{lower index} = \max\left(\frac{N-T}{N} \cdot 1000 - 50,\ 1\right)$$

$$\text{upper index} = \min\left(\frac{N-T}{N} \cdot 1000 + 50,\ 1000\right)$$

$$\text{lower cutoff} = Y(\text{lower index})$$

$$\text{upper cutoff} = Y(\text{upper index}).$$

(3) Extract from the original data all records between the lower and upper cutoffs. This data will contain approximately 10% of the original data and will be searched for the true T^{th} largest record.

(4) Apply the Blum algorithm (Ref. 5) to this reduced data set to compute the Tth largest element.

A second program which was developed computes the covariance matrix efficiently and accurately for a large set of data. The algorithm is a one-pass algorithm similar to the one-pass algorithm of West (Ref. 6) for computing the variance of a set of data. Our algorithm is summarized here.

```
AMX = X(1)
AMY = Y(1)
T = 0
For I = 2, 3, ..., N Do
     QX = X(I) - AMX
     QY = Y(I) - AMY
     RX = QX/I
     RY = QY/I
     AMX = AMX + RX
     AMY = AMY + RY
     T = T + (I-1) * QX * QY
End
COV = T/(N-I)
```

Because of the sizes of the data bases under study, it is not practical to try to read in all of the data and process it in one step as shown above. Instead, we input (from a mass storage device) 10,000 elements at a time for each of the 2 arrays, X and Y, and accumulate the necessary information in T.

The two codes just described have been sent to BFEC, Grand Junction, Colorado, for their use.

B. Large Data Base

For each of the 22 quadrangles we have computed 265 statistics based on the following 10 variables: thallium, bismuth, potassium, gross count, Tl/gross, Bi/gross, K/gross, Bi/K, Bi/Tl, and Tl/K. For each variable we have computed the median, mean, and standard deviation; the 99th, 95th, 90th, 1st, 5th, and 10th percentiles; the average and standard deviation of all the

points above the 99th, 95th, and 90th percentiles; and the same for all points below the 1st, 5th, and 10th percentiles. In addition, we have computed the 55 upper triangular elements of the 10 by 10 covariance matrix for the variables listed above. Thus, there are 10 variables times 21 statistics computed on them plus the 55 elements of the covariance matrix for a total of 265 statistics.

Work is underway to begin analysis of this data in order to find a uranium favorability index.

C. Problems With the Data

Some of the quadrangles listed in Ref. 7 are not being used in this study. Some of the reasons for their omission are missing data, incomplete quadrangle coverage, and difficulties in reading the data tapes.

REFERENCES

1. T. R. Bement and M. S. Waterman, "Locating Maximum Variance Segments in Sequential Data," Math. Geol. 9, 55-61 (1977).

2. A. G. Journel and C. J. Huijbregts, Mining Geostatistics (Academic Press, London, 1978).

3. J. Tammenmaa, R. L. Grasty and M. Peltoniemi, "The Reduction of Statistical Noise in Airborne Radiometric Data," Canadian Journal of Earth Sciences 13, 1351-1357 (1976).

4. K. L. Kosanke and C. D. Koch, "An Aerial Radiometric Data Modeling Program," IEEE Trans. Nuclear Science NS-25, No. 1, 767 (1978).

5. M. Blum, R. W. Floyd, V. Pratt, R. L. Rivest and R. E. Tarjan, "Time Bounds for Selection," Journal Computer and Syst. Sci. 7, 448-461 (1973).

6. D. H. D. West, "Updating and Variance Estimates: An Improved Method," Comm. ACM 22,9, 532-535 (September 1979).

7. J. A. Howell, T. R. Bement and P. L. Buslee, "Geostatistics Project of the National Uranium Resource Evaluation Program, April-September 1979," Los Alamos Scientific Laboratory report LA-8175-PR (December 1979).