LA-9046-PR
Progress Report

UC-51
Issued: October 1981

Geostatistics Project of the National Uranium Resource Evaluation Program

October 1980—March 1981

T. R. Bement
J. A. Howell
M. D. McKay
M. E. Johnson
G. L. Tietjen
G. W. Wecksung
C. K. Jackson

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

GEOSTATISTICS PROJECT OF THE NATIONAL URANIUM RESOURCE EVALUATION PROGRAM OCTOBER 1980--MARCH 1981

by

T. R. Bement, J. A. Howell, M. D. McKay, M. E. Johnson, G. L. Tietjen, G. W. Wecksung, and C. K. Jackson

ABSTRACT

During the period covered by this report, we analyzed the radiometric data collected along the Texas Gulf Coast using ten discriminant analysis techniques to establish radiometric signatures and classify new observations. We conducted a survey of several methods for computing the covariance matrix of large data sets, with particular interest to one-pass algorithms. An investigation of methods of estimating upper-tail percentiles for aerial radiometric data was begun. A feasibility study was conducted concerning the design of ground-based sampling plans using a statistical model for the correlation between observations taken along a flight line. A study of the use of cluster analysis in aerial radiometric data analysis was initiated. Two short courses on statistical methods were presented in Grand Junction, Colorado, and more are planned.

I. INTRODUCTION

This report outlines the activities and progress of the Los Alamos National Laboratory on the Geostatistics project during the first half of FY81. 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 Los Alamos applies statistical methods to the analysis of data collected by airborne instrumentation. To handle a broad range of problems related to the NURE, Los Alamos maintains a close statistical consulting relationship with the DOE Grand Junction Office and the Bendix Field Engineering Corporation (BFEC) in Grand Junction.

We established signatures for favorable and unfavorable units along the Texas Gulf Coast using partial discriminant analysis with the linear discriminant function.

A survey of several methods for computing the covariance matrix of large data sets was completed with particular interest given to one-pass algorithms.

We investigated several methods of cluster analysis placing special emphasis on applications requiring the analysis of large data sets.

We performed an evaluation of the use of aerial data in designing ground sampling plans, concentrating attention on the problem of isolating various components of variance in the aerial data.

Our report on methods of estimating upper-tail percentiles for aerial radiometric data is summarized in this report.

In addition, our project included a review of the Texas Instruments report, "Interpretation Methods Test Report for NURE Aerial Radiometric and Geochemical Data," Vol. I (Ref. 1).

II. DESIGNING GROUND-SAMPLING PLANS

A. Background

A statistical model for the correlation among observations taken from an aircraft along a flight line is discussed in Ref. 2. That report shows how estimates of certain ratios of variances can be obtained and how these estimates might apply in the detection of differences among flight lines using ground sampling.

A common method for detecting differences in the concentration of uranium in two areas is based on the ratio of the variances of the concentration for the two areas. In aerial surveys, counts of gamma rays are recorded. After adjusting for various sources of radiation and for altitudes, some number of the counts can be attributed to ground-source radiation. Letting E denote the ground-source counts and Q denote the ground concentrations,

$$E = cQ \quad .$$

That is, we assume that the gamma ray counts and the uranium concentration are related by a proportionality constant, c. Let Q_1 and Q_2 be concentrations in two areas. Because the variance of Q and that of E are related by

$$V[E] = c^2 V[Q] \quad ,$$

we see that the ratio of the variance of concentrations is equal to the ratio of variances of counts. That is,

$$V[Q_1]/V[Q_2] = V[E_1]/V[E_2] \quad .$$

Hence, an estimate of the concentration variance ratio can be obtained from the aerial data.

A related problem concerns the use of aerial data in designing ground-based sampling plans for estimating actual concentrations (as opposed to detecting differences in concentrations). In this problem, we are concerned about

$$V[Q] = (1/c^2)V[E] \quad .$$

Thus, we need specific information about the proportionality constant, c, and the variance of counts.

The following three sections describe, through simplified models, the approach we are using to design ground-based sampling plans for concentration estimation.

B. Ground-Sampling Model

Suppose that we are going to take ground samples along a flight line. Let x denote a location on the flight line and $Q(x)$ the concentration at the point x. We will use the model

$$Q(x) = \mu + P(x) + S(x) \quad ,$$

where

μ is a constant over the flight line;

$P(x)$ is a relatively slowly varying function;

$S(x)$ is a rapidly varying function.

Suppose that the sampling occurs along the flight line at several sites and that several samples are collected at each site. $P(x)$ is constant within a site but is different at each site. $S(x)$ varies for each sample at a site.

The precision of ground-sample estimates of uranium concentration is determined by the covariance functions

$$C_p(h) = \mathrm{Cov}[P(x), P(x + h)]$$

and

$$C_s(h) = \mathrm{Cov}[S(x), S(x + h)] \quad .$$

We want to obtain estimates of these functions from the aerial data.

C. Aerial-Collection Model

Counts are recorded by aircraft for fixed periods of time Δt (say, 1 s). During the interval $(t - \Delta t/2,\ t + \Delta t/2)$, the aircraft flying with speed N along the flight line would move the distance $N\Delta t$ and receive counts from the entire flight line according to a weighting function (point-spread function), $p(\cdot)$. Let $K(t)$ denote the number of photons reaching the aircraft from the ground during the time interval $(t - \Delta t/2,\ t + \Delta t/2)$. Then

$$K(t) = \upsilon \int_{t-\Delta t/2}^{t+\Delta t/2} ds \int_{-\infty}^{+\infty} dx\, p(x)\, Q(Ns - x) \quad .$$

Letting ε denote the detector efficiency and $n(t)$ all other sources of counts, the number of counts recorded during $(t - \Delta t/2,\ t + \Delta t/2)$ is

$$\mathcal{L}(t) = \varepsilon K(t) + n(t) \quad .$$

(We are ignoring the stochastic or Poisson behavior of the gamma emission process.)

D. Signal Reconstruction

Using the aerial data, Z(t), together with data on aircraft background and atmospheric radon, we are reconstructing effective "ground-level" spectra. The reconstructed spectra should reflect the spatial trends in the mean count rate as well as the spatial variability in the count rate. To extract information from Z(t), one must identify the different components of the signal, the statistical process that generated them, and the method in which they were merged (convolved) to form the signal. As these points are discussed in Ref. 2, we will proceed to outline the method of signal reconstruction.

The Fourier transform of the aerial data can be factored into parts related to the ground signal, the point-spread or altitude-attenuation function, the aircraft background, and the atmospheric radon. The transforms for the last three parts can be estimated from instrument calibration experiments and data from an upward-facing detector on the aircraft. Thus, the transform associated with the ground signal can be estimated. Having obtained this transform, one can invert it to produce the reconstructed ground signal.

The process of determining the point-spread-function component of the Fourier transform of Z(t) is being investigated using Lake Mead and Walker Field test data. This work will continue until a satisfactory Gaussian approximation to the point-spread function is found.

Arminto area flight line data are being studied to see how well the background and atmospheric components can be estimated.

Even with these components of the Fourier transform, we will encounter difficulties using the reconstructed spectra because of the limit of resolution of the ground signal. There is a theoretical limit to the ground detail we can expect. This limit, arising from the 1-s counting period, means that a reconstructed point value is really an average of what the aircraft saw during a small number of counting periods. Several ground samples will be taken at each site, and it appears that information on sampling variation within a site may not be available from the aerial data.

Two approaches to estimating within site variation are being investigated. One of these involves using the aerial data to estimate an average concentration and then constructing distributions of concentrations in the site area consistent with the aerial average and "looking like" what might be expected from other considerations. These hypothesized distributions would be studied to see the type of variance they induce on the ground-sampling process.

A second approach is to consider the variation in concentration along a flight line as the sum of a slowly changing trend and a more quickly varying stochastic component. If this model can be supported by the data, then estimation of the trend will yield estimates of an average stochastic component. This estimate, in turn, can be adjusted for the relatively small area of a ground-sample site.

III. PERCENTILE ESTIMATION

We are preparing a report on methods of estimating upper-tail percentiles for aerial radiometric data. In the past, the data have been assumed to follow either a normal or log-normal distribution. The 90th, 95th, and 99th percentiles were calculated from a fit to one of these distributions. Bement and Pirkle[3] discuss the errors that can result when these are the only two possibilities. The report now being prepared gives consideration to other distributional possibilities. In particular, the Johnson and Pearson systems of distributions are considered as well as the use of ordinary sample percentiles. Techniques were tested on data from the Copper Mountain and Owl Creek Mountains, Wyoming, as well as on simulated data.

IV. DISCRIMINANT ANALYSIS

A paper, "Discriminant Analysis Applied to Aerial Radiometric Data and Its Application to Uranium Favorability in South Texas," has been accepted for publication in Mathematical Geology.[4] Aerial radiometric data collected along the Texas Gulf Coast was analyzed using ten discriminant analysis techniques. The purpose of the analyses was to address the following questions:

1. Do particular geologic formations appear homogeneous along strike in the South Texas Central Plain?
2. Do favorable geologic formations exhibit a common aerial radiometric signature?
3. Are there sets of observations from known favorable and unfavorable formations that can be used in classifying observations from other formations (whose favorability is undetermined)?

Discriminant analysis procedures that were applied include the classical linear and quadratic discriminant analyses as well as the use of robust estimators or ranked data with the classical procedures. Partial and forced discriminant analysis procedures were also used. Partial discrimination methods allow for nonclassification of observations whereas forced methods classify every observation.

Our study suggests that partial linear discriminant analysis using raw (rather than ranked) data is adequate.[5] Geologic results allow one to differentiate between the Catahoula Formation in the Houston embayment and the Catahoula Formation in the Rio Grande embayment. In addition, three signatures were established for four formations known to be favorable for uranium resources.

V. COMPUTING THE COVARIANCE MATRIX

A. Introduction

We completed a survey of several methods for computing the covariance matrix of large data sets. With a large amount of data, it is desirable to make as few passes through the data as possible. Therefore, one-pass algorithms were of special interest.

The data used as input to our test programs were generated using a uniform random-number generator. We generated three correlated vectors: x, w, and u. They were each of length 150 000 and were generated as shown in the following algorithm. The function RAND(0) generates random numbers from a uniform (0,1) distribution.

```
for i=1,n do
  x(i) = RAND(0)
  w(i) = RAND(0) + a * x(i) + 2000.0
  u(i) = RAND(0) + x(i) + 2000.0
end
```

The 2 x 2 covariance matrix for w and u is approximately

$$\begin{bmatrix} 12.083333 & 1.0 \\ 1.0 & 0.166667 \end{bmatrix} .$$

Because of the large size of the vectors, they were buffered to and from memory in blocks of size 10 000.

B. The Algorithms

The algorithms that we used are now described. Methods 1 through 5 are variations of the standard textbook two-pass algorithm. It is as follows:

$$\text{xbar} = \sum_{i=1}^{n} x_i/n \quad ,$$

$$\text{ybar} = \sum_{i=1}^{n} y_i/n \quad ,$$

$$\text{cov} = \sum_{i=1}^{n} (x_i - \text{xbar})(y_i - \text{ybar})/(n - 1) \quad .$$

Method 1 uses this algorithm as shown. In Method 2, double precision is used in the accumulation of the mean. Method 3 normalizes the data to lie between 0 and 1 in absolute value before computing the mean. Method 4 combines the normalizing and double precision. Method 5 consists of quadruple-precision arithmetic.

Methods 6 and 7 are the West algorithm[6] that is given by:

```
xbar = x1
ybar = y1
t = 0
for i=2,n do
  qx = xi - xbar
  rx = qx/i
  xbar = xbar + rx
  qy = yi - ybar
  ry = qy/i
  ybar = ybar + ry
  t = t + (i-1) * qx * ry
```

```
end
  cov = t/(n - 1)  .
```

Method 6 uses single precision, and Method 7 uses double precision.

Methods 8 and 9 use the Youngs and Cramer updating algorithm as described in Refs. 5 and 7. This algorithm is as follows:

$$tx_{1,j} = \sum_{i=1}^{j} x_i$$

$$tx_{j+1,j+k} = \sum_{i=j+1}^{j+k} x_i$$

$$ty_{1,j} = \sum_{i=1}^{j} y_i$$

$$ty_{j+1,j+k} = \sum_{i=j+1}^{j+k} y_i$$

$$s_{1,j} = \sum_{i=1}^{j} [(x_i - tx_{1,j})/j] \cdot [(y_i - ty_{1,j})/j]$$

$$s_{j+1,j+k} = \sum_{i=j+1}^{j+k} [(x_i - tx_{j+1,j+k})/k] \cdot [(y_i - ty_{j+1,j+k})/k]$$

$$s_{1,j+k} = s_{1,j} + s_{j+1,j+k} + c_1 \cdot (c_2 \cdot tx_{1,j} - tx_{1,j+k}) \cdot (c_2 \cdot ty_{1,j} - ty_{1,j+k}) \ ,$$

where $c_1 = j/[k \cdot (j + k)]$ and $c_2 = (j + k)/j$.

When $j + k = n$, then $cov = s_{1,n}/(n - 1)$. In fact, we selected j and k so that there would be an inner loop in which the formulae were updated in pairs to 10 000 and an outer loop in which the formulae were updated in units

of 10 000. Method 8 uses this algorithm in single precision, and Method 9 is double precision.

C. Numerical Results

Results of the different algorithms are given in Table I. The "correct" answers were taken to be those resulting from using the two-pass algorithm with quadruple-precision arithmetic.

D. Conclusions

From these results we see that the standard two-pass algorithm performs poorly without double precision. Examining only the first seven digits of each answer (rounding the double-precision answers to single-precision answers), we note the following rankings of performance. Methods 2 and 4 give the same results as the standard method (Method 5). Method 9 outperforms Method 8 slightly. Method 6 gives better results than Method 7, which has normalization. Methods 1 and 3 give wrong answers.

The one-pass Youngs and Cramer algorithm with double precision appears to give very good results in much less time than the two-pass with double precision. Therefore, we recommend it for use on large data sets.

TABLE I

EXPERIMENTAL RESULTS AND TIMINGS

Type of Algorithm	CPU Time (s)	Covariance Matrix		
Textbook (Two Pass)	0.9356000E+02	0.1264869E+02	0.1553510E+01	0.8080876E+00
Textbook (Two Pass, Double Precision)	0.1042500E+03	0.121869869232178E+02	0.100957870483398E+01	0.167746096849442E+00
Textbook (Two Pass, Normalized Data)	0.9885999E+02	0.1307400E+02	0.7953988E+00	0.2194436E+00
Textbook (Two Pass, Double Precision, Normalized Data)	0.1143500E+03	0.121869869232178E+02	0.100957870483398E+01	0.167746096849442E+00
Textbook (Two Pass, Quad Precision)	0.3331269E+05	0.121869869232178E+02	0.100957870483398E+01	0.167746096849442E+00
West (One Pass)	0.5683000E+02	0.1218664E+02	0.1009572E+01	0.1677420E+00
West (One Pass, Normalized Data)	0.6214000E+02	0.1218665E+02	0.1009614E+01	0.1677799E+00
Youngs and Cramer (One Pass)	0.5712000E+02	0.1218694E+02	0.1009568E+01	0.1677454E+00
Youngs and Cramer (One Pass, Double Precision)	0.5967000E+02	0.121869821548462E+02	0.100957882404327E+01	0.167746439576149E+00

VI. CLUSTER ANALYSIS

An effort to apply clustering techniques was started late in the reporting period. The objective of the study is to determine if cluster analysis can be used to determine the number of distinct radiometric signatures present in a given set of data (quadrangle, geologic unit, group of geologic units, etc.). Several methods of clustering were considered with respect to their suitability for analysis of the aerial radiometric data and for computer code availability. The basic methods include hierarchical clustering, data partitioning, and density searching. Although the number of basic procedures is small, a wide variety of techniques is possible by permuting various choices of distance functions and similiarity measures. At this point, we believe a combination of methods will be required to handle the large quantity of data associated with the aerial program.

VII. STATISTICS SHORT COURSES

Short courses in general statistics and regression analysis were presented in Grand Junction. Additional courses on factor analysis, discriminant analysis, cluster analysis. time series, and sampling are planned.

REFERENCES

1. "Interpretation Methods Test Report for NURE Aerial Radiometric and Geochemical Data," Texas Instruments, Inc., prepared for the US Department of Energy GJBX-138 (March 1980) Vol. I-text.

2. K. Campbell, T. R. Bement, J. A. Howell, R. J. Beckman, C. K. Jackson, and P. Buslee, "Geostatistics Project of the National Uranium Resource Evaluation Program, October 1979-March 1980," Los Alamos Scientific Laboratory report LA-8508-PR (August 1980).

3. T. R. Bement and F. L. Pirkle, "Estimating Percentiles in Aerial Radiometric Data Using Normal and Lognormal Distributional Assumptions," accepted by Journal of Mathematical Geology (LA-UR-80-3124).

4. D. A. Patterson, F. L. Pirkle, M. E. Johnson, T. R. Bement, C. K. Jackson, and N. K. Stablein, "Discriminant Analysis Applied to Aerial Radiometric Data and Its Application of Uranium Favorability in South Texas," accepted by Journal of Mathematical Geology (LA-UR-80-906).

5. E. A. Youngs and E. M. Cramer, "Some Results Relevant to Choice of Sum and Sum-of-Product Algorithms," Technometrics 13, No. 3, 657-665 (August 1971).

6. D. H. D. West, "Updating and Variance Estimates: An Improved Method," Comm. ACM 22, No. 9, 532-535 (September 1979).

7. T. F. Chan, G. H. Golub, and R. J. LeVeque, "Updating Formulae and a Pairwise Algorithm for Computing Sample Variances," Dept. of Computer Science, Stanford University, STAN-CS-79-773 (November 1979).