April 1981—September 1981

NOTICE

PORTIONS OF THIS REPORT ARE ILLEGIBLE

It has been reproduced from the best available copy to permit the broadest possible availability.

M. D. McKay
J. A. Howell
D. E. Whiteman
B. A. Duran*
C. K. Jackson
G. W. Wecksung
T. R. Bement
F. A. Pirkle**

— DISCLAIMER

[illegible]

*Visiting Staff Member. Department of Mathematics, Texas Technical University, Lubbock, TX 79409.

**Bendix Field Engineering Corporation, Grand Junction, CO 81501.

GEOSTATISTICS PROJECT OF THE
NATIONAL URANIUM RESOURCE EVALUATION PROGRAM
April 1981--September 1981

by

M. D. McKay, J. A. Howell, D. E. Whiteman, B. A. Duran,
C. K. Jackson, G. W. Wecksung, T. R. Bement, and F. A. Pirkle

ABSTRACT

During the period covered by this report, we proposed a method of comparing aerial and ground data as a means of assessing the quality of aerial data. We also compared two methods of partitioning count rates among several isotopes. Time-series analysis and analysis of variance were considered as tools for using aerial radiometric data to aid in designing ground-based sampling experiments. Several methods of computing covariance matrices were compared for use with very large data sets. A study showed a potential for using aerial radiometric data to rank quadrangles according to the Department of Energy's estimated uranium inventories. A discriminant analysis code was transferred to Grand Junction, and several statistics short courses were presented there. Recommended cluster analysis procedures were developed and applied to aerial radiometric data. Reports and papers were prepared on topics such as outlier detection, percentile estimation, discriminant analysis, and statistical package comparison.

I. INTRODUCTION

This report outlines the activities and progress of the Los Alamos National Laboratory on the Geostatistics project during the second 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 considered the problem of relating the aerial data to ground data, thereby assessing the quality of the aerial data.

In response to a request from BFEC, we considered two methods of partitioning count rates in a given energy window among several isotopes and derived conditions when the methods are equivalent.

Time-series and standard analysis of variance procedures were tools for estimating variances. In each case the objective was to develop methods for using aerial data as a basis for designing ground-based sampling experiments.

Several methods of computing covariance matrices were compared.

We evaluated the possibility of using aerial radiometric data to rank quadrangles in order of their DOE uranium inventory estimates. The results are promising.

A discriminant analysis code was modified for transfer to BFEC and a Los Alamos-BFEC open-file report describing the code and its application is in preparation.

We developed suggested procedures for applying cluster analysis to aerial radiometric data. An open-file report on the topic is in preparation.

Papers or reports were prepared on outlier detection, percentile estimation, discriminant analysis, and statistical package comparison.

II. VERIFICATION OF AERIAL DATA

The problem of assessing the quality of data obtained in an aerial survey is not simple. Consider the following questions.

- Do the aerial data accurately predict the concentration of uranium in the area of interest?
- Can the aerial data be used to estimate accurately the trends in concentration over the area of interest?
- Can the aerial data be used to identify accurately regions of unusually high concentration (hot spots)?

- Do the aerial data lead to the same or similar interpretations as ground data would if taken from the same region?

The following discussion addresses the question of consistency between aerial- and ground-sampled data. Rather than trying to assess the quality of aerial data relative to unknown ground "truth," we will try to assess it relative to what would be obtained through ground sampling.

Let

$x(z)$ = ground-sampled value at location z .

and

$y(z)$ = aerial-sampled value over location z .

Assume that

$$x(z) \sim n[\mu_x(z), \sigma_x^2(z)]$$

and

$$y(z) \sim n[\mu_y(z), \sigma_y^2(z)] \quad .$$

Suppose we observe x and y for points

$$z \in \{z_1, z_2, \dots, z_n\} \quad .$$

Assume

$$x(z_1)$$

and

$$x(z_j)$$

are independent when

$$z_i \neq z_j \quad ,$$

and likewise for

$$y(z_i) \text{ and } y(z_j) \quad .$$

A first criterion for consistency between x and y might be to require

$$\mu_y(z) = f[\mu_x(z)] \quad ,$$

with the function $f(\cdot)$ having an inverse. This criterion says that the expected values of aerial and ground data are related in a one-to-one manner. For example, we could ask that the aerial data be a scaled value of ground data plus a bias or background. In this case, we could use the relationship

$$\mu_y(z) = \alpha + \beta \mu_x(z) \quad ,$$

where the constants α and β do not depend on the location coordinate z .

The best way to model the relationship between aerially obtained values $y(z)$ and ground-obtained values $x(z)$ currently is unknown. Initially we propose to use the linear relationship between the means just described, combined with the assumption that the variances are independent of location z . The unknown constants, α , β , σ_x^2 , and σ_y^2 , will be estimated by the method of maximum likelihood.¹ The adequacy of the model will be tested and refinements are expected. The estimation equations have been derived. Analysis is expected to begin in early FY82 when ground sample data from the Arminto, Wyoming, quadrangle is received.

III. CONCENTRATION ESTIMATES AND STRIPPING RATIOS

When determining concentrations of radioisotopes with gamma-ray detectors, certain energy bands (windows) are associated with particular isotopes. Concentrations are functions of count rates in these windows. Because a given isotope may generate counts in more than one window, some method is needed to partition the total counts in a window among the several isotopes. In response to a request from BFEC, we compared two partitioning methods.

We consider a sample of material containing unknown amounts of three isotopes. Let

- c_k = amount of isotope k , where $k = 1, 2, 3$;
- R_j = counts in window j , where $j = 1, 2, 3$;
- a_{kj} = counts in window j per unit (weight) of isotope k ,
where $j, k = 1, 2, 3$; and
- b_{jk} = fraction of counts in window k that are added to counts in
window j , where $j, k = 1, 2, 3$; $j \neq k$.

The b_{jk} are commonly called "stripping ratios."

Assuming no background sources, the relationship between counts and amount of isotope is

$$R_j = \sum_{k=1}^3 a_{kj} c_k \quad , \quad j = 1, 2, 3 \quad . \quad (1)$$

When stripping ratios are used, one assumes that the counts in the windows are related by

$$R_j = a_{jj} c_j + \sum_{\substack{k=1 \\ k \neq j}}^3 b_{jk} R_k \quad , \quad j = 1, 2, 3 \quad . \quad (2)$$

If we let

- c = 3-by-1 vector of c_k ,
- r = 3-by-1 vector of R_j ,
- A = 3-by-3 matrix of a_{kj} , and
- B = 3-by-3 matrix of b_{jk} with $b_{jj} = 0$,

then Eq. (1) can be written in matrix notation as

$$r = Ac \quad , \quad (3)$$

and Eq. (2) becomes

$$r = D_a c + Br \quad , \quad (4)$$

where

D_a = diagonal 3-by-3 matrix of a_{jj} .

If A and B are known, c can be determined from either Eq. (3) or (4). The question is: when are the two methods equivalent. As is shown below, the methods are equivalent when there is only downward scatter in the energy of the gamma rays.

Substituting Ac for r in the right side of Eq. (4), we have

$$r = D_a c + BAc \quad ,$$

or

$$r = (D_a + BA)c \quad . \quad (5)$$

Comparing Eq. (5) with Eq. (3) we have

$$A = D_a + BA \quad ,$$

and

$$D_a A^{-1} = I - B \quad , \quad (6)$$

where I is the 3-by-3 identity matrix. We assume the existence of A^{-1} .

For Eq. (6) to hold, the diagonal elements of the left side must be 1 because the diagonal of B is 0. A sufficient condition for these values to be 1 is that the matrix A is triangular. To see this, consider the i^{th} diagonal element of A^{-1} when A is triangular.

$$(A^{-1})_{ii} = (\prod_{j \neq i} a_{jj}) / |A| \quad ,$$

$$(D_a A^{-1})_{ii} = a_{ii} (\prod_{j \neq i} a_{jj}) / |A| \quad ,$$

$$= |A|/|A|$$

$$= 1$$

The symbol π denotes product, and the symbol $| \cdot |$ denotes determinant.

Therefore, a sufficient condition for the equivalence of the two methods is that A be triangular, which it will be if there is only downward scatter.

IV. DETERMINING GROUND SAMPLE SIZES

In the NURE program report published in August, 1980, an analysis of variance (AOV) model is used to determine sample sizes for ground-based sampling plans.² That procedure can be improved by considering the power of the associated test of hypothesis. We assume that samples will be collected at several positions (sites) along several flight lines. Let

Y_{ijk} = value associated with sample number k at position j on flight line i .

We use a random effects linear model for Y_{ijk} :

$$Y_{ijk} = \mu + L_i + P_{ij} + e_{ijk} \quad , \text{ for } i = 1, \dots, \ell; j = 1, \dots, p; \text{ and } k = 1, \dots, n$$

where

- μ = constant,
- L_i = effect of line i (a normally distributed random variable with mean 0 and variance σ_L^2),
- P_{ij} = effect of position j within line i (a normally distributed random variable with mean 0 and variance σ_p^2), and
- e_{ijk} = residual effect including within site variation and possible measurement error (a normally distributed random variable with mean 0 and variance σ^2).

Moreover, all effects are independently distributed.

The AOV table for the model is given in Table I. The distribution of the test statistic for line effects can be obtained through the following argument. We observe that

$$(\lambda - 1)MS_L / (\sigma^2 + n\sigma_p^2 + np\sigma_L^2) \sim \chi^2(\lambda - 1) ,$$

and

$$\lambda(p - 1)MS_p / (\sigma^2 + n\sigma_p^2) \sim \chi^2[\lambda(p - 1)] ,$$

where $\chi^2(v)$ denotes a chi-square distribution with v degrees of freedom. Therefore,

$$(MS_L / MS_p) \cdot [(\sigma^2 + n\sigma_p^2) / (\sigma^2 + n\sigma_p^2 + np\sigma_L^2)] \sim F[\lambda - 1, \lambda(p - 1)] ,$$

where $F(u, v)$ denotes an F distribution with u and v degrees of freedom. Letting

$$c = (\sigma^2 + n\sigma_p^2 + np\sigma_L^2) / (\sigma^2 + n\sigma_p^2) ,$$

and

$$X = MS_L / MS_p ,$$

TABLE I
AOV TABLE FOR RANDOM EFFECTS MODEL

<u>Source of Variation</u>	<u>Degrees of Freedom</u>	<u>Mean Square</u>	<u>Expected Mean Square</u>
Lines	$\lambda - 1$	MS_L	$\sigma^2 + n\sigma_p^2 + np\sigma_L^2$
Positions	$\lambda(p - 1)$	MS_p	$\sigma^2 + n\sigma_p^2$
Samples	$\lambda p(n - 1)$	MS_s	σ^2

we have

$$X \sim cF[\lambda - 1, \lambda(p - 1)] \quad .$$

X is the test statistic for testing the hypothesis of no differences between lines, and it will be compared to

$$F_{\alpha} = F_{\alpha}[\lambda - 1, \lambda(p - 1)] \quad ,$$

where

$$P\{F > F_{\alpha}\} = \alpha \quad .$$

Let

$$\begin{aligned} \rho &= P\{X > F_{\alpha}\} \\ &= \text{power of the test} \quad . \end{aligned}$$

We want to choose the sample-size determinants λ , p , and n so that the specified power, ρ , of the test is achieved. For example, suppose we plan a test with significance level $\alpha = 0.05$, and we want to reject the hypothesis of no line effect ($\sigma_L^2 = 0$) with probability at least β when there actually is a line effect, (that is, when $\sigma_L^2 > 0$). Then we want to choose λ , p , and n such that

$$P\{X > F_{0.05}\} = 0.95 \quad .$$

We also require that

$$\begin{aligned} P\{X > F_{\alpha}\} &= P\{(1/c)X > (1/c)F_{\alpha}\} \\ &= P\{F[\lambda - 1, \lambda(p - 1)] > (1/c)F_{\alpha}\} \\ &\geq \beta \quad . \end{aligned}$$

Therefore, we want to choose λ , p , and n so that c satisfies

$$F_p \geq 1/c F_\alpha ,$$

or

$$F_\alpha / F_p \leq c .$$

We can write c as

$$c = 1 + p(\sigma_L^2/\sigma_p^2)[1/(1 + \sigma^2/n\sigma_p^2)] .$$

Letting

$$p' = p/(1 + \sigma^2/n\sigma_p^2) ,$$

we have

$$c = 1 + (\sigma_L^2/\sigma_p^2)p' .$$

Then, we want

$$F_\alpha(\lambda, \lambda(p-1)/F_p[\lambda, \lambda(p-1)]) \leq (1 + \sigma_L^2/\sigma_p^2)p' .$$

The inequality above involves not only λ , p , and n , but also the variance ratios σ_L^2/σ_p^2 and σ^2/σ_p^2 . An estimate of the first ratio can be obtained from aerial data. Determination of the second ratio is more difficult. It should be noted that p' approaches p as n becomes large or σ^2 becomes small. We expect σ^2/σ_p^2 to be less than 1. Hence, p' is about p for $n \geq 5$. Table II shows solutions to the inequality for $\beta = 0.70$, $\sigma_L^2/\sigma_p^2 = 0.5$ and $\sigma^2/\sigma_p^2 = 0.6$. These variance ratios roughly correspond to the red sandstone area of the Arminto quadrangle. Many combinations of λ , p , and n values satisfy the inequality. In Table II, LP is the number of sites that would have to be sampled.

Table II shows the relative insensitivity of the solution sample size to n . For the red sandstone area of the Arminto quadrangle, we see that $\lambda = 15$ lines, $p = 4$ positions on each line, and $n = 2$ samples should be adequate.

TABLE II

SAMPLE SIZES FOR VARIOUS CONDITIONS

Note that: $\rho = 0.70$; $\sigma_L^2/\sigma_P^2 = 0.5$; and $\sigma^2/\sigma_P^2 = 0.6$.

N = 2			N = 5			N = 8			N = ∞		
L	P	LP	L	P	LP	L	P	LP	L	P	LP
2	66	132	2	57	114	2	55	110	2	51	102
3	21	63	3	18	54	3	18	54	3	16	48
4	13	52	4	12	48	4	11	44	4	11	44
5	10	50	5	9	45	5	9	45	5	8	40
6	8	48	6	8	48	6	7	42	6	7	42
7	7	49	7	7	49	7	6	42	7	6	42
8	6	48	8	6	48	8	6	48	8	5	40
11	5	55	11	5	55	11	5	55	11	4	44
18	4	72	18	4	72	18	4	72	18	3	54

V. TIME-SERIES ANALYSIS OF AERIAL RADIOMETRIC DATA

Data from aerial radiometric surveys, as in many other fields of science, are collected sequentially in space or time. As a result, the observations are frequently not independent, thus invalidating many of the commonly used statistical techniques. A study was performed by Data Information Associates to determine the usefulness of time-series analysis methodology in the analysis of aerial radiometric data.

The particular application addressed by the study was the estimation of components of variation for use in designing ground-based sampling experiments. Methodology for sample-size determination was presented and demonstrated using data from the Armino, Wyoming, quadrangle. A report presenting the details of the study was provided to BFEC.

VI. COMPUTING THE COVARIANCE MATRIX OF LARGE DATA SETS

A. The Algorithms

We completed additional comparisons of methods for computing the covariance matrix of a large data set. Six algorithms were compared on three types of data. The algorithms are as follows:

- I: textbook two pass,
- II: textbook two pass (double precision),

- III: West,³
- IV: West³ (double precision),
- V: generalized Youngs and Cramer,⁴ and
- VI: generalized Youngs and Cramer⁴ (double precision).

Algorithm I is described computationally as follows:

ALGORITHM I

```

 $\bar{x} = 0$ 
 $\bar{y} = 0$ 
for  $i = 1, n$  do
     $\bar{x} = \bar{x} + x(i)$ 
     $\bar{y} = \bar{y} + y(i)$ 
end
 $\bar{x} = \bar{x}/n$ 
 $\bar{y} = \bar{y}/n$ 
 $c = 0$ 
for  $i = 1, n$  do
     $c = c + [x(i) - \bar{x}] * [y(i) - \bar{y}]$ 
end
 $cov = c/(n - 1)$ 

```

Algorithm III, the method described by West,³ is as follows:

ALGORITHM III

```

 $\bar{x} = x(1)$ 
 $\bar{y} = y(1)$ 
 $c = 0$ 
for  $i = 2, n$  do
     $ax = x(i) - \bar{x}$ 
     $rx = ax/i$ 

```

```

 $\bar{x} = \bar{x} + rx$ 
 $ay = y(i) - \bar{y}$ 
 $ry = ay/i$ 
 $\bar{y} = \bar{y} + ry$ 
 $c = c + (i - 1) * ax * ry$ 
end
 $cov = c/(n - 1)$ 

```

Algorithm V is the Chan et al.⁵ generalization of the Youngs and Cramer⁴ algorithm.

Computationally we carry out this algorithm with an inner loop in which c is computed and updated two at a time and an outer loop in which the value is updated KBUF at a time. KBUF is some buffer size such that KBUF of the elements x_i and KBUF of the elements y_i will fit into memory at once, KBUF in our test was 10 000.

The inner loop for Algorithm V is as follows:

ALGORITHM V

```

txold = x(1) + x(2)
tyold = y(1) + y(2)
 $\bar{x} = txold/2$ 
 $\bar{y} = tyold/2$ 
m = 2
 $c = [x(1) - \bar{x}] * [y(1) - \bar{y}] - [x(2) - \bar{x}] * [y(2) - \bar{y}]$ 
for k = 3, KBUF - 1 (in steps of 2)
    tx = x(k) + x(k + 1)
    ty = y(k) + y(k + 1)
    cnew = [x(k) + tx/2] * [y(k) - ty/2]
           + [x(k + 1) - tx/2] * [y(k + 1) - ty/2]
    c1 = m/[2 * (m + 2)]
    c2 = 2/m
     $c = c + cnew + c1 * (c2 * txold - tx)$ 
            $* (c2 * tyold - ty)$ 
    txold = txold + tx
    tyold = tyold + ty
    m = m + 2

```

end
 $\text{cov} = c/(n - 1)$.

The outer loop is analogous and accumulates c in steps of KBUF. If KBUF is odd, we see in the above loop that not all elements from 3 to KBUF - 1 are included in the computations. Hence, some "clean up" is required if KBUF is not even or if n is not an exact multiple of KBUF.

B. Computational Results

The results in this section were obtained on a Digital Equipment Corporation VAX 11/780 computer. Single-precision accuracy is 7 to 8 digits and double-precision accuracy is typically 16 digits. Quadruple-precision accuracy with user-supplied software is 32 digits. The data used to derive the results given in Table III were normal random variates having a mean of 1.0 and several different specified variances.

TABLE III
 NUMBER OF CORRECT DIGITS^a FOR DATA TAKEN FROM $n(1.0, \sigma^2)$

σ^2	Algorithm					
	I	II	III	IV	V	VI
1.0	4.75	7.00	4.81	6.95	6.19	6.30
10^{-1}	4.97	7.00	4.81	6.85	6.094	6.28
10^{-2}	5.00	7.00	4.52	6.76	6.13	6.38
10^{-3}	4.57	7.00	4.45	6.92	5.07	5.19
10^{-4}	4.33	7.00	3.41	6.83	3.54	4.21
10^{-5}	3.49	7.00	3.25	6.88	3.10	3.20
10^{-6}	3.76	7.00	1.56	6.82	1.83	1.84
10^{-7}	2.57	7.00	2.60	6.78	-1.26	-1.24
10^{-8}	2.38	7.00	1.53	6.65	-1.46	-1.36
Ave. Time (s)	93.99	103.43	57.09	71.74	57.12	64.41

^aThe number of correct digits is $-\log_{10}$ of the relative error.

In each case, the results were averaged over elements of a 2-by-2 covariance matrix. The "correct" answers were taken to be those resulting from using the two-pass algorithm with quadruple-precision arithmetic.

The data summarized in Table IV were generated from a uniform random number generator. We generated three vectors x, w, and y using the following algorithm:

```

for i = 1, n do
  x(i) = RAND(0)
  w(i) = RAND(0) + 12.0 * x(i) + 2000.0
  y(i) = RAND(0) + x(i) + 2000.0
end

```

Thus, the 2-by-2 covariance matrix for w and y 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 (KBUF = 10 000).

TABLE IV
NUMBER OF CORRECT DIGITS^a FOR DATA TAKEN FROM UNIFORM (0,1)

	Algorithm					
	I	II	III	IV	V	VI
	0.37	7.00	4.57	6.82	5.10	6.20
Ave. Time (s)	93.55	103.96	48.09	74.21	57.08	62.69

^aThe number of correct digits is $-\log_{10}$ of the relative error.

The entries in both tables are reported in terms of correct digits. This number is computed as $-\log_{10}(E)$, where E is the relative error. The relative error is computed as

$$| \text{computed answer} - \text{correct answer} | / | \text{correct answer} |.$$

Table V contains the results obtained from running the six codes described above on NURE aerial radiometric data collected over a portion of the Rawlins 1:250000 National Topographic Map Service (NTMS) quadrangle. The data analyzed consist of that portion of the Rawlins quadrangle surveyed with rotary wing aircraft and consist of 56 039 data points. The data were collected by Geometrics, Inc.⁶

Table VI contains the actual elements of the upper triangle of the 3-by-3 covariance matrix computed for Rawlins data (T_{jk} , B_i , k).

VII. CONCLUSIONS

From the results in Tables III-VI, we see that the double-precision West algorithm (method IV) performed well in all cases for our data sets of size 150 000. The double- and single-precision Youngs and Cramer method (methods V and VI) performed well unless the standard deviation of the data was less than about half the single-word length accuracy. The two-pass algorithm gave good results only when double precision was used, and then at the expense of time (about 49% more time than the West double precision).

TABLE V
NUMBER OF CORRECT DIGITS^a FOR RAWLINS DATA

	Algorithm					
	<u>I</u>	<u>II</u>	<u>III</u>	<u>IV</u>	<u>V</u>	<u>VI</u>
	5.15	7.00	5.33	6.88	5.73	6.00
Time (s)	75.11	82.38	45.59	55.98	45.13	47.14

^aThe number of correct digits is $-\log_{10}$ of the relative error.

TABLE VI
ELEMENTS OF THE COVARIANCE MATRIX FOR RAWLINS
(T_l, B_i, K) DATA^a

I	(textbook two pass)	$\begin{bmatrix} 173.9744 & 32.93773 & 251.0809 \\ & 112.7869 & 76.73734 \\ & & 1109.824 \end{bmatrix}$
II	(textbook two pass, double precision)	$\begin{bmatrix} 173.9788 & 32.93769 & 251.0818 \\ & 112.7848 & 76.73762 \\ & & 1109.842 \end{bmatrix}$
III	(West)	$\begin{bmatrix} 173.9764 & 32.93757 & 251.0817 \\ & 112.7834 & 76.73788 \\ & & 1109.832 \end{bmatrix}$
IV	(West, double precision)	$\begin{bmatrix} 173.9788 & 32.93769 & 251.0818 \\ & 112.7848 & 76.73762 \\ & & 1109.842 \end{bmatrix}$
V	(generalized Youngs and Cramer)	$\begin{bmatrix} 173.9789 & 32.93773 & 251.0821 \\ & 112.7855 & 76.73796 \\ & & 1109.838 \end{bmatrix}$
VI	(generalized Youngs and Cramer, double precision)	$\begin{bmatrix} 173.9789 & 32.93771 & 251.0821 \\ & 112.7856 & 76.73767 \\ & & 1109.838 \end{bmatrix}$

---

^aThe matrix elements are rounded to 7 digits each.

For large data sets it is evident that double precision may be necessary to achieve high accuracy. When analyzing data sets of more than 150 000 data points, even higher precision may be necessary.

Because of the consistently good results given by the double-precision West algorithm, we recommend its use for large data sets.

VIII. RELATIONSHIP BETWEEN DOE ESTIMATES OF QUADRANGLE URANIUM INVENTORY AND AERIAL RADIOMETRIC RECONNAISSANCE DATA

The objective of the study was to examine the feasibility of using the Aerial Radiometric Reconnaissance (ARR) data to rank quadrangles in an order similar to that of the Department of Energy whole quadrangle estimates of uranium inventory (UNIV). Quadrangle uranium inventory is defined to be all

uranium (past production, reserves, and potential resources above 0.1% U_3O_8) expressed in tons of U_3O_8 , for a given quadrangle.

A. Background

Data tapes were obtained for 41 quadrangles. Because of difficulties in extracting data from many of the tapes, only 18 quadrangles were used.

The UNIV estimates were obtained for 15 of the 18 quadrangles from the 1980 Texas Instruments, Inc., report.⁷ The UNIV estimates for the remaining three quadrangles were unavailable. Five of the 15 UNIV estimates were zero. BFEC personnel informed us that a zero did not indicate zero inventory, but indicated that the quadrangle was not evaluated. Therefore, these five quadrangles were omitted from the analysis. Table VII lists the 10 quadrangles and their UNIV estimates.

B. Analysis

One hundred twenty statistics were chosen by Los Alamos and BFEC personnel for possible use in ranking the quadrangles. The statistics are based on the following 10 variables: ^{214}Bi (Bi), ^{208}Tl (Tl), ^{40}K (K), gross counts (G), Bi/Tl, Bi/K, Tl/K, Bi/G, Tl/G, and K/G. The following 12 statistics were computed for each of the 10 variables: A_{90}/M , A_{95}/M , A_{99}/M , S_{90}/A_{90} , S_{95}/A_{95} , S_{99}/A_{99} , P_1 , P_5 , P_{10} , P_{90} , P_{95} , and P_{99} where A_i and S_i are the mean and standard deviation of all records larger than the i^{th} percentile, M is the median of all records and P_i is the i^{th} percentile.

TABLE VII
DATA USED IN RANKING STUDY

<u>Quadrangle</u>	<u>Quadrangle Identification</u>	<u>UNIV (tons)</u>
Greeley	1	39 511
Mesa	4	37 941
Lamar	6	915
Green Bay	7	4
Hutchinson	8	311
Rice Lake	9	417
Manhattan	11	379
Sherman	13	5 974
Rock Springs	14	28 181
Casper	15	975 107

These 120 statistics were individually plotted against UNIV. Figures 1-5 show five of the several plots that demonstrate a monotone relationship between UNIV and a statistic. The numbers plotted on these figures are the quadrangle identification numbers shown in Table VII. These few plots show that there is a strong relationship between UNIV ranking and the ranking of some of the statistics under consideration. In all five figures, the Lamar and/or Green Bay UNIV estimates (6 and 7) appear to be outliers.

From this preliminary study, there appears to be a good possibility of using aerial radiometric data to rank quadrangles approximately according to their UNIV values.

IX. DISCRIMINANT ANALYSIS

During the reporting period, the discriminant analysis code discussed in the Los Alamos Laboratory report of M. E. Johnson⁸ was modified for transfer to BFEC. An open-file report documenting the procedure and presenting examples is currently in preparation.

Fig. 1.
Plot of P_{99} versus UNIV for T2.

Fig. 2.
Plot of A_{95}/M versus UNIV for Tx.

Fig. 3.
Plot of A_{99}/M versus UNIV for gross counts.

Fig. 4.
Plot of P_{95} versus UNIV for Tl.

Fig. 5.
Plot of Agg/M versus UNIV for Tl.

X. CLUSTER ANALYSIS

A study of the applicability of cluster analysis to the analysis of aerial radiometric data was conducted during the reporting period. The investigation centered on the development of a procedure for use on large data sets where there is little or no prior knowledge about the region being studied.

Some clustering techniques (such as k-means) are capable of handling large numbers of data points, but these techniques assume prior knowledge concerning the number of clusters and their centroids. Other techniques (for example, hierarchical) require no prior knowledge, but are severely limited in the number of data points they can handle. Two methods of dealing with these limitations were considered.

The first involves using a hierarchical method on a random sample from the large data set to estimate the number of clusters and their centroids. This information is then used as input to a k-means procedure and applied to the entire data set. For hierarchical clustering to be efficient, the sample must be relatively small (fewer than 200 points). However, it appears that if the sample is less than about 10% of the entire data set, the results will not be stable. The factors affecting stability are the number of clusters and the number of data points per cluster. In general, this method is not recommended if the data set being analyzed contains more than 2000 points.

The second method involves the use of principal components analysis to estimate the number and locations of clusters. This method does not have the sample-size limitations of the first method and appears to lead to more stable results.

The two methods have been tested and illustrated using the Copper Mountain* detail aerial radiometric data and the Precambrian data from the same survey. These two data sets contain 9331 and 6991 data points, respectively. Analysis of the Precambrian data showed four clusters, two of which are interesting. One is the metasediments present in the area and the other, contained entirely within the Canning Stock, contains the highest mean uranium values in the area.

---

*"Aerial Gamma Ray and Magnetic Survey, Copper Mountain Area of the Owl Creek Mountains, Wyoming," prepared for the US Department of Energy by Geometrics, Inc., 1978.

A joint Los Alamos-BFEC open-file report giving details of this study is being prepared. Publication is expected early in 1982.

XI. SHORT COURSES

Short courses on factor analysis, discriminant analysis, cluster analysis, time-series analysis and analysis of variance were presented in Grand Junction.

XII. REPORTS AND PAPERS

In addition to the reports mentioned above, open-file reports on the following topics were written estimating percentiles in aerial radiometric data with and without distributional assumptions, outlier detection, and comparison of commercial statistics software packages.

REFERENCES

1. M. G. Kendall and A. Stewart, The Advanced Theorey of Statistics, Vol. 2, (Hafner Publishing Co., New York, 1967).
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 National Laboratory report LA-8508-PR (August 1980).
3. D. H. D. West, "Updating and Variance Estimates: An Improved Method," Comm ACM 22, No. 9, 532-535 (September 1979).
4. E. A. Youngs and E. M. Cramer, "Some Results Relevant to Choice of Sum and Sum of Product Algorithms," Technometrics 3, No. 3, 657-665 (August 1971).
5. 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 report STAN-CS-79-773 (November 1979).
6. Geometrics, Inc., "Aerial Gamma Ray and Magnetic Survey: Rock Springs, Rawlins, and Cheyenne Quadrangles, Wyoming, and the Greeley Quadrangle, Colorado," US Department of Energy open-file report GJBX-17(79) (1979).
7. Texas Instruments, Inc., "Interpretation of NURE Aerial Radiometric and Hydrogeochemical and Stream Sediment Reconnaissance Data," US Department of Energy open-file report GJBX-32(81) (October 1980).
8. M. E. Johnson, J. A. Howell, C. K. Jackson, D. Patterson, R. J. Beckman, K. Campbell, and T. R. Bement, "Geostatistics Project of the National Uranium Resource Evaluation Program, April-September 1980," Los Alamos National Laboratory report LA-8734-PR (March 1981).