141
4-14-81

Dr. 2520

LA-8734-PR

Progress Report

MASTER

R-3516

Geostatistics Project of the National Uranium Resource Evaluation Program

April—September 1980

University of California

LOS ALAMOS SCIENTIFIC LABORATORY
Post Office Box 1663 Los Alamos, New Mexico 87545

GEOSTATISTICS PROJECT OF
THE NATIONAL URANIUM RESOURCE EVALUATION PROGRAM
APRIL-SEPTEMBER 1980

by

M. E. Johnson, J. A. Howell, C. K. Jackson, D. Patterson, R. J. Beckman,
K. Campbell, and T. R. Bement

ABSTRACT

A large computer code was written to perform a number of discriminant analysis procedures on aerial radiometric data. Work on percentile estimation, using the normal and log-normal probability distributions, was extended. Additional work was performed on methods of computing with large data sets. Attempts are being made to evaluate the behavior of principal components analysis relative to element distribution in a survey area. We also provided general statistical consulting in such areas as discriminant analysis, filtering, and kriging.

I. INTRODUCTION

This report outlines the activities and progress of the Los Alamos National Laboratory on the Geostatistics Project during the April-September 1980 time period. 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.

As part of the NURE program, the Grand Junction office has been conducting aerial radiometric surveys over various portions of the United States since 1974. The data collected in this part of the program include multichannel observations in the gamma-ray portion of the spectrum (0.4 to 3.0 MeV). From this radiometric data the contribution of uranium (from ^{214}Bi), thorium (from ^{208}Tl) and potassium (from ^{40}K) to the total gamma-ray activity is determinable.

The purpose of the aerial radiometric surveys[1] is to:

(1) map the regional distribution of the near-surface abundances of potassium, uranium, and thorium;

(2) disseminate basic data to industry and/or interested parties;

(3) identify lithologic limits favorable for uranium deposition; and

(4) indicate those areas of the country where there is the greatest probability of finding new (uranium) deposits, including new types of deposits.

The Geostatistics Project at Los Alamos, in close cooperation with the Grand Junction office of DOE, applies statistical methods to the analysis of aerial radiometric data. To handle a broad range of problems related to the NURE, the Laboratory maintains a close statistical consulting relationship with the DOE Grand Junction office and the Bendix Field Engineering Corporation (BFEC) in Grand Junction.

During this reporting period, we did additional work on methods of computing with large data sets, with particular attention given to the problem of obtaining selected sample percentiles.

A large computer code was written to perform a variety of discriminant analysis procedures on aerial radiometric data. This code includes the standard linear and quadratic discriminant analysis procedure as well as some robust and partial discriminant techniques.

Work on percentile estimation using the normal and log-normal probability distributions was extended,[2] as was previous work involving the use of principal components analysis.[3]

We also provided general statistical consulting in such areas as discriminant analysis, kriging, and filtering.

II. LARGE DATA BASE ANALYSIS

During the reporting period we made modifications and improvements to our program that computes percentiles of large data sets. These modifications make allowances for special data problems, such as data values equal to 0 and many equal data points.

We began work on a paper describing the algorithm for finding percentiles of large data sets. This paper, expected to be submitted early next year, describes the algorithm, its use, and its limitations. There is a detailed description of the input parameters that the user can vary to meet the particular needs of the problem.

We completed the computation and storage of 265 statistics for 20 quadrangles based on the following 10 variables: thallium, bismuth, potassium, gross count, Tl/gross, Bi/gross, K/gross, Bi/K, Bi/Tl, and Tl/K. These statistics are stored in a data base on the VAX 11/780, and we have begun plotting portions of the data to find a uranium favorability index.

III. DISCRIMINANT ANALYSIS

Discriminant analysis methods have been employed previously[4] in the analysis of aerial radiometric data to classify regions as favorable or unfavorable. These procedures are based on the development of data training consisting of records corresponding to geologic formations known to contain uranium deposits as well as barren formations. Subsequent to the effort reported in Ref. 4, several new approaches and methods in discriminant analysis have emerged in the statistical literature. These offer considerable potential in the analysis of aerial radiometric data.

Data from survey areas in the south Texas coastal plain are being used to test the applicability of these new techniques. Examples of geological questions being addressed are:

(1) Do the favorable geologic formations, such as Catahoula, Whitsett, Goliad, and Oakville, have a common radiometric signature?

(2) Do individual geologic formations vary throughout the south Texas coast? For example, does the Catahoula formation in the Seguin-Houston quadrangles have the same signature as in the Catahoula formation in the Corpus Christi-Laredo quadrangles?

(3) Can favorable and unfavorable formations be identified and their signatures employed to classify unknown formations? In other words, can discriminant analysis be used to identify formations worthy of intensive

exploration and to avoid formations with limited uranium mineralization potential?

Thus far we have concluded that discriminant analysis can be usefully employed to address these questions. The next section briefly describes the procedures that have been implemented into a computer code for automating the analyses. The final section describes preliminary findings and work in progress.

A. Discriminant Analysis Procedures

Until recently, linear and quadratic discriminant analyses have been the dominant procedures used in classification problems. These techniques have optimality properties if the populations are governed by multivariate normal distributions, but otherwise, these techniques are not very robust. When applying these methods, one must exercise considerable caution and must investigate the appropriateness of the assumptions.

New approaches to extend the range of applicability of the classical discriminant procedures have appeared in the statistical literature. Both linear and quadratic discriminant procedures involve the estimation of first- and second-order moments. Broffitt, Randles and Hogg[5] suggested the use of robust Huber-type estimates for use in the linear discriminant function (LDF) and the quadratic discriminant function (QDF). This approach works well on training sets with observations that appear to be "outliers." Conover and Iman[6] advocate another approach that is designed to control for outliers. They suggest replacing the raw data with the marginal ranks and then applying the usual LDF or QDF methods. This approach offers high efficiency under normal situations and superior performance under many alternatives.

Competing discrimination procedures are typically evaluated on the basis of the estimated misclassification rates. In the two-population problem, these error rates are defined in terms of the conditional probabilities

$$P_1 = P(\underline{Z} \text{ is classified in } \pi_2 / \underline{Z} \epsilon \pi_1)$$

and

$$P_2 = P(\underline{Z} \text{ is classified in } \pi_1 / \underline{Z} \epsilon \pi_2) \quad ,$$

where π_i denotes population i and $\underline{Z}$ denotes the observation to be classified. If the estimated error rates are excessive, then partial discriminant analysis should be considered. Here, an additional option, that of not classifying $\underline{Z}$, is included.

Johnson and Beckman[7] presented an asymptotically distribution-free procedure that controls the conditional probabilities of misclassification. That is, it controls the error rates among those $\underline{Z}$s for which classification was attempted, but ignores observations that were not classified. For applying partial discriminant analysis to the aerial radiometric data, Johnson and Beckman conducted a simulation study to estimate the sample sizes required. They determined that in the two-population, two-component problem, sample sizes of 100 or larger are sufficient.

A computer code was written for applying these techniques to the aerial radiometric data. The forced discriminant procedures that are included in the code are the classical LDF and QDF and robust estimates in the LDF and QDF. These procedures can be applied to both the raw data and the ranks. Partial discriminant analysis can be used with each of these discriminant functions. Other information produced by the code includes

(1) sample first- and second-order moments;
(2) canonical variable analysis;
(3) estimates of misclassification rates;
(4) some normal theory results; and
(5) classification lists.

The next section presents preliminary findings and describes work in progress.

B. Preliminary Conclusions and Current Effects

The computer code is currently involved in several efforts to address some of the previously mentioned geologic questions. We are preparing a joint Los Alamos-BFEC report describing the computer code and the interpretation of its output. This report will be concluded in early 1981.

Preliminary testing of the code has been carried out with data from the south Texas coast. Preliminary conclusions, subject to considerably more verification, include the following:

(1) The Catahoula and Whitsett formations, both favorable for uranium, appear to be indistinguishable. The Goliad and Oakville formations have some

statistical overlap, but can be distinguished from the Catahoula and Whitsett formations.

(2) The Catahoula formation seems to have a homogenous radiometric signature through the south Texas coast, within the Rio Grande embayment.

(3) Some areas of the Beaumont formation look like the Catahoula.

A joint Los Alamos-BFEC paper is being prepared to present these findings and to highlight the strengths and weaknesses of the discriminant analysis methods employed.

IV. COMPARISON OF METHODS OF PERCENTILE ESTIMATION

The comparison of normal and log-normal methods of percentile estimation that was reported in Ref. 2 was presented at a joint session of the American Statistical Association and the International Association for Mathematical Geology at the Joint Statistical Meetings in Houston, Texas, in August 1980.

This work was expanded to include the use of sample percentiles, independent of any distributional assumptions. We examined the Browns Park and Battle Spring formations in the Rawlins, Wyoming, quadrangle[8] and compared the three percentile estimation methods. A general conclusion was that, when sample size permits, the sample percentiles should be used, rather than run the risk of making an incorrect distributional assumption. The percentile estimation code discussed in Sec. II of this report allows one to obtain a limited number of sample percentiles without a complete sorting of the data.

A joint Los Alamos-BFEC paper on this subject has been submitted to the Journal of the International Association for Mathematical Geology for consideration for publication.

V. PRINCIPAL COMPONENTS ANALYSIS

We are conducting a study to analytically assess the behavior of the eigenvalues and vectors in a principal components analysis relative to the distribution of potassium, uranium, and thorium in a survey area. The study is based on a simple model where it is assumed that there are three regions in

TABLE I

MODEL FOR STUDYING BEHAVIOR OF PRINCIPAL COMPONENTS ANALYSIS

Region in Survey Area	Proportion	Count Rate Due To K	U	Th
1	P_1	B_1	B_2	B_3
2	P_2	$B_1 + S_1$	$B_2 + S_2$	$B_3 + S_3$
3	P_3	B_1	$B_2 + S_2$	B_3

the survey area and the count rates obtained over each of these regions is due to a background effect B plus a signal S as indicated in Table I. It is assumed that the region is a fraction P_i of the entire survey area. From Table I, it can be seen that in Region 1, each element is present only at background level. Region 2 is enriched with all three elements, and Region 3 is enriched with uranium only.

Under the above model, an analytic expression has been derived for the expected value of the sample covariance matrix. The eigenvalues and vectors of this matrix can be obtained numerically for various values of the backgrounds and signals. A joint Los Alamos-BFEC report on the results of this study, along with geologic implications, is being prepared.

VI. GENERAL STATISTICAL CONSULTING

A. Spectral Enhancement

We provided statistical consulting to DOE and BFEC relative to the Science Applications, Inc. spectral enhancement contract. In particular, aspects of the MAZAS computer code for spectral enhancement were examined, and a report is being prepared for BFEC.

B. Effect of Soils on Radiometric Data

Statistical consulting is being provided to BFEC in an effort to determine if radiometric signatures can be developed for various soil types. A joint BFEC-Los Alamos report is planned when this study is completed.

C. Kriging

A kriging code was written to permit the mathematical overlaying of hydrogeochemical stream and sediment reconnaissance (HSSR) data with aerial radiometric data. This supports the Geostatistics Project as well as an effort of the Los Alamos Resource Characterization and Materials and Image Evaluation groups to merge multivariant data sets.

The overlaying was accomplished by interpolating each type of data to a rectangular grid using universal kriging. Universal kriging is a statistical method used to obtain spatially interpolated values from a set of irregularly spaced data points. The standard assumption is that the spatial covariance in a small region is dependent only on the distance between points. If this covariance function were known, an unbiased estimate of the elemental concentration at an unsampled location could be formed, together with an estimate of the variance. In practice, of course, the spatial covariance function (actually, a related function, the variogram) must be estimated.

For the HSSR data, an estimate of the concentration was formed at each point on a 1-km grid as a weighted average of nearby samples. The nearest five samples were used, provided there were five within 5 km of the grid point; also, all samples within 500 m were used, even when there were more than five. When there were fewer than two samples within 5 km of a grid point, no estimate was computed. The weighting of these samples depended on their distance from the grid point and on the estimated variogram.

For the aerial radiometric data, the first step in getting from the densely sampled flight lines to the 1-km rectangular grid was to smooth and subsample the noisy data along flight lines. This was done using an automated kriging procedure, by which the parameters of the variogram were estimated locally, depending on the local signal-to-noise ratio. The result was an adaptive filter. That is, in regions where there was little change in the data relative to the noise, the smoothed estimate was almost an unweighted average of measurements in a relatively long segment of the flight line (about 1.5 km). In regions where the signal was changing rapidly, the technique weighted nearby measurements much more heavily than more distant ones. These kriged estimates, together with error estimates, were computed every 0.4 km along each flight line. That is, they were computed at about one-tenth the density of the original one-second measurements.

From this point, interpolation onto the rectangular grid proceeded much as for the hydrogeochemical data. A variogram was computed and used to form an estimate at each grid point as a weighted average of the smoothed data (with associated errors). All of the smoothed data within 3.5 km of a given grid point were used in forming an estimate at that grid point.

These kriging codes have been used successfully by the Los Alamos Resource Characterization Group.

REFERENCES

1. P. H. Dodd, "Airborne Radiometric Reconnaissance: The Whys and Wherefores," in Summaries and Visual Presentations: Uranium Geophysical Technical Symposium, Grand Junction, Colorado, September 14-16, 1976 (US Energy Research and Development Administration and Bendix Field Engineering Corporation, Grand Junction, Colorado, 1976), pp. 13-20.

2. K. Campbell, T. R. Bement, J. A. Howell, R. J. Beckman, 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. F. L. Pirkle, K. Campbell, and G. W. Wecksung, "Principal Component Analysis as a Tool for Interpreting Radiometric Data," J. of Geology 88, 57-67 (1980).

4. T. R. Bement and D. E. Whiteman, "Geostatistics Project of the National Uranium Resource Evaluation Program, July-September 1978," Los Alamos National Laboratory report LA-7806-PR (January 1979).

5. J. D. Broffitt, R. H. Randles, and R. V. Hogg, "Distribution-Free Partial Discriminant Analysis," J. A. Stat. Assoc. 71, 934-939 (1976).

6. W. J. Conover and R. L. Iman, "The Rank Transformation as a Method of Discrimination with Some Examples," Communications in Statistics A9, 465-487 (1980).

7. R. J. Beckman and M. E. Johnson, "A Ranking Procedure for Partial Discriminant Analysis," Los Alamos National Laboratory, unpublished data 1980.

8. Geometrics, Inc., "Aerial Gamma-Ray and Magnetic Survey, Rock Springs, Rawlins, and Cheyenne Quadrangles, Wyoming, and the Greeley Quadrangle, Colorado," US Department of Energy report GJBX-17(79) (1979).