Mathematical Geology, Vol. 15, No. 1, 1983

SIMSAG: Integrated Computer System for Use in Evaluation of Mineral and Energy Resources[1]

Chang-Jo F. Chung[2]

A system of interactive graphic computer programs for multivariate statistical analysis of geoscience data (SIMSAG) has been developed to facilitate the construction of statistical models to evaluate potential mineral and energy resources from geoscience data. The system provides an integrated interactive package for graphic display, data management, and multivariate statistical analysis. It is specifically designed to analyze and display spatially distributed information which includes the geographic locations of observations. SIMSAG enables the users not only to perform several different types of multivariate statistical analysis but also to display the data selected or the results of analyses in map form. In the analyses of spatial data, graphic displays are particularly useful for interpretation, because the results can be easily compared with known spatial characteristics of the data. The system also permits the user to modify variables and select subareas imposed by cursor. All operations and commands are performed interactively via a graphic computer terminal. A case study is presented as an example. It consists of the construction of a statistical model for evaluating potential areas for explorations of uranium from geological, geophysical, geochemical, and mineral occurrence map data quantified for equal-area cells in Kasmere Lake area in Manitoba, Canada.

KEY WORDS: Mineral resources, graphic display.

INTRODUCTION

In the study of the quantitative evaluation of the mineral and energy resources of a region, the computer graphic and data management techniques link geoscience information and statistical analysis. Integrated interactive computer systems for data management, graphic display, and statistical analysis facilitate construction of a quantitative model for resource appraisal from geoscience data.

With such systems, an initial model is statistically tested on the data, and then the results of the analysis are immediately returned and display in map form on the user's terminal. Consequently, they can be used to guide the choice

[1]Manuscript received 21 Dec 1981. Paper presented at the Workshop on Interactive Graphic Computer Programs, 20–22 October 1981, Ottawa, Canada.

[2]Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario, Canada K1A 0E8.

0020-5958/83/0200-0047$03.00/0

of new models or modifications of the initial model. These new models then can be tested again. This interactive procedure for the construction of models can not be performed without a graphic display of the data and the results of analysis in map form, because the results can not be compared with known geological features usually published in map form.

The system of interactive graphic computer programs for multivariate statistical analysis for geoscience data, SIMSAG, is specifically designed to meet the preceding requirements. The system provides an integrated package that enables the user not only to perform several different types of multivariate statistical analysis but also to modify, transform, and recode variables, to display the data or the results of the analysis, and to select subareas by imposing conditions on the variables or by drawing polygons with the cursor. All operations and commands in the system are interactively performed on the user's terminal.

A case study concerning uranium mineralization in the Kasmere Lake area in Manitoba, Canada is presented to illustrate the use of SIMSAG.

OVERVIEW OF SIMSAG

Data Preparation

A region under study is first divided into nonoverlapping polygons and each polygon is called a cell. A cell is the basic unit of analysis for which measurements (observations) have been obtained. The data to be entered into SIMSAG consists of cells and each cell is characterized by the x and y coordinates for the geographic location of its central point and by the observed values of the variables it contains.

First the SIMSAG system file is generated by a separate program from the user's raw input data and information defining and describing the data such as file name, names of variables, and structure of the data. The SIMSAG system file contains the user's input data, the variable names and their formats supplemented by other information on the data as required for SIMSAG, and the boundary of the area to be used for graphic display. The system file may be permanently retained as a computer disc file. Thereafter the user has to attach to the system file whenever processing is desired. Then all information on the data is automatically entered into the system along with the data itself.

Data Modification and Selection

All data modifications such as selecting, deleting, modifying and adding variables, or selecting and deleting subareas are interactively performed. Before any modifications are made the user must enter the name of subfile from which subsequent modifications will be derived. This subfile could be the system file itself or one of the subfiles created in the current session. Upon completion, the system creates a new subfile with all modifications that have been made. If the

user intends to use this subfile later, he must name it also. Any subfile generated can be accessed, modified, or read by simply referring to it by its name.

At any one time in a session, up to 30 subfiles can be stored in the system. However, when some of subfiles are no longer required, the user can erase them by entering their names.

The system also provides the user with the capacity for generating a listing of the complete contents of the data in any subfile, including variable names and information such as the number of variables used. Furthermore, all cells in any subfile can be displayed on the user's terminal.

In order to create, access, and modify subfiles easily without regard to their physical positioning, a random access file is used instead of a conventional sequential file.

In the variable modifications, the system provides the user with the following capacities

(1) selecting or deleting variables
(2) defining new variables under new names and functions of the existing variables
(3) transforming variables into any functional form
(4) recording variables into binary form

The following two facilities to select or delete cells from a subfile are also available in the system

(1) specifying a criterion; those cells for which the criterion is met are selected or deleted
(2) drawing a polygon by using the cursor on the user's terminal; cells with centers within the polygon are selected or deleted

Multivariate Statistical Analysis

The system contains at present seven multivariate statistical techniques: principal components analysis, discriminant analysis, multiple regression analysis, stepwise regression analysis, logistic regression analysis, and Poisson regression analysis. Each type of analysis will be briefly described followed by a brief description of the operational procedures.

In the principal components analysis, the sample correlation matrix is used for computing the eigenvalues and their corresponding eigenvectors. The sample means and variances are printed together with all eigenvalues and eigenvectors. The discriminant analysis is for two populations only and the coefficients of the linear discriminant function are estimated from the sample means and the sample covariance matrix as suggested by Anderson (1958).

In the multiple regression and stepwise regression analysis, the general linear model is specified as

$$E(y_i) = X_i'b, \qquad \text{for} \quad i = 1, 2, \ldots, n \tag{1}$$

where $X_i' = (1, X_{1i}, X_{2i}, \ldots, X_{pi})$ represents the p explanatory variables for the ith cell; $b' = (b_0, b_1, \ldots, b_p)$ is the vector of regression coefficients to be estimated; y_i is for the dependent variable.

For the multiple regression analysis, the regression coefficients b are estimated by the ordinary least-squares (OLS) method. Simple statistics such as sample means and variances of the variables and standard deviations of the OLS estimates of the regression coefficients are computed. The analysis of variance table is also printed. When the sample correlation matrix used for estimating the regression coefficients is singular, a generalized inverse matrix is used instead and the analysis of variance is performed on the basis of "model not of full rank." In this case, the rank of the matrix is also printed as well as a warning message.

For the stepwise regression, the computational procedures of Efroymson (1962) are used and summary statistics are printed after each step. In addition to the estimates of regression coefficients and their corresponding standard deviations, an analysis of variance table is printed.

A drawback of the general linear model in (1) for estimating the probability of occurrence of an event is that the OLS estimates $E(y_i)$ may not lie between 0 and 1 as they should. An alternative model of representing the probability of occurrence of an event in terms of p explanatory variables, so that the estimates do lie between 0 and 1, is to postulate the logistic form

$$E(y_i) = e^{X_i'b}/(1 + e^{X_i'b}), \qquad \text{for} \quad i = 1, 2, \ldots, n \tag{2}$$

where X_i and b are as in eq. (1).

This model has been advocated by, among others, Cox (1970), to estimate the probability of occurrence of an event. To estimate b, Cox (1970) proposed to use the maximum likelihood (ML) estimator. In SIMSAG, the ML estimates are obtained by the scoring method (Rao, 1973).

In the Poisson regression analysis, a general linear model is specified as in eq. (1) without the constant term. However, the assumption of this model is that the dependent variable has a Poisson density. For example, the dependent variable has a Poisson density if the dependent variable is the sum of several independent Poisson variables.

To estimate the regression coefficients b, Jorgenson (1961) suggested a maximum likelihood method. The ML estimates are obtained by the scoring method as in the logistic analysis.

The OLS estimates are used as initial values for the iterations required for the scoring method. The analysis of variance table is computed on the basis of the estimated weights, and the individual observations have different variances depending upon the expected values, as the Poisson assumption of the model implies.

Upon choosing one of the types of analysis described, the title of the

analysis chosen is printed and a message asking for the name of subfile to be analyzed prompts the user to enter the name of this subfile. Then the names of variables to be analyzed have to be specified. For the regression analysis, the dependent variable and the independent (explanatory) variables are entered separately.

In addition, the critical probability levels of variables entering into the model and of variables being deleted from the model have to be specified for the stepwise regression analysis. In the logistic and Poisson regression analysis, the level of convergence and the maximum number of iterations should be specified for the interative procedure of the scoring method.

As soon as the user completes entering the necessary information, the system responds by printing the statistics and output for the analysis chosen. At the completion of the printout, one can obtain a graphic display of the results such as the expected values or residuals for the regression analysis, the discriminant scores for the discriminant analysis, or the first principal component scores for the principal components analysis.

CASE STUDY

Background

The Kasmere Lake area chosen for this study is in the northwest corner of the province of Manitoba, Canada. The specific focus of interest centers on the potential for uranium mineralization in the area. These experiments were part of the study by Bonham-Carter and Chung (1982) for integrating geoscience data for quantitative resource appraisals. The area was chosen mainly because of the availability of detailed coverage of regional geophysics, geology, lake sediment geochemistry, and mineral occurrences data.

The area has been studied by the Manitoba Department of Mines (Weber, Schledewitz, Lambs, and Thomas, 1975) and the geology background is described in detail in their report.

Data Base

The major part of the data was digitized and edited by Fabbri (1981) using an image processing computer system, GIAPP. From these computer processible images, a grid of square cells with a size of 3.17 km X 3.17 km shown in Fig. 1 was superimposed and the data were coded cell by cell.

Besides the geological map, which contains 33 lithological units (Weber, Schledewitz, Lambs, and Thomas, 1975), geophysical maps (Fabbri, 1981) containing topographic, aeromagnetic, gravity and radiometric information,

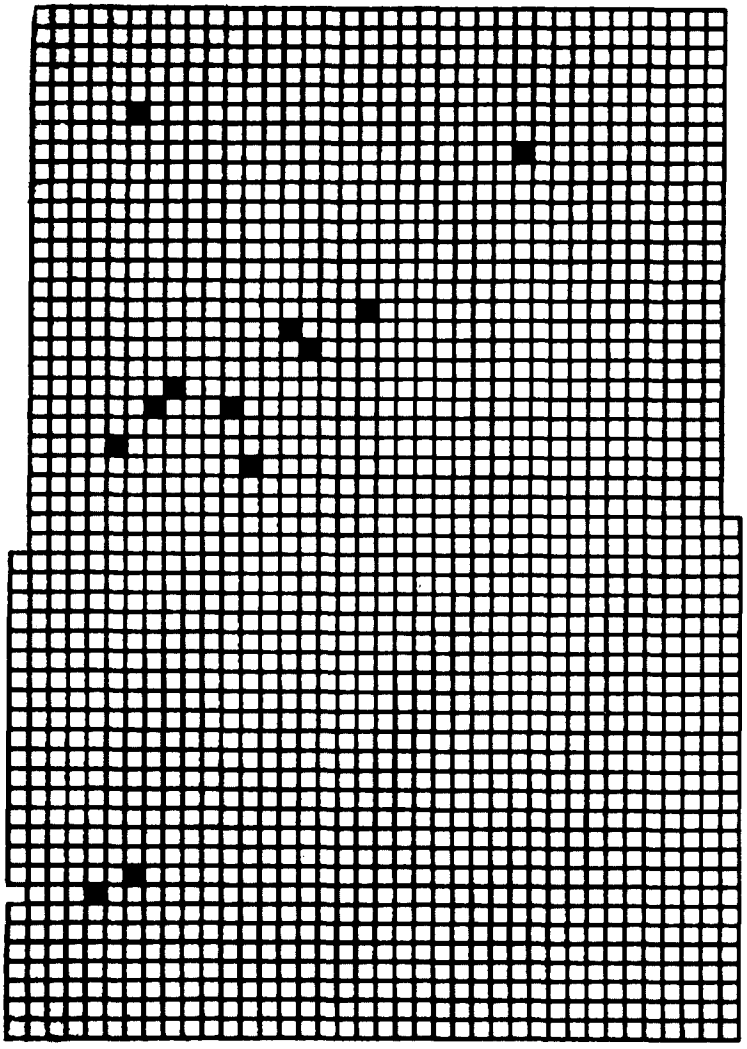

Fig. 1. Kasmere Lake area in northwest manitoba, Canada. A grid of square cells with a size of 3.17 X 3.17 km was superimposed. The shaded cells contain one or more of the known uranium occurrences.

11 lake-sediment geochemical maps displaying nickel, copper, uranium, cobalt, molybdenum, zinc, mercury, lead arsenic, iron, and manganese anomalies, and a uranium "mineral occurrence" map were quantified and stored in the data base (Bonham-Carter and Chung, 1982).

The area is covered by 1959 cells and among them, 12 cells contain one or more of the known uranium occurrences and shown in Fig. 1.

Experiment 1

This experiment deals with the problem of separating the cells with high uranium anomalies into those that fall on Aphebian metasediments and those that fall on Archean or Helikian igneous rocks by using all geochemical and geophysical data. The purpose of this experiment was to determine how well the geochemical and geophysical variables can be used to discriminate between these two types of cells.

The following steps were taken in a session with SIMSAG

(i) select cells with high uranium anomalies
(ii) define an indicator variable Y such that

$Y = 1$ for selected cells underlain Aphebian metasediments (see Fig. 2)
$Y = 0$ for the other cells

(iii) take all 17 geochemical and geophysical variables, call them V1-V17
(iv) perform stepwise regression analysis to select only the statistically significant variables among V1-V17 and to estimate the unknown coefficients
(v) compute $\hat{Y}$ as a linear combination of the selected variables and the corresponding estimated coefficients. Compare the results with the original Y.

All operations for computing and display were performed interactively. From the analysis it follows that by far the strongest indicators for Aphebian metasediments among the geochemical and geophysical variables, are aeromagnetic, thorium equivalent, and gravity data. These three variables are positive in the Archean and Helikian igneous rocks, negative in the Aphebian metasediments. Among the geochemical variables the metasediments are characterized by a high nickel-uranium-zinc association whereas the igneous rocks are characterized by a lead-molybdenum association. Figure 3 illustrates the predicted Aphebian metasediment environments (printed as 6, 7, 8, 9).

From this experiment, we may conclude that, using the geochemical and geophysical data, areas with high uranium anomalies underlain by Aphebian metasediments can be distinguished from those with high anomalies but underlain by igneous rocks. This information may also be useful in other areas because the geochemical and geophysical data can be obtained even when the bedrock is not exposed.

Experiment 2

In this experiment, uranium mineralization was predicted by relating all geochemical and geophysical variables with the known uranium occurrences.

The steps in this experiment are

(i) select the relevant geological variables, all 17 geochemical and geophysical variables, and uranium occurrence data

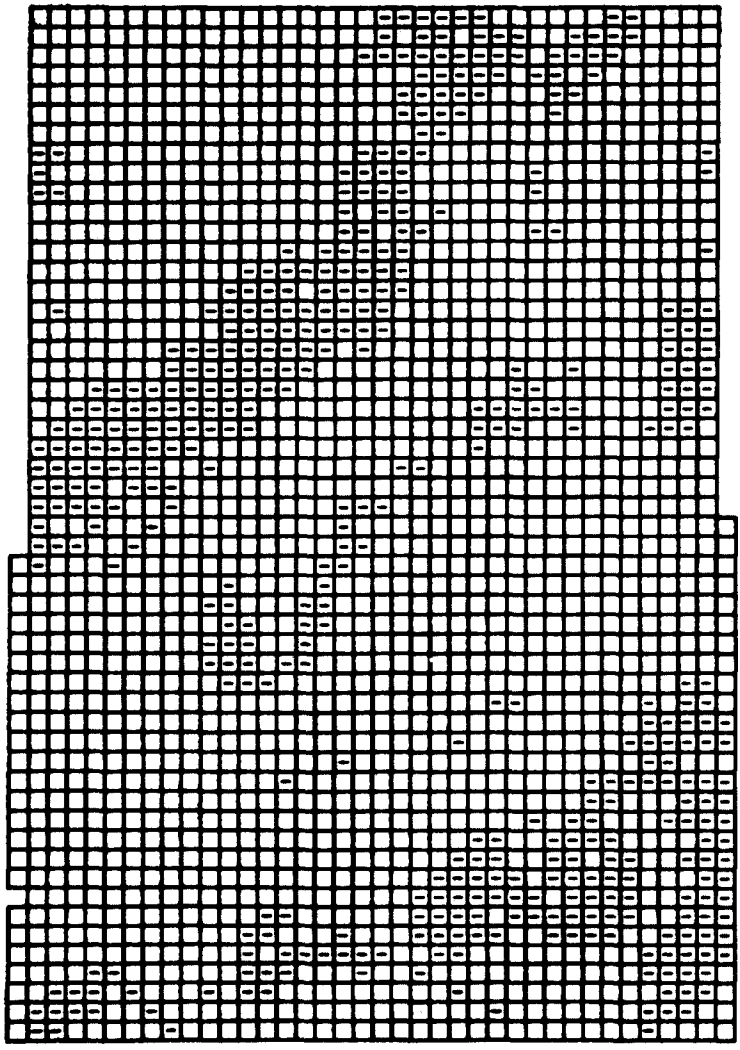

Fig. 2. The cells marked with 1 contain Aphebian metasediments.

(ii) select those cell with known uranium occurrences and display as reference (see Fig. 1).

(iii) select a control area as shown in Fig. 4 which includes most of the known uranium occurrences that are in Aphebian metasediments

(iv) define an indicator variable

$Y = 1$ for cells with the known occurrence

$Y = 0$ otherwise

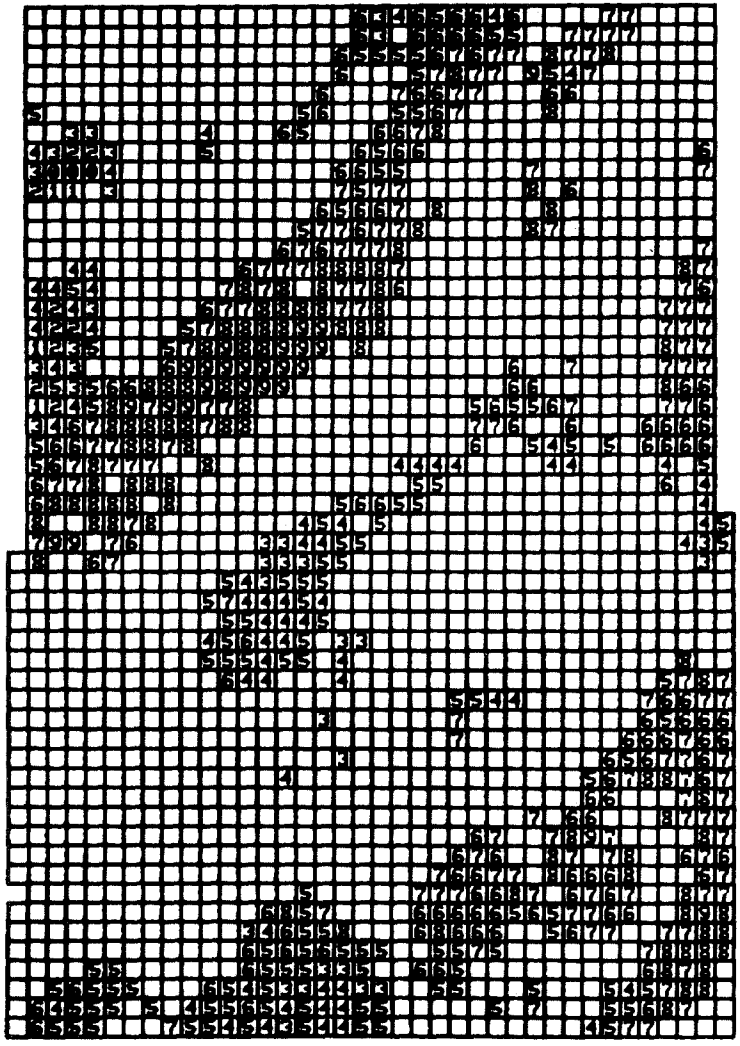

Fig. 3. Using the geochemical and geophysical data in Experiment 1, those cells with high uranium anomaly underlain by Aphebian metasediments are predicted in terms of probability. The cells with 6, 7, 8, and 9 are likely underlain by Aphebian metasediments. The predicted results can be compared with Fig. 2.

(v) perform multiple regression analysis and logistic analysis to estimate the relationships.

(vi) using the estimated equations, predict uranium mineralizations by means of probability index maps throughout the entire area (not only the control area) and display in Figs. 4 and 5.

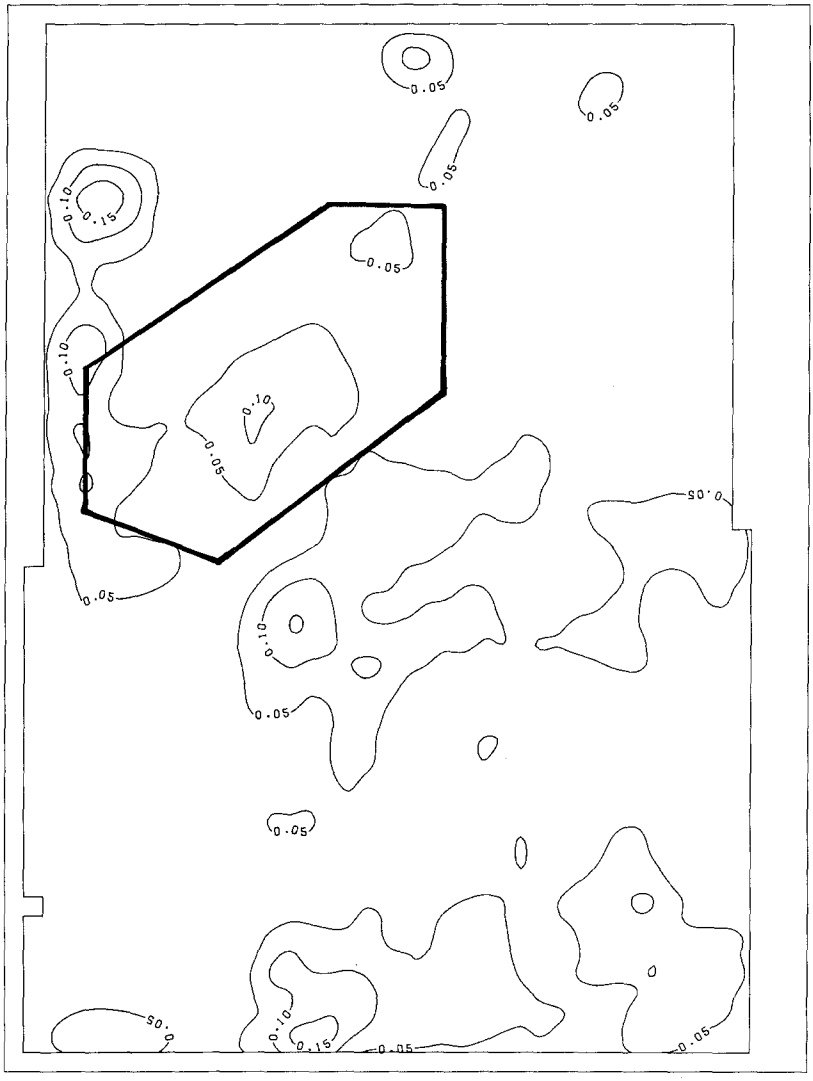

Fig. 4. The potential of uranium mineralizations is predicted in terms of probabilities estimated by multiple regression analysis using geochemical and geophysical data. The statistical analysis was performed within the control area enclosed by the polygon drawn.

The areas with high contoured value indicate higher potential of uranium mineralizations.

CONCLUDING REMARKS

Although the case study shown in this paper has demonstrated that useful results can be obtained by using SIMSAG in resource appraisal, the use of the

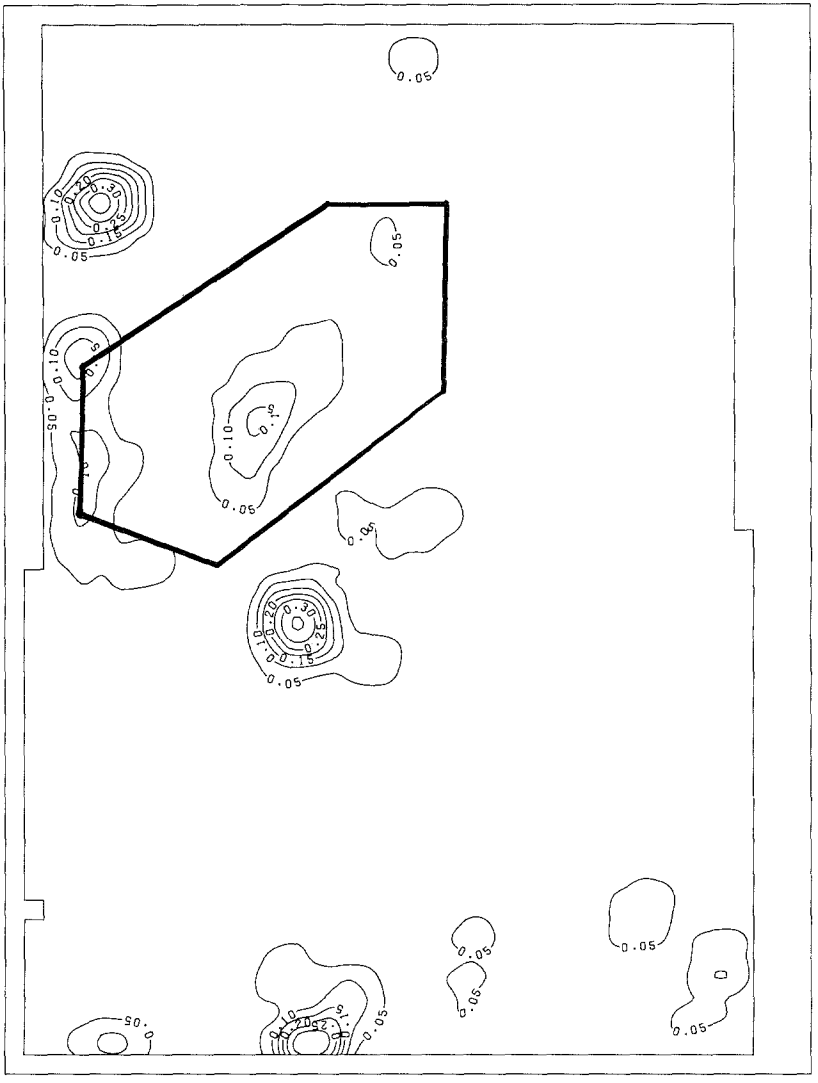

Fig. 5. Same as Fig. 4; constructed, however, by logistic regression analysis.

statistical models for evaluation of natural resources is still in the developing stage. In practice, most mineral and energy resource appraisals are performed by means of subjective methods–opinions are usually formulated by panels of experts on the basis of evidence from many sources and their experiences.

One of the difficulties in using the statistical techniques in resource appraisals is that the construction of a computer processible data file is a difficult task mainly because (1) it is laborious to quantify all available geoscience data; (2) the intensity of the exploration differs from one to another part of the

study area so that the level of detail of the relevant information differs from one subarea to another within the study area. For example, a part of the study area may be covered by a geological map at the scale of 1:50,000 but another part may be covered by that of 1:500,000; (3) some of the relevant data may occur outside of the study area.

SIMSAG has been developed as part of continuing effort by the Geological Survey of Canada into the methodology and application of multivariate statistical analysis to quantitative resource evaluation (see Agterberg, 1981).

The system is in FORTRAN and, at present, operational with a 4014/5 Telectronix Graphic terminal on a CDC CYBER 730 computer. A user's guide is in preparation and further information may be obtained from the author upon request.

REFERENCES

Agterberg, F. P., 1981, Application of image analysis and multivariate analysis to the mineral appraisal: Econ. Geol., v. 76, p. 1016–1031.

Anderson, T. W., 1958, An introduction to multivariate statistical analysis: Wiley, New York.

Bonham-Carter, G. F. and Chung, C. F., 1983, Integration of mineral resource data for Kasmere Lake area, Northwest Manitoba, with emphasis on uranium: Math. Geol., v. 15, n. 6, pp. 25–45.

Chung, C. F., and Agterberg, F. P., 1980, Regression models for estimating mineral resources from geological map data: Math. Geol., v. 12, no. 5, p. 473–488.

Cox, D. R., 1970, The analysis of binary data: Methuen,

Efroymson, M. A., 1962, Multiple regression analysis, *in* A. Ralston and H. S. Wilk (Eds.), Mathematical methods for digital computers: Wiley, New York.

Fabbri, A. G., 1981, Image processing of geological data, unpublished Ph.D. thesis, University of Ottawa.

Jorgenson, D. W., 1961, Multiple regression analysis of a Poisson process: JASA, v. 56, p. 253–255.

Rao, C. R., 1973, Linear statistical inferences and its applications, 2nd ed.: Wiley, New York, p. 366–374.

Weber, W., Schledewitz, D. C. P., Lambs, C. F., and Thomas, K. A., 1975, Geology of the Kasmere Lake Whiskey Jack area (Kasmere project): Pub. No. 74–2, Manitoba Department of Mines, Resources, and Environmental Management, 163 p.