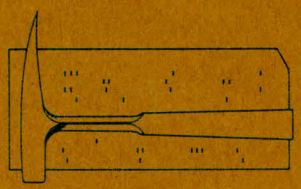
COMPUTER APPLICATIONS IN THE EARTH SCIENCES:

COLLOQUIUM ON TREND ANALYSIS

Edited by
DANIEL F. MERRIAM
and
NAN CARNAHAN COCKE



COMPUTER CONTRIBUTION 12
State Geological Survey
The University of Kansas, Lawrence
1967

Editor's Remarks

The "Colloquium on Trend Analysis" is the second in a series of meetings on Computer Applications in the Earth Sciences to be held at The University of Kansas. The Colloquium, sponsored by the Kansas Geological Survey, Department of Chemical and Petroleum Engineering, Department of Geography, and University Extension, affords participants an opportunity to convene and converse on a subject of mutual interest at an advanced level in an interdisciplinary atmosphere. By definition a colloquium is "a conversation".

Oral presentations are designed to encourage discussion - discussion, of course, is a basis for exchange of ideas, which is one of the chief purposes of this Colloquium. Many oral presentations may not coincide with written ones as recorded in these proceedings. To allow latest developments to be transmitted, however, discussion leaders cannot be confined to findings of 6 weeks ago. Such is the way of the computer!

Trend analysis was an obvious choice as one of the subjects. Geologists have long sought trends and it was only natural that this quantitative technique was one of the first to be utilized by them. Many papers dealing with applications of trend analysis in two, three, and four dimensions to geological, geophysical, and geochemical data have been published - many are cited in references of papers presented here. Most papers are concerned with distribution of various constituents in igneous and sedimentary rocks and geologic structure. Today approximately ninety papers can be found that treat this subject - this number of references serves to emphasize the importance of trend analysis as a sophisticated quantitative method.

The sponsoring organizations take this opportunity to thank all participants. Hopefully everyone will benefit from the interaction of active participation. The editors thank the authors for complete and wholehearted cooperation. Many have helped with various "chores" involved with preparations of the Colloquium, including Mrs. Alberta E. Bonnett, Mr. John C. Davis, Mr. Owen E. Spitz, and Mr. Richard F. Treece.

Indeed, conferences of this type seemingly serve a definite purpose and fill a particular need. By co-sponsoring the Colloquium, the Survey is fulfilling yet another obligation to industry and the profession, that of disseminating information of current and immediate interest and providing the avenue of exhange of information between people with mutual interests.

Comments and suggestions concerning the COMPUTER CONTRIBUTION series are welcome and should be addressed to the Editor. An up-to-date list of publications and available decks can be obtained by writing the Editor.

COMPUTER APPLICATIONS IN THE EARTH SCIENCES:

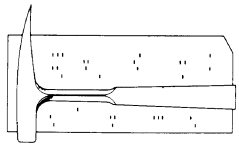
COLLOQUIUM ON TREND ANALYSIS

Edited by

DANIEL F. MERRIAM

and

NAN CARNAHAN COCKE



1967

CONTENTS

SELECTING THE "BEST" REGRESSION EQUATION[1/]

by

Norman R. Draper
University of Wisconsin

and

Harry Smith
University of North Carolina

INTRODUCTION

Suppose we wish to establish a linear regression equation for a particular response Y in terms of "independent" or predictor variables X_1, X_2, ... X_k. We assume this is the complete set of variables from which the equation is to be chosen and includes any functions such as squares and cross products or transformations such as log, roots, etc., thought to be desirable and necessary. Two opposite criteria of selecting a resultant equation are usually involved. They are as follow:

1. To make the equation useful for predictive purposes we should want our model to include as many X's as possible so that reliable fitted values can be determined.
2. Because of the costs involved in obtaining information on a large number of X's and subsequently monitoring them, we should like the equation to include as few X's as possible.

The compromise between these extremes is what is usually called "selecting the best regression equation." There is no unique statistical procedure for doing this, and personal judgment will be a necessary part of any of the statistical methods discussed. We shall describe two main procedures which have been proposed; both of these are in current use. Other current procedures which, in our opinion, are of less value in general, will be mentioned also. All these procedures do not necessarily lead to the same solution when applied to the same problem, although for many problems they will achieve the same answer. The two procedures to be discussed are:

1. Backward elimination
2. Stepwise regression.

To illustrate the procedures we shall use the data in a four-variable (k=4) problem given by A. Hald (1952, p. 647). The data are given in the first section of Table 4. The independent variables here are X_1, X_2, X_3 and X_4, and the dependent variable Y is X_5.

BACKWARD ELIMINATION PROCEDURE

The basic steps in the backward elimination procedure are:

1. A regression equation containing all variables is computed.
2. The "partial F-test" value is calculated for every variable treated as though it were the last variable to enter the regression equation.
3. The lowest partial F-test value, F_L is compared with a pre-selected significance level F_0;

(a) If $F_L < F_0$ remove the variable X_L which gave rise to F_L, from consideration and recompute the regression equation in the remaining variables; reenter stage 2.

(b) If $F_L > F_0$ adopt the regression equation as calculated.

Using the Hald data, we can illustrate this procedure as follows. (It is not intended that the section numbers below should correspond to those above.)

1. First, do the complete regression on all independent variables. In the example this means find the least-squares equation $\hat{Y} = f(X_1, X_2, X_3, X_4)$. The regression procedure forces an ordering of the variables into regression. In Table 1, note that the complete model was obtained by fitting X_4 first, then X_3, then X_2 and finally X_1. In order to eliminate variables at this point, one must determine

---

[1/] This material consists of selected portions of Chapter 6 of Applied Regression Analysis, by Norman R. Draper and Harry Smith, John Wiley and Sons, Inc., New York.

the contribution of each of the variables X_1, X_2, X_3 and X_4 to the regression sum of squares as if each were in the last position. The partial F-test shown in the last column of this printout indicates just this.

2. Using the partial F-test, choose the smallest value and compare it to some critical value of F based on a predetermined α-risk. In this case, the critical F value for, $\alpha = .10$ is $F(1,8,0.90) = 3.46$. The smallest partial F is for variables, X_3; i.e., .0182345. Since the calculated F is smaller than the critical F, reject X_3.

3. Next, find the least-squares equation, $\hat{Y} = f(X_1, X_2, X_4)$. This is shown in Table 2. The overall F value for the equation is 166.83 which is statistically significant. Examining this equation for potential elimination, one sees that X_4 should be eliminated. The procedure for this elimination is similar to the preceding elimination with one change; namely, the critical F value is $F(1,9,0.90) = 3.36$.

4. Find the least-squares equation $\hat{Y} = f(X_1, X_2)$. This is shown in Table 3. This indicates a statistically significant overall equation with an F of 229.50. Both variables X_1 and X_2 are significant regardless of position, as indicated by the significant partial F's. Thus, the backward elimination selection procedure is terminated and yields the equation,

$$\hat{Y} = 52.5773400 + 1.4683057\, X_1 + 0.6622507\, X_2 .$$

Opinion.-This is a very satisfactory procedure in general especially for statisticians who like to see all the variables in the equation once in order "not to miss anything." However, if the input data yields an X'X matrix which is ill-conditioned, i.e., nearly singular, then this procedure may yield nonsense due to rounding errors. With new computing equipment this is not usually a serious problem. We believe this method to be slightly inferior to the stepwise regression procedure below. On the whole, though, it is an excellent procedure.

THE STEPWISE REGRESSION PROCEDURE

The backward elimination method begins with the largest regression, using all variables, and subsequently reduces the number of variables in the equation until a decision is reached on the equation to use. The stepwise procedure attempts to achieve a similar conclusion working from the other direction, i.e. to insert variables in turn until the regression equation is satisfactory. The order of insertion is determined by using the partial correlation coefficient as a measure of the importance of variables not yet in the equation. The basic procedure is as follows. First we select the X most correlated with Y (suppose it is X_1) and find the first order, linear regression equation $\hat{Y} = f(X_1)$. We next find the partial correlation coefficient of X_j ($j \neq 1$) and Y (after allowance for X_1). Mathematically this is equivalent to finding the correlation between (i) the residuals from the regression $\hat{Y} = f(X_1)$ and (ii) the residuals from another regression $\hat{X}_j = f_j(X_1)$ (which we have not actually performed). The X_j with the highest partial correlation coefficient with Y is now selected (suppose this is X_2) and a second regression equation $\hat{Y} = f(X_1, X_2)$ is fitted. The partial F criteria for each variable X_1 and X_2 is now evaluated and compared with a pre-selected percentage point of the appropriate F distribution. This provides a judgment on the contribution made by each variable as though it had been the most recent variable entered. If either variable provides a nonsignificant contribution it is removed from the model. This process continues. After $X_1, X_2, \ldots X_k$ are in the regression, the partial correlation coefficients are the correlations between (i) the residuals from the regression $\hat{Y} = f(X_1, X_2, \ldots, X_k)$ and (ii) the residuals from a regression $\hat{X}_j = f_j(X_1, X_2, \ldots, X_k)$ $(j > k)$. As each variable is entered into the regression, the partial F values for every variable are examined and compared with a preselected percentage point of the appropriate F distribution and any variable which provides a nonsignificant contribution is removed from the model. The process continues until no more variables will be admitted to the equation and no more are rejected. We shall use the Hald data once again, to illustrate the workings of the forward selection procedure. The analysis would proceed as follows (see Table 4):

1. The stepwise procedure starts with the simple correlation matrix and enters into regression that X variable most highly correlated with the response. Here X_4 is entered as in Step No. 1 of Table 4.

2. Using the partial correlation coefficients, it now selects as the next variable to enter regression that X variable whose partial correlation with the response is highest. In this problem it is X_1, with a partial correlation coefficient of 0.91541.

3. Given the regression equation $\hat{Y} = f(X_4, X_1)$ shown in Step No. 2 of Table 4, the

method now examines the contribution X_4 would have made if X_1 had been entered first and X_4 entered second. Because the value of the partial F is 159.295 which is statistically significant, X_4 is retained. So is X_1 because it has a partial F value of 108.224. The stepwise method now selects as the next variable to enter, the one most highly partially correlated with the response (given that variables X_4 and X_1 are already in regression). This is seen to be variable X_2. (The partial correlation coefficient of X_2 with the response is .35833 shown at the bottom of Step No. 2 of Table 4.)

4. A regression equation of form $\hat{Y} = f(X_4, X_1, X_2)$ is now determined by least squares. The variable X_2 enters with a significant partial F value of 5.026. At this point partial F-tests for the variables X_1 and X_4 are made to determine if they should remain in the regression equation. As a consequence, X_4 is rejected since its partial F value 1.863 given in Step No. 3 of Table 4 is not significant compared with the $F(1, 9, 0.95) = 5.12$.

5. The only remaining variable is X_3. Because this variable is immediately rejected, the stepwise regression procedure terminates and chooses as its best regression equation $\hat{Y} = f(X_1, X_2)$ as shown in Step No. 4 of Table 4

$$\hat{Y} = 52.58 + 1.47\, X_1 + 0.66\, X_2 .$$

Opinion.-We believe this to be the best of the variable selection procedures in current use and recommend it. However, "stepwise regression" can easily be abused by the "amateur" statistician. As in all selection procedures, sensible judgment is still required in the initial selection of variables and in the critical examination (through residual anslysis) of the model. It is easy to rely too heavily on the automatic selection performed in the computer.

A discussion of the method is given by M.A. Efroymson (1960). A complete account of the computations needed for the Hald data are given in Draper and Smith (1966).

OTHER PROCEDURES

Other selection procedures in current use include

1. All possible regressions
2. The forward selection procedure
3. The stagewise procedure
4. Variations on the methods above.

For a more complete discussion see Applied Regression Analysis, Chapter 6, by Draper and Smith.

REFERENCES

Draper, N.R., and Smith, H., 1966, Applied regression analysis: John Wiley and Sons, New York, 407 p.

Efroymson, M.A., 1960, Multiple regression analysis, in Mathematical methods for digital computers; Ralston, A., and Wilf, H.S., eds.: John Wiley and Sons, New York, p. 191-203.

Hald, A., 1952, Statistical theory with engineering applications: John Wiley and Sons, New York, 783 p.

Table 1.-$\hat{X}_5 = f(X_1, X_2, X_3, X_4)$

Control Information

No. of observations	13
Response variable is no.	5
Risk level for B conf. interval	5%
Variable entering	1
Sequential F-test	4.3375998
Percent variation explained R-SQ	98.2375700
Standard deviation of residuals	2.4460044
Mean of the response	95.4230750
Std. dev. as % of response mean	2.563%
Degrees of freedom	8
Determinant value	.0010377

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	4	2667.9000000	666.9750000	111.4795200
Residual	8	47.8634980	5.9829372	

B Coefficients and Confidence Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
4	29.9999990	-.1440588	1.4909970 -1.7791144	.7090441	.0412794
3	11.7692300	.1019111	1.8422494 -1.6384272	.7547001	.0182345
2	48.1538450	.5101700	2.1792063 -1.1588665	.7237799	.4968402
1	7.4615383	1.5511043	3.2685233 -.1663147	.7447611	4.3375858

Constant Term in Prediction Equation 62.4051530

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Square of Partials
5	1.00000

Table 2.-$\hat{X}_5 = f(X_1, X_2, X_4)$

Control Information

No. of observations	13
Response variable is no.	5
Risk level for B conf. interval	5%
List of excluded variables	3
Variable entering	4
Sequential F-test	1.8632545
Percent variation explained R-SQ	98.2335600
Standard deviation of residuals	2.3087418
Mean of the response	95.4230750
Std. dev. as % of response mean	2.419%
Degrees of freedom	9
Determinant value	.0500394

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	3	2667.7911000	889.2637000	166.8321800
Residual	9	47.9725980	5.3302886	

B Coefficients and Confidence Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
2	48.1538450	.4161107	.8359611 -.0037398	.1856103	5.0258974
1	7.4615383	1.4519380	1.7165861 1.1872899	.1169974	154.0080400
4	29.9999990	-.2365395	.1554371 -.6285160	.1732876	1.8632548

Constant Term in Prediction Equation 71.6482410

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Square of Partials
3	.00227
5	1.00000

Table 3.-$\hat{X}_5 = f(X_1, X_2)$

Control Information

No. of observations	13
Response variable is no.	5
Risk level for B conf. interval	5%
List of excluded variables	3, 4
Variable entering	1
Sequential F-test	146.5229400
Percent variation explained R-SQ	97.8678500
Standard deviation of residuals	2.4063327
Mean of the response	95.4230750
Std. dev. as % of response mean	2.522%
Degrees of freedom	10
Determinant value	.9477514

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	2	2657.8593000	1328.9296000	229.5042100
Residual	10	57.9043680	5.7904368	

B Coefficients and Confidence Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
2	48.1538450	.6622507	.7644149 .5600865	.0458547	208.5823200
1	7.4615383	1.4683057	1.7385638 1.1980476	.1213008	146.5229400

Constant Term in Prediction Equation 52.5773400

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Square of Partials
3	.16914
4	.17152
5	1.00000

Table 4.-Stepwise solution for the Hald data.

Original and/ or Transformed Data

	X_1	X_2	X_3	X_4	X_5
1	7.00000000	26.00000000	6.00000000	60.00000000	78.50000000
2	1.00000000	29.00000000	15.00000000	52.00000000	74.30000000
3	11.00000000	56.00000000	8.00000000	20.00000000	104.30000000
4	11.00000000	31.00000000	8.00000000	47.00000000	87.60000000
5	7.00000000	52.00000000	6.00000000	33.00000000	95.90000000
6	11.00000000	55.00000000	9.00000000	22.00000000	109.20000000
7	3.00000000	71.00000000	17.00000000	6.00000000	102.70000000
8	1.00000000	31.00000000	22.00000000	44.00000000	72.50000000
9	2.00000000	54.00000000	18.00000000	22.00000000	93.10000000
10	21.00000000	47.00000000	4.00000000	26.00000000	115.90000000
11	1.00000000	40.00000000	23.00000000	34.00000000	83.80000000
12	11.00000000	66.00000000	9.00000000	12.00000000	113.30000000
13	10.00000000	68.00000000	8.00000000	12.00000000	109.40000000

Means of Transformed Variables

1	7.46153830	48.15384500	11.76923000	29.99999900	95.42307500

Std. Deviations of Transformed Variables

1	5.88239440	15.56087900	6.40512590	16.73817800	15.04372400

Correlation Matrix

1	.99999991	.22857948	-.82413372	-.24544512	.73071745
2	.22857948	1.00000010	-.13924238	-.97295516	.81625268
3	-.82413372	-.13924238	.99999991	.02953701	-.53467065
4	-.24544512	-.97295516	.02953701	1.00000010	-.82130513
5	.73071745	.81625268	-.53467065	-.82130513	.99999999

Control Information

No. of observations	13
F level for entering a variable	3.28
F level for deleting a variable	3.28
Response variable is no.	5
Risk level for B conf. interval	5%

Step No. 1

Variable entering	4
Sequential F-test	22.7985280
Percent variation explained R-SQ	67.4542100
Standard error of Y	8.9639014
Mean of the response	95.4230750
Std. error as a % of mean response	9.394%
Degrees of freedom	11
Determinant value	1.0000001

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	1	1831.8968000	1831.8968000	22.7985300
Residual	11	883.8668200	80.3515290	

B Coefficients and Confidence Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
4	29.9999990	-.7381620	-.3978962 -1.0784277	.1545960	22.7985270

Constant Term in Prediction Equation 117.5679300

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Square of Partials
1	.91541
2	.01696
3	.80117
5	1.00000

Table 4.-continued.

Step No. 2

Variable entering	1
Sequential F-test	108.2240500
Percent variation explained R-SQ	97.2471100
Standard error of Y	2.7342642
Mean of the response	95.4230750
Std. error as a % of mean response	2.865%
Degrees of freedom	10
Determinant value	.9397567

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	2	2641.0015000	1320.5007000	176.6272400
Residual	10	74.7620080	7.4762008	

B Coefficients and Confidence Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
4	29.9999990	-.6139538	-.5055738 -.7223338	.0486445	159.2954900
1	7.4615383	1.4399582	1.7483502 1.1315662	.1384165	108.2240500

Constant Term in Prediction Equation 103.0973800

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Square of Partials
2	.35833
3	.32003
5	1.00000

Step No. 3

Variable entering	2
Sequential F-test	5.0258747
Percent variation explained R-SQ	98.2335500
Standard error of Y	2.3087426
Mean of the response	95.4230750
Std. error as a % of mean response	2.419%
Degrees of freedom	9
Determinant value	.0500394

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	3	2667.7908000	889.2636000	166.8320500
Residual	9	47.9726310	5.3302923	

B Coefficients and Confidence Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
4	29.9999990	-.2365401	.1554367 -.6285170	.1732877	1.8632619
1	7.4615383	1.4519379	1.7165861 1.1872897	.1169975	154.0079500
2	48.1538450	.4161100	.8359608 -.0037408	.1856104	5.0258730

Constant Term in Prediction Equation 71.6482910

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Square of Partials
3	.00227
5	1.00000

Table 4.-concluded.

Step No. 4

Variable leaving is	4
Sequential F-test	1.8632611
Percent variation explained R-SQ	97.8678500
Standard error of Y	2.4063325
Mean of the response	95.4230750
Std. error as a % of mean response	2.522%
Degrees of freedom	10
Determinant value	.9477514

ANOVA

Source	d.f.	Sums sqs.	Mean sq.	Overall F
Total	12	2715.7635000		
Regression	2	2657.8593000	1328.9296000	229.5042500
Residual	10	57.9043570	5.7904357	

B Coefficients and Confidence Limits

Var No.	Mean	Decoded B Coefficient	Limits Upper/ Lower	Standard Error	Partial F-test
1	7.4615383	1.4683057	1.7385638 1.1980476	.1213008	146.5229500
2	48.1538450	.6622507	.7644149 .5600864	.0458547	208.5821200

Constant Term in Prediction Equation 52.5773400

Squares of Partial Correlation Coefficients of Variables Not in Regression

Variables	Square of Partials
3	.16914
4	.17152
5	1.00000

Residual Analysis

Obs. No.	Observed Y	Predicted Y	Residual	Normal Deviate
1	78.5000000	80.0739960	-1.5739960	-.6541058
2	74.3000000	73.2509140	1.0490860	.4359688
3	104.3000000	105.8147300	-1.5147300	-.6294766
4	87.6000000	89.2584720	-1.6584720	-.6892115
5	95.9000000	97.2925130	-1.3925130	-.5786869
6	109.2000000	105.1524800	4.0475200	1.6820285
7	102.7000000	104.0020500	-1.3020500	-.5410931
8	72.5000000	74.5754150	-2.0754150	-.8624806
9	93.1000000	91.2754870	1.8245130	.7582132
10	115.9000000	114.5375400	1.3624600	.5661977
11	83.8000000	80.5356710	3.2643290	1.3565577
12	113.3000000	112.4372400	.8627600	.3585373
13	109.4000000	112.2934400	-2.8934400	-1.2024273

FOURIER TREND-SURFACE ANALYSIS IN THE GEOMETRICAL ANALYSIS OF SUBSURFACE FOLDS OF THE MICHIGAN BASIN

by

E.H. Timothy Whitten

Northwestern University

ABSTRACT

The folded top of the subsurface Dundee Limestone has been simulated by double Fourier trend-surface analysis on the basis of 504 well logs. Dips were calculated from tangent planes to this simulated surface and were used to calculate scalar descriptors of the folds. The areal variability of the actual folds can be analyzed by trend-surface analysis of the scalar fold descriptors.

INTRODUCTION

This paper briefly describes a method used successfully to analyze the nature and spatial variability of fold geometry in the Dundee Limestone, central Michigan, on the basis of well-log data.

FOLD DESCRIPTION

While measurement of fold axis orientation and fold size are routine, fold shapes have commonly been described in qualitative terms only. However, Loudon (1964) and Whitten (1966a, 1966b) demonstrated that the nature of fold shape can be quantitatively described by scalar quantities. To describe the geometry of a series of folds quantitatively it is necessary to measure (a) fold axis orientation, (b) fold size, and (c) overall fold shape (Whitten, 1966a, 1966b). Description of fold shape by scalars can automatically specify the orientation of the axial surface and the nature of the fold profile and its variation along the fold axis. Such description permits the complete nature and areal variability of fold geometry to be mapped in a systematic manner, although an actual example has not been analyzed previously.

To illustrate the method, a large number of datum points on the folded surface must be measured. An intensely drilled subsurface area seemed more promising for a preliminary analysis than an area of natural surface outcrop. An area of apparently simple folds (and no known faults) was selected in the Basin oil and gas district of central Michigan (Isabella, Clare, and Necosta Counties: Townships 13-17 North, Ranges 3-7 West). The upper surface of the Middle Devonian Dundee Limestone was cut by 504 drill holes at an approximate mean depth of 2,800 feet within the 900 square mile sample area; the central 324 square mile area was analyzed separately on the basis of 244 drill logs. Actual dip of the stratum at each site could not be determined from the well logs.

DETERMINATION OF FOLD SURFACE AND ITS ATTRIBUTES

Loudon (1964) suggested that the three-dimensional orientation of lines joining all possible pairs of observation points on a folded surface provides an adequate basis for describing the fold geometry. His method has several serious disadvantages that make its use impractical and inadvisable.

The present approach involves use of a trend surface to simulate the folded stratum on the basis of the well-log data. For this purpose the double Fourier series model (James, 1966a) has been used rather than the polynomial trend-surface model (Whitten and others, 1965) because (a) intuitively, it seemed likely that the fold structures are harmonic, and (b) when adequate safeguards are taken, the 'boundary effects' at the periphery of the sampled area appear less severe. Preliminary work with the computer program published by James (1966b) shows that 90 to 95 percent of the total sum of squares is accounted for when some 30 terms of the double Fourier series are used for different portions of the test data. Available tests suggest that the calculated trend surfaces are close approximations to the actual folded bedding plane geometry.

The dip of the simulated bedding plane (Fourier surface) can be calculated at as many points as necessary to yield an adequate sample for analysis of the varying fold geometry. A new computer program in CDC-FORTRAN computes the three-dimensional orientation of the normals to the simulated surface at an operator specified grid of 'sample points'; the computation involves locating the upward-directed normal to the tangent plane (defined by the first derivative of the double Fourier series) at each sample point. The program uses the array of normals to calculate the scalar descriptors necessary to specify all features of the fold geometry (Whitten, 1966a, 1966b), and to prepare equal-area (Schmidt) projections and cross sections through the structure.

By selection of appropriate subareas from the

total region, scalar descriptors for attributes of the fold geometry in each subarea can be calculated and mapped. Trend-surface analysis of these derived data (either by polynomial or Fourier trend-surface analysis) permits the regional changes of fold geometry to be identified and separated from the local deviations.

SIGNIFICANCE OF THE TECHNIQUE

The method apparently provides the first quantitative method of describing the nature and spatial variability of subsurface fold geometry. While this is useful in itself, the availability of scalar descriptors will enable several other significant studies to be undertaken. For example, quantitative assessment of the differences in fold geometry in successive members of the subsurface succession could be analyzed. The method appears to offer a significant exploration tool because regional trends in the changing fold geometry of an area could be identified objectively.

A full analysis of the results obtained from the Dundee Limestone structures, Michigan, and a listing of the computer program will form the subject of a subsequent paper.

REFERENCES

James, W.R., 1966a, The Fourier series model in map analysis: Office of Naval Research, Geography Branch, Tech. Rept. 1, ONR Task No. 388-078, 37 p.

James, W.R., 1966b, FORTRAN IV program using double Fourier series for surface fitting of irregularly spaced data: Kansas Geol. Survey Computer Contr. 5, 19 p.

Loudon, T.V., 1964, Computer analysis of orientation data in structural geology: Office of Naval Research, Geography Branch, Tech. Rept. 13, ONR Task No. 389-135, 129 p.

Whitten, E.H.T., 1966a, Structural geology of folded rocks: Rand McNally and Co., Chicago, 663 p.

Whitten, E.H.T., 1966b, Sequential multivariate regression methods and scalars in the study of fold-geometry variability: Jour. Geology, v. 74, no. 5, pt. 2, p. 744-763.

Whitten, E.H.T., Krumbein, W.C., Waye, I., and Beckman, W.A., Jr., 1965, A surface-fitting program for areally-distributed data from the earth sciences and remote sensing: NASA Contractor Rept., CR-318, 146 p.

THE USE OF EIGENVECTOR METHODS IN DESCRIBING SURFACES

by

T. Victor Loudon

University of Reading (UK)

INTRODUCTION

Any statistical method is, of course, concerned with a set of numerical data. Conventional trend-surface methods are generally applied to a set of values of a mappable variable which describes a surface; either a real surface located in space, or a conceptual surface describing, for instance, the sand/shale ratio of a particular formation. The methods described below, on the other hand, are more applicable to sets of surfaces, where the aim is not so much to study individual measurements of a variable, and their relationship to the surface of which they form a part, but rather to study an entire surface and its relations to the set of surfaces to which it belongs.

Geologists are already familiar with the concept of a set of surfaces, that is with a collection of surfaces that have some property in common. For example, there is the set of all surfaces that show ripple marking, or the set of polished and striated surfaces produced by erosion of rock faces by ice. On a larger scale, there is the set of surfaces that are marked by a dendritic pattern, which might reflect a system of river valleys or submarine canyons, or the set showing the linear pattern of offshore bars or the arcuate ridges of wind-blown dunes. A surface may be assigned to a set on the basis of a number of complex properties, as in deciding, for example, whether a surface in a metasediment belongs to the set of bedding planes or the set of cleavage planes. The sets of surfaces that seem to be most important from a geological point of view are those that indicate the process or environment of formation of the surface.

The fact that sets of surfaces have an important part, implicitly at least, in geological thought, suggests that it may be rewarding to attempt to describe the properties of the sets in quantitative terms. There are strong arguments for making the attempt on economic grounds also. The value of being able to examine the statistical distribution of surfaces belonging to the set describing porosity-feet in Devonian reefs of Alberta and of making comparisons with the set describing similar properties of shoestring sands in the Cretaceous need not be stressed. Perhaps the most urgent need for a quantitative approach, however, arises from the development of methods of simulating geological processes on the computer (Harbaugh, 1966). Frequently the outcome or result of the simulation procedure is in the form of a description of a surface. The aim of simulation is frequently to determine the statistical properties of a number of results which differ because of random variation in the process. The computer is capable of generating large numbers of surfaces, and methods are required for summarizing the results.

The approach to the problem of quantitative description of sets of surfaces that is described here uses correlation methods rather than the regression methods employed by various other workers, for example, Merriam and Sneath (1966) and Sneath (1966). It seems likely that future work will show that the two approaches to the problem are to a large extent equivalent. In the meantime, there may be advantages in looking at the problem from more than one point of view. The actual procedures described below involve well known statistical methods that have been described elsewhere in the geological literature (Whitten, 1966; Scheidegger, 1965; Loudon, 1964), and are not therefore considered in detail here. Application of the techniques described below is believed to break some new ground. Methods are based on procedures for the analysis of orientation data in structural geology, which were developed at Northwestern University. Work on the application of the techniques to sedimentology has formed part of a project undertaken by Professor P. Allen at Reading University.

PROPERTIES OF A SURFACE

The problem of describing a set of surfaces is essentially one of finding for each surface properties that can be described statistically and that are, as far as possible, independent of one another. Statistical measures must be additive, so that measures for the entire set or for subsets can be obtained, and should preferably be easily interpretable in terms of geometrical or geological properties. Orientation, size, and shape of features on a surface are examples of properties that could be measured. The pattern on a surface is frequently repetitive, as in the ripple-marked surface or the dendritic pattern of river valleys mentioned above. It seems desirable that measures describing the shape of features on a ripple-marked bed should not be affected by the relative size of different ripples, nor by their number or completeness. If a set of ripple-marked beds

is described, it should be possible to combine and compare measures derived from different beds. The problem of describing orientations, shapes, sizes and spatial relationships is a geometrical one, and matrix algebra offers a means of representing geometrical operations within the computer.

STATISTICAL DESCRIPTION

In order to separate the concept of shape from that of size and position in space, it is convenient to consider measurements of a surface in terms of orientations rather than locations. In some fields, structural geology for example, the original measurements of a surface are commonly recorded in the form of orientations, such as strike and dip. Otherwise, where the original measurements specify the value of a variable at a number of points, it is always possible to derive orientation measurements from them. Any two points on a surface define a vector, namely the line which joins the points. The orientation of the vector is measured by three direction ratios, which are simply the differences between the x, the y, and the z coordinates of the two points. The length of the vector, by the theorem of Pythagoras, is equal to the root of the sum of squares of the three direction ratios. Direction cosines are the direction ratios of a vector of unit length, and are obtained by dividing each of the direction ratios by the length of the vector. A vector has sense as well as direction, that is, it points one way rather than the other. An arbitrary decision may be necessary about the sense of vectors used to describe a surface, and one may consider that each pair of points defines two vectors, both joining the points, but in opposite senses. One vector is obtained from the other by reversing the sign of each direction ratio. Alternatively, one might choose to consider only those vectors that are directed upwards.

Size

The apparent form of a surface depends on the scale on which it is observed. A bed of sandstone which appears flat and smooth through a pair of binoculars may look very rough under the microscope. Different geological controls operate on different scales, and it may be desirable to try to separate their effects. A whole range of partly independent factors on different scales may control the deposition of a layer of sedimentary rock, for example. On a small scale, the frequency distribution of available grain sizes may determine grain to grain relationships. On a larger scale, properties of the depositing current may cause sedimentary structures to develop. The general environment might control the development and position of tidal channels and offshore bars, while variation in the rate of subsidence might control the large-scale distribution of sediment. In order to separate such features for descriptive purposes, one might consider first only vectors less than a centimeter in length, secondly, vectors between a centimeter and a meter, thirdly, those between a meter and a kilometer, and fourthly, those over one kilometer.

Orientation

In comparing different surfaces, their orientation in space is frequently irrelevant to the comparison. In comparing the topography of river valleys, for example, it is desirable to align them in the same direction before the comparison is made. The alignment should depend on internal properties of the surfaces, not on accidental external features such as their position relative to the North Pole. Suitable internal reference axes are provided by the principal axes of the distribution of vectors. There are three principal axes, three directions at right angles that have the following properties: the correlation between direction cosines measured about the principal axes is zero; the second moment of the vectors measured about the first principal axis is larger than the second moment measured about any other direction in space, while the second moment measured about the third principal axis is a minimum. In geological terms, the principal axes are axes of symmetry, or are as symmetrically placed as is possible in the particular surface. One axis is normal to the plane of the surface, the other two lie on the surface, one in the direction in which the slopes are steepest, the other at right angles to it. In many cases it is likely that a geologist who is shown a map of a surface and asked to align it by eye in directions that reflect the internal properties of the surface, would choose an alignment very close to the principal axes.

Principal axes are computed by matrix algebra (Whitten, 1966) using exactly the same methods as are used in factor analysis. A covariance matrix is calculated from the direction cosines of the vectors. The three eigenvectors of the covariance matrix are the orientation of the principal axes, measured as direction cosines. If, as described above, the vectors were divided into groups according to their length, different groups might prove to have different principal axes. This might occur, for example, where small ripple marks were aligned obliquely on the sides of larger sand bars. When the principal axes have been found, original measurements of location, in the form of coordinates, or of orientation, in the form of direction cosines, can be transformed to refer to the principal axes as coordinate axes by multiplying the data matrix, of which each measurement forms one row, by the eigenvector matrix, which has the direction cosines of the three principal axes as its columns.

Shape

It is usual to compute eigenvalues and eigenvectors at the same time. Eigenvalues are variances of the transformed distribution of direction cosines. They therefore measure the variation in amount of slope in direction parallel to the principal axes. An almost flat surface would have a low variance and an irregular surface a high variance. The variance of the transformed direction cosines is thus a measure of one aspect of the form of a surface. Skewness and kurtosis (Whitten, 1966) can also be computed for the distribution of direction cosines about each principal axis in turn. Skewness measures the degree of asymmetry of slopes on the surface, and indicates whether the steeper slopes face to the right or left along each axis. Kurtosis measures the relative abundance of steep and gentle slopes. A surface in which most of the slopes have approximately the same gradient has a low kurtosis. Surfaces which have a wide range of gradients in a particular direction have a high value for the kurtosis about that axis. Because of the way principal axes are chosen, covariance terms are zero. Higher order correlations, however, and correlations between slope and distance along an axis may be significant in surfaces of complicated form.

Measures of shape derived in this way are independent of size, and if relations between shape and size are of interest, orientation vectors can be divided into groups of different lengths and each group analyzed separately. However, if more exact information is required on the size distribution of features on a surface, some form of series analysis, such as serial correlation or Fourier analysis, can be performed on direction cosines about each principal axis separately.

Change of Scale

It is possible to alter the scale in which length is measured in such a way that the unit of length in a particular direction depends on the variability of slope in that direction. The computational method is to multiply the data matrix by a diagonal matrix in which the eigenvalues are the diagonal elements. The distortion of scale is similar to that used by geologists when the vertical scale of a cross section is exaggerated to show vertical variation more clearly. A map in which measurements were plotted in this distorted scale could be used as a guide to the collection of additional information about the surface. Intuitively, the sampling pattern should be evenly distributed in this space.

Estimation

The descriptive statistics described above can, of course, be combined for a number of surfaces belonging to the same set, weighting the statistics for each surface, if necessary, according to the accuracy with which they are known. Material from different sources can also be brought together in one composite description. For example, information obtained at outcrop about small-scale variation in the thickness of a formation could be stored in the same manner as information about the large-scale distribution pattern of the formation, derived from subsurface information. Data from observation, computer simulation, and knowledge about similar situations elsewhere could be combined to form a composite, quantitative description of the pattern of the thickness distribution of the formation. If therefore, an estimate was required of the thickness at a proposed well site, information from various sources could be used in the prediction.

In the vicinity of the proposed well site, there may be a number of boreholes at which the thickness of the formation is known. Information as described above is available about the expected frequency distribution of slopes on the surface that maps the formation thickness. Vectors can be drawn joining each of the known values to the point at which an estimate is required. From the known value at one end of a vector, and from the frequency distribution of slopes, the probability distribution of values at the other end of the vector can be predicted. Each borehole in turn can be used to compute a probability distribution of values at the proposed well site. Short vectors, from nearby points, are likely to give precise estimates. Long vectors, from distant points may, even though the same distribution of slopes is used, merely indicate that a wide range of values is equally likely, as far as that vector is concerned. A composite probability distribution, taking a number of estimates into account, can be obtained by multiplying together the various estimates of probability for each value. If vectors of different length had been found to behave in different ways, each group could be treated separately, and results combined by multiplication. The calculations are all, of course, performed by computer, and the final result is an estimate of the probability distribution of thicknesses at the site of the proposed well.

Assumptions in the Model

To a large extent, methods described above are purely descriptive, and no constraining assumptions are involved. The interpretation of the results, however, as in the estimation procedures described above, implies a mathematical model which may be an oversimplification of the geological situation. The estimation methods neglect the effect of high-order correlations in the surface, thus implying that the surface has a simpler mathematical form than in fact may be the case. The assumption

is also made that estimates based on different measurements can be considered to be independent, although in general, this may not be true. As the method is still in an experimental stage, great caution is needed in making any interpretation of the results.

SUMMARY AND CONCLUSIONS

The computer's ability to store large quantities of data from many sources and the development of computer techniques that produce numbers of simulated surfaces suggest that methods that can summarize the properties of a set of surfaces are required. One approach is to consider a surface in terms of the orientations of vectors tangential to it. Eigenvector methods can be used to rotate the distribution of vectors to principal axes which provide an internal frame of reference. Conventional descriptive statistics, such as the variance, skewness and kurtosis can be used to describe the distribution of slopes relative to the principal axes and thus measure properties of the shape of the surface. Descriptive statistics from several surfaces can be combined to give a composite picture of the set.

REFERENCES

Harbaugh, J.W., 1966, Mathematical simulation of marine sedimentation with IBM 7090/7094 computers: Kansas Geol. Survey Computer Contr. 1, 52 p.

Loudon, T.V., 1964, Computer analysis of orientation data in structural geology: Office of Naval Research, Geography Branch, Tech. Rept. 13, ONR Task No. 389-135, 129 p.

Merriam, D.F., and Sneath, P.H.A., 1966, Quantitative comparison of contour maps: Jour. Geophysical Res., v. 71, no. 4, p. 1105-1115.

Scheidegger, A.E., 1965, On the statistics of the orientation of bedding planes, grain axes, and similar sedimentological data: U.S. Geol. Survey Prof. Paper 525-C, p. C164-C167.

Sneath, P.H.A., 1966, Estimating concordance between geographical trends: Systematic Zoology, v. 15, no. 3, p. 250-252.

Whitten, E.H.T., 1966, Structural geology of folded rocks: Rand McNally and Company, Chicago, 663 p.

STEPWISE REGRESSION IN TREND ANALYSIS[1/]

by

A.T. Miesch

and

J.J. Connor

U.S. Geological Survey

Trend analysis is an empirical method used in the examination of numerical map data. It consists of fitting mathematical surfaces (models) to the data by means of least-squares techniques in an effort to either (1) obtain a regression equation that can be used to interpolate or predict values between the map control points, or (2) separate components of variation in the map. Selection of a mathematical surface is, in large part, a matter of personal judgment; most work so far has employed either polynomial or Fourier models. According to Krumbein (1966, p. 28), the selection of the trend model depends, to a large extent, on the objectives of the map analysis.

The selection of a trend model for use in prediction or interpolation presents fewer difficulties than selection of a model for use in separating components of variation. In prediction problems a model is needed which will yield very low or zero autocorrelation in the trend residuals (i.e., the deviations of the data from the model should be unpredictable from one map control point to another). The model with the lowest autocorrelation in the trend residuals might, in fact, be accepted as the best predicting device, as long as the model behaves well between data points, so that no more maxima and minima are present in the model than are called for by the data being analyzed.

Use of trend analysis in separating components of map variance is more difficult, and considerably more subjectivity is involved in the selection of a trend model. The variance in any set of map data may be thought of as having resulted from two groups of geologic processes (regional and local) plus some contribution (noise) due to sampling and measurement error. Although such partitioning of the variance seems reasonable, the separation of the conceptual components may be extremely difficult. Nevertheless, this is the objective of a very large amount of numerical map interpretation in geology, whether trend analysis methods are used or not.

The approach offered by trend analysis is to fit a mathematical surface to the data by the method of least squares. This surface is then considered to be an estimate of the regional component of variance (or trend). The deviations of the observed data from the trend describe the local component plus the noise. Where the deviations cluster on the map into areas of positive or negative values (i.e., where they are autocorrelated) they are thought to reflect mostly local variation, or variation on a scale less than the map area but broader than the average distance between map control points. Where the deviations are not so clustered (autocorrelated) they are interpreted to reflect a very local variation (on a scale less than the average distance between control points) and noise due to sampling and measurement errors.

The principal shortcomings of trend analysis are that the correct mathematical form of the actual regional gradient is unknown and that even if it were known its fit to the observed data by least squares would be affected, and therefore, biased by the presence of local components of variation. Because of these shortcomings the methods of trend analysis are essentially empirical. Yet, trend analysis offers a powerful means of examining the data more thoroughly than can be done by visual inspection. Numerous examples of its successful application to a wide diversity of map interpretation problems are present in recent geologic literature (Merriam and Harbaugh, 1964, p. 2).

In a large number of map interpretation problems to which trend analysis has been applied, the trend model accounts for only a small portion of the total variance in the data. In such cases, where the trend is obviously weak, the final geologic conclusions based on the configuration of either the trend or the residual maps are not highly dependent on the mathematical form of the equation used to describe the trend. Two surfaces fitted to a set of map data that are different in mathematical form but about equal in terms of their fit to the data (i.e., sums of squares accounted for) will yield deviations of roughly the same configuration. In other cases, however, where most of the map variation is contained in the trend, residual maps are more highly

---

[1/] Publication authorized by the Director, U.S. Geological Survey.

dependent on the mathematical form of the trend equation. Without a good theoretical basis for deciding which trend equation may be more nearly correct for a given problem it becomes necessary to give equal consideration to a number of equations. These may be of the polynomial or Fourier type, but other equations describing smooth surfaces with gentle flexure should not be excluded. Trend analysis, used in this manner, then becomes an exploratory tool for the examination of map data; various trend models are used and, although mathematical tests of fit may be applied, the value of a model is determined entirely from the geologic information or suggestions it may produce.

Any trend model which can be arranged into the general linear model (Krumbein and Graybill, 1965, p. 283) can be fitted to a set of observed map data by least-squares methods. The general linear model is

$$T_i = b_0 + b_1 w_{i1} + b_2 w_{i2} + \ldots + b_n w_{in} , \qquad (1)$$

where T_i is the trend value at the ith map control point and the w_i's are functions of the map coordinates, X and Y, at the ith point, and constitute the terms of the general model. Terms in X and Y which we have used are listed in Table 1.

Table 1.-Terms used in X and Y.

Polynomial terms	
Linear	X, Y
Quadratic	X^2, XY, Y^2
Cubic	X^3, X^2Y, XY^2, Y^3
Quartic	X^4, X^3Y, X^2Y^2, XY^3, Y^4
Quintic	X^5, X^4Y, X^3Y^2, X^2Y^3, XY^4, Y^5
Other terms	
Root	$\sqrt{X}$, $\sqrt{XY}$, $\sqrt{Y}$
Exponential	e^x, e^y, e^{2x}, e^{x+y}, e^{2y}
Logarithmic	log X, log Y, $(\log X)^2$, log X·log Y, $(\log Y)^2$
Reciprocal	$1/X$, $1/Y$, $1/X^2$, $1/XY$, $1/Y^2$

The terms to be entered into a particular trend equation can be selected by means of stepwise regression. We have used a technique essentially the same as that described by Efroymson (1960), modified so that terms are selected or rejected at a probability level specified by the user rather than using constant critical F values. The modified technique was programmed for a Burroughs 5500 computer by D.S. Handwerker of the U.S. Geological Survey.

In the stepwise regression procedure the standardized partial regression coefficients are estimated as the independent variables to be used in the regression equation are selected, but the method is equivalent to deriving the estimates by

$$B = R^{-1}C , \qquad (2)$$

where B is an array of n standardized regression coefficients used to derive the b's of Equation 1, R^{-1} is the inverse of a matrix of correlation coefficients among the selected w's (of Equation 1), and C is an array of n correlation coefficients between the selected w's of Equation 1 and observed values at the corresponding map control points. If R, the matrix of correlation coefficients among the w's, is singular, R^{-1} in Equation 2 does not exist. Also, if the determinant of R is small the system of equations, in (2), is said to be ill conditioned, and the solutions for B may be highly sensitive to roundoff errors and to small changes which can occur in R with the addition or deletion of even a single map control point. Difficulties attendant with poorly conditioned matrices in trend analysis were considered by Krumbein (1959, p. 828) and Mandelbaum (1963, p. 507). In order to measure matrix condition, we have used the determinant of the normalized matrix, as suggested, for example, by Booth (1957, p. 85) and by Macon (1963, p. 66), and refer to this as the "condition value." The condition value may range from minus one to plus one; the higher the absolute quantity of the condition value, the better condition is the system of equations.

The condition value of an R matrix depends on (1) the distribution of the X and Y control points on the map, and (2) the particular terms in X and Y (Table 1) used in the trend equation.

For example, if the map control points were along some straight line nonparallel to the map coordinate system, and only the linear polynomial terms were selected for the regression equation, the R matrix would consist of

$$R = \begin{bmatrix} r_{x \cdot x} & r_{x \cdot y} \\ r_{y \cdot x} & r_{y \cdot y} \end{bmatrix} = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 1.0 \end{bmatrix}$$

$$\text{or} \quad \begin{bmatrix} 1.0 & -1.0 \\ -1.0 & 1.0 \end{bmatrix} \qquad (3)$$

and would have a condition value of zero, indicating singularity. If one or more control points occurred as outliers from those occurring in a straight line, the $r_{x \cdot y}$ and $r_{y \cdot x}$ elements in (3) would be less than 1.0 in absolute value and the condition value would not equal zero. As sufficient control points are added to form a rectangular grid the R matrix will

converge to

$$\begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix} \quad (4)$$

with a condition value of 1.0.

If the map control points are on a rectangular grid and terms in X and Y that have nonzero linear correlations are used in the trend equation, the condition value of the R matrix is, again, less than 1.0 in absolute value. The specific value, in this case, will depend on what terms are used and the degree to which they are linearly correlated in a particular map problem. In general, the condition value decreases as the number of terms in X and Y is increased. It seems desirable, therefore, to exclude all terms from the trend equation which do not account for a statistically significant proportion of the variance in the dependent variable. This can be accomplished through stepwise regression procedures.

The terms selected by the stepwise regression procedure depend in part on the scale and the origin of the X-Y coordinate system. In one problem we have investigated, for example, the only polynomial terms, linear through quartic, found to be statistically significant were X and X^2; these two terms alone accounted for 95.8 percent of the variance in the dependent variable, whereas the remaining 12 terms accounted for only an additional 1.5 percent. However, when the origin of the coordinate system was moved away from the data points (by adding 10.0 to all X and Y coordinate values) 95.7 percent of the variance was explained by a single term, X^3.

The condition value of the R matrix used to compute the complete quartic surface was 10^{-35}, whereas the R matrix used to fit the X plus X^2 surface had a condition value of 0.025.

Translation and scale changes on the coordinate system may affect the selection of terms in Table 1 by stepwise regression, whether terms are polynomials or not. By changing the coordinate system and repeating the stepwise regression procedure, a number of different trend maps, having varying degrees of fit to the observed data, can be obtained and examined for geologic significance. Where the trends show a high degree of fit to the observed data, the differences among the corresponding residual maps may be large. This circumstance emphasizes the empirical nature of trend analysis and the fact that residual maps, in such cases, cannot be interpreted in a useful way without heavy reliance on other geologic information.

REFERENCES

Booth, A.D., 1957, Numerical methods: 2d ed., Academic Press, Inc., New York, 195 p.

Efroymson, M.A., 1960, Multiple regression analysis, in Mathematical methods for digital computers; Ralston, A., and Wilf, H.S., eds.: John Wiley and Sons, New York, p. 191-203.

Krumbein, W.C., 1959, Trend surface analysis of contour-type maps with irregular control-point spacing: Jour. Geophysical Res., v. 64, no. 7, p. 823-834.

Krumbein, W.C., 1966, A comparison of polynomial and Fourier models in map analysis: Office of Naval Research, Geography Branch, Tech. Rept. 2, ONR Task No. 388-078, 45 p.

Krumbein, W.C., and Graybill, F.A., 1965, An introduction to statistical models in geology: McGraw-Hill Book Co., New York, 475 p.

Macon, N., 1963, Numerical analysis: John Wiley and Sons, New York, 161 p.

Mandelbaum, H., 1963, Statistical and geological implications of trend mapping with nonorthogonal polynomials: Jour. Geophysical Res., v. 68, no. 2, p. 505-519.

Merriam, D.F., and Harbaugh, J.W., 1964, Trend-surface analysis of regional and local components of geologic structure in Kansas: Kansas Geol. Survey Sp. Dist. Publ. 11, 27 p.

APPLICATION OF CANONICAL CORRELATION TO TREND ANALYSIS

by

P.J. Lee

and

G.V. Middleton

McMaster University

INTRODUCTION

Recent years have seen the growth of an intense interest in the application of numerical techniques to the processing of areally distributed data. The most thoroughly studied technique has been trend analysis, which is basically the fitting of a polynomial surface to the observed data, by the method of least squares. Trend analysis appears to be open to a number of practical criticisms:

(i) In common with other univariate or bivariate statistical techniques, applications of trend analysis to geological data rarely reveal information that is not apparent from a close examination of the original data. It has been claimed that the separation of trends from residuals yields information not obtainable from hand- (or machine-) contoured maps, but the few examples presented in the literature are not convincing. Thus far, it appears that conventional trend analysis is useful mainly as a routine data-processing technique, rather than as a research tool.

(ii) Almost all geological data is inherently multivariate: it is rare that the geologist measures only one property at each observation point. Trend analysis, however, is essentially a univariate technique, though it makes use of the apparatus of multivariate statistics to process locational variables.

One approach to the problem of dealing with spatially distributed multivariate data is to process the data first, using a technique such as component or factor analysis, and then use trend analysis to determine trends and residuals for the components or factors. Another technique will be discussed in this paper, namely the application of canonical correlation to trend analysis.

DESCRIPTION OF TECHNIQUE

Canonical correlation is a technique introduced by Hotelling (1936) and hitherto applied mainly in the social sciences, especially psychology (Horst, 1961a, 1961b) and economics. It is basically a technique which seeks to relate two sets of variates to each other, by finding two linear combinations of the variates which maximize the correlation between the two sets. As such, it may be considered to be an extension of multiple correlation or regression, i.e. instead of seeking the b_i which maximizes the multiple correlation between y and the x_i, or setting up the regression equation

$$y = \Sigma b_i x_i + \varepsilon \qquad j = 1,2,\ldots,n$$

where ε is a random variable, we set up the equation

$$U = \Sigma a_i y_i\,, \quad V = \Sigma b_i x_i \qquad i = 1,2,\ldots,m;$$

$$j = 1,2,\ldots,n \qquad (m \geq n, \text{ or } m \leq n) \qquad (1)$$

and maximize the correlation between U and V.

The technique clearly has potential applications in geology, quite apart from trend analysis. For example, the two sets of variates might be chemical and modal rock analysis, or trace and major element analyses of rocks and minerals, or biological and environmental characteristics. The technique may also be used in relation with the discriminant function (Bartlett, 1947).

In order to apply the technique to trend analysis, the spatial coordinates and their various powers and cross-products constitute one set of variates (say, the x_i), and the geological variates constitute the second set. The geological variates are standardized (to zero mean and unit variance) and the covariance matrix is generated for all combinations of the geological and spatial variates, up to the highest order of polynomial surface which is to be investigated. Let this covariance matrix be $\hat{R}$

$$\hat{R} = \begin{bmatrix} \hat{R}_{11} & \hat{R}_{12} \\ \hat{R}_{21} & \hat{R}_{22} \end{bmatrix}$$

where $\hat{R}_{11}$ is the standardized covariance matrix of the geological variates, $\hat{R}_{22}$ is the covariance matrix of spatial variates (including linear, quadratic,

...., up to 6th order polynomial terms) $\hat{R}_{12} = \hat{R}'_{21}$. We require a_i and b_i to be such that U and V have unit variance and a maximum correlation, i. e., maximum E [UV]. These requirements lead to the matrix equation

$$\begin{bmatrix} -\lambda \hat{R}_{11} & \hat{R}_{12} \\ \hat{R}_{21} & -\lambda \hat{R}_{22} \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = 0 \qquad (2)$$

where λ is Lagrange multiplier. In order that there be a nontrivial solution, the matrix on the left of (2) must be singular. Thus, we have

$$(\hat{R}_{22}^{-1} \hat{R}_{21} \hat{R}_{11}^{-1} \hat{R}_{12} - \lambda^2 \hat{I}) = 0 \qquad (3)$$

A Jacobi-like method (Eberlein, 1962) was used to solve for the λ^2's (the eigenvalues). The λ's are called the canonical roots by Hotelling (1936). The largest root is the one of greatest interest. Values of the $\hat{b}_i$ in equation (1) are the eigenvectors associated with λ_i^2, whereas $\hat{a}_i$ may be obtained by

$$\hat{a}_i = \hat{R}_{11}^{-1} \hat{R}_{12} \hat{b}_i / \lambda_i \qquad (\lambda_i \neq 0) \qquad (4)$$

Details of the standard technique are given by Anderson (1958, p. 288-306). In order to obtain a scale dependent trend-surface equation and equal variance for the geological variates, it is necessary to modify the standard technique by the use of the covariance matrix, after prior standardization of the geological variates. It should be noted that the technique selects the surface that has maximum correlation with the weighted geological variates. This surface shows the correct trends, but its slopes and intercept on the vertical axis need to be adjusted before it can be considered as a surface of "best fit" or before meaningful residuals can be calculated.

Interpretation

To understand the possibilities of the technique, consider one possible application. Suppose a geologist wishes to predict areas favorable to oil occurrence on the basis of a number of mapped variables. For simplicity, suppose that only two variates are of interest, formation thickness and percent porosity. Basically the problem is to determine in which area the combination of these two variates reaches a maximum. However, it is clear that predictions can be made only on the basis of the trends, not of the random residuals. Thus in assessing the trends displayed by each variate considered separately, greater weight should be attached to the variate which shows the most clearly defined trend, the variate which can be most reliably predicted.

The application of canonical correlation to this example will result in the determination of that trend which shows (for a given order of polynomial surface) the greatest correlation with a linear combination of the two variates. If the two variates are positively correlated, the weightings will be positive and inversely proportional to the variances of the error terms in their individual trend surfaces. Thus, if there is a strong correlation between spatial position and one of the variates (formation thickness) and only a weak correlation between position and the other variate (porosity), the trend given by canonical correlation will be close to that of the formation thickness alone, and formation thickness will be strongly weighted in the equation for U. This appears to be geologically reasonable because although porosity may be very important for oil accumulation, the assumption has been made in this example that its regional variation is very difficult to predict. Under such circumstances the geologist would naturally tend to outline favorable areas mainly on the basis of variates such as formation thickness, which might be less important as controls of oil accumulation, but had the advantage that they could be reliably predicted.

An illustration of some of the principles of interpretation given above can be given by making use of hypothetical examples. Data were constructed for these examples by using 150 points spaced over a rectangular area, the points being located by a stratified random sampling method. Values for the hypothetical variates at the sampling points were computed from equations incorporating an independent, random, normally distributed, "error" term.

Example 1.-The constructed variates were

$$y_1 = 0.7071\, x_1 + 0.7071\, x_2 + \epsilon_1$$

$$y_2 = 0.7071\, x_1 - 0.7071\, x_2 + \epsilon_2$$

where x_1 is the E-W direction, x_2 is the N-S direction, and ϵ_1 and ϵ_2 are independent random variables with N(0, 1) distribution. The canonical trend was

$$U = 0.67\, y_1 + 0.74\, y_2$$

$$V = 1.0\, x_1$$

The canonical correlation was 0.976.

In this case the strike of the two trends was NW-SE and NE-SW, with both trends increasing towards the east, and with the error terms equal. The canonical trend weights each of the two variates equally, and establishes an "average" trend which strikes N-S and increases towards the east.

Example 2.-The constructed variates were

$$y_1 = 0.7071\, x_1 + 0.7071\, x_2 + \varepsilon_1$$

$$y_2 = -1.0\, x_1 + \varepsilon_2$$

ε_1 and ε_2 were N(0, 1). The canonical trend was

$$U = -0.53\, y_1 + 0.85\, y_2$$

$$V = -0.94\, x_1 - 0.34\, x_2$$

The canonical correlation was 0.985

In this case the variates were negatively correlated, so that the weightings in the canonical trend are of different signs. Because the variance in x_1 was larger than that in x_2 (the area studied was rectangular, with the larger dimension in the E-W direction), the variance of y_2 was larger than that of y_1. After standardization, therefore, the contribution of the variance of the random variable ε_2 to the variance of y_2 was proportionally less than the contribution of ε_1 to y_1. Thus the weighting of y_2 is greater than the weighting of y_1 in the equation for U. If this is taken into account, however, it can be seen that the two variates have approximately equal, though opposite weightings, and that the canonical trend (NNW-SSE) is intermediate between the trends of the two original variates (N-S and NW-SE).

Example 3.-The constructed variates were the same as in example 2 except that ε_2 was N(0, 4). The canonical trend was

$$U = 0.996\, y_1 - 0.093\, y_2$$

$$V = 0.74\, x_1 + 0.68\, x_2$$

The canonical correlation was 0.967.

It is clear that y_1 is so strongly weighted that y_2 has only a very small effect on the trend. The opposite effect was obtained in another experiment in which ε_1 was N(0, 4) and ε_2 was N(0, 1).

If one of the constructed variates has no random variable term, it will receive a unit weight and the other variates will be given zero weight. The canonical correlation is equal to one. If more than one of the constructed variates has no random term, the computer program selects one of the variates, more or less at random, and assigns unit weight to that variate and zero weight to the others.

The following is a summary of the principles of interpretation which emerge from these experiments:

(1) Positively correlated variates have weightings of similar sign in the equation for U. Large weightings are assigned to those variates which show least deviation from the specified order of polynomial surface.

(2) Negatively correlated variates have weightings of opposite sign.

(3) The canonical correlation cannot be less than the largest multiple correlation between a single variable and the specified order of polynomial surface, and may be much larger. U is the most predictable linear combination of the variates and the equation for V specifies its trend surface (subject to intercept and slope corrections).

CONCLUSIONS

This paper should be considered to be a report on work in progress. Further studies are necessary, both on constructed and real geological examples, in order to gain experience of the potentialities and limitations of the method. It appears, however, that canonical correlation may be used to extend the application of trend analysis to multivariate data, and that the trends which result may be expected to have both geological meaning and practical application.

REFERENCES

Anderson, T.W., 1958, An introduction to multivariate statistical analysis: John Wiley and Sons, Inc., New York, 374 p.

Bartlett, S.M., 1947, Multivariate analysis: Jour. Roy. Stat. Soc. Supple., v. 9, p. 176-197.

Eberlein, P.J., 1962, A Jacobi-like method for the automatic computation of eigenvalues and eigenvectors of an arbitrary matrix: Jour. Soc. Industrial and App. Math., v. 10, p. 74-88.

Horst, P., 1961a, Generalized canonical correlations and their applications to experimental data: Jour. Clinical Psychology, no. 14, p. 331-347.

Horst, P., 1961b, Relations among m sets of measures: Psychometrika, v, 26, p. 129-149.

Hotelling, H., 1936, Relations between two sets of variates: Biometrika, v. 28, p. 321-377.

Hotelling, H., 1957, The relations of the newer multivariate statistical methods to factor analysis: British Jour. Stat. Psychology, v. 10, pt. 2, p. 69-79.

A SIMULATION OF GHOST STRATIGRAPHY

by

George S. Koch
U.S. Bureau of Mines

and

Richard F. Link
Princeton University

ABSTRACT

A quadratic equation is fitted to illustrative data simulating ghost stratigraphy. The patterns found among the residual variations indicate the nature of the rock before metamorphism. Thus, the complications of a specific model are shown, and an instructive use of simulation is illustrated.

A familiar geological problem in deciphering the history of metamorphic rocks is to identify the nature of the rock before metamorphism. The same problem exists for rocks changed by processes not always considered to be metamorphic, such as diagenesis, weathering, and hydrothermal alteration. In studying such rocks, it is sometimes possible to identify statistically numerical information that has not been oblitereated by the metamorphism, just as a mineral pseudomorph can be recognized even though the mineralogy has been changed. Thus, Whitten (1959, 1960) has recognized "ghost stratigraphy," stating that, in the Donegal area, Ireland "the deviations (residuals) have geological significance and may be closely correlated with the metasediments extant prior to the emplacement of the granitoid rocks."

We shall look for indications of premetamorphic phenomena by studying residual variations to a trend surface fitted to simulated illustrative data. When a quadratic or other mathematical equation is fitted to data, the resulting surface seldom, if ever, corresponds exactly to the actual observations at the sample points. This lack of agreement results from variation within sample points, if there is more than one observation at each sample point, and from the fact that the fit to the summary-sample points is not perfect. Rather, there is a residual variation between the surface and the observations, whether original or summarized, measured by the vertical distance at each point between the elevation of the point and the elevation of the fitted surface, both with reference to the datum plane. This residual variation, named simply the residual, is positive if the actual value is above the fitted surface and is negative if the actual value is below.

Residuals are studied for several reasons. The principal one is that if a generalized trend, say a quadratic trend, is removed, the basic, underlying behavior of the dependent variable may be recognized. Geologists are already accustomed to generalized linear trends. In structural petrology, data may be plotted on a stereographic net and then rotated to make some element, such as a hypothetical "b" axis or fold axis horizontal, regional dip may be removed for a problem in petroleum geology, or, details of vein structure may be plotted. In each of these examples, linear trend is removed. The same method can be applied mathematically, with the advantage that quadratic or higher order trend can also be removed.

In our simulated illustration we will assume a square surface area, as outlined in Figure 1. The area, of arbitrary size, is subdivided into four equal parts, which may be imagined to represent outcropping sedimentary beds striking due north. The mean value of an imaginary constituent is different in the four beds, as shown in the figure (20 in bed 1, for example). The area is sampled at 256 points on a 16 by 16 grid. However, the observations in each bed are not equal to the mean, because two additional elements are introduced. First, to simulate natural fluctuation, a random number from a normal distribution with a mean of 0 and a standard deviation of 2 is added to each of the original mean values. Second, to simulate a regional metamorphic front or other phenomenon, a northeast linear trend has been added. This trend is numerically equal to 0 at the southwest corner of the map area and increases linearly to 10 on the northwest-southeast diagonal and to 20 at the northeast corner. The resulting observations are plotted in Figure 1.

If we now pretend that the construction of the model is unknown and that only the observations at the sample points are available, we can test whether the 256 observations provide evidence of

northward-trending ghost stratigraphy in the simulated metamorphic area. When a quadratic equation is fitted to the data, the surface of Figure 2 and the analysis of variance of Table 1 are obtained. As the element of linear northeast gradient is strong, most of the regression is explained by the linear terms, and strike of 141 degrees (N 39 W) is estimated. The plotted surface shows the general increase in gradient to the northeast, complicated by a quadratic trend introduced because of the initially different means in the four beds.

The next step is to determine the pattern of residuals from the quadratic trend. When the residuals are calculated and plotted, the pattern shown in Figure 3 and Table 2 is obtained. For simplicity on this figure, the residuals are plotted as plus or minus signs, with blanks indicating those from plus 1 to minus 1. The residuals indeed define a pattern, for example, in bed 1 there are 37 positive residuals and only 20 negative ones. To carry the analysis further, we could have compared the residuals among beds by one of several F- or t-tests depending on the details of our hypotheses, and then could have confirmed the result, already evident from the counting of positive and negative residuals.

In summary, this simulaton of ghost stratigraphy shows the sort of assumptions that can be made, and the complications of a specific model. The presentation also gives an idea of the use of simulation from which some insight can conveniently be gained about how real data might behave under certain model conditions. While preparing this example, we simulated several sets of ghost stratigraphy varying the parameters, and found it instructive to see how, as the trend, the initial bed differentiation, and the amount of randomness changed, the"ghost" got fainter and fainter and then disappeared.

REFERENCES

Whitten, E.H.T., 1959, Composition trends in a granite: Modal variation and ghost-stratigraphy in part of the Donegal granite, Eire: Jour. Geophysical Res., v. 64, p. 835-848.

Whitten, E.H.T., 1960, Quantitative evidence of palimpsestic ghost-stratigraphy from modal analysis of a granitic complex: XXI Internat. Geol. Congr., Norden, pt. 14, p. 182-193.

Table 1.-Analysis of variance for quadratic regression analysis of simulated ghost stratigraphy data.

Source of variation	Sum of squares	Degrees of freedom	Mean square	F	F 10%
Linear terms	4,698.23	2	2349.11	119.0	2.30
Quadratic terms	2,533.90	3	844.63	42.8	2.08
Residual	4,935.71	250	19.74		
Total	12,167.84				

Table 2.-Distribution of positive, zero, and negative residuals from fitted quadratic surface.

Bed	Number +	Number 0	Number -
1	37	6	20
2	11	8	45
3	28	17	19
4	27	10	27

28.6	34.0	30.2	29.7	24.7	16.1	24.3	30.2	28.6	28.4	27.4	32.9	36.1	35.5	42.5	36.0
27.6	33.0	31.2	32.6	19.3	23.1	24.1	24.3	28.4	29.0	30.4	26.6	27.9	29.7	39.9	33.7
26.5	25.6	28.2	31.2	17.0	18.9	24.7	19.7	33.3	27.2	36.2	34.8	31.6	34.3	40.1	37.1
25.9	29.3	22.8	29.8	27.2	15.0	28.1	25.1	29.1	22.1	30.4	24.2	37.8	32.0	37.9	32.1
30.1	24.7	28.5	25.2	20.2	22.8	19.8	20.9	23.7	24.7	22.9	22.0	26.8	43.0	33.6	36.7
21.1	31.3	27.9	28.1	16.0	26.6	18.8	19.8	21.5	32.5	22.3	32.1	38.6	34.8	31.8	40.4
36.8	21.9	29.3	25.1	19.7	17.6	23.4	20.8	21.2	24.3	32.1	24.7	33.5	33.0	33.6	39.0
23.4	21.6	22.3	24.5	17.5	16.9	16.8	20.2	21.8	26.0	28.7	32.1	33.3	28.3	30.2	34.8
23.6	19.4	25.2	24.7	22.1	15.5	20.1	33.2	22.8	29.8	25.2	24.3	35.9	34.1	29.5	33.0
27.4	23.2	21.0	26.8	16.7	12.1	19.9	26.9	25.0	25.6	23.4	25.6	32.1	26.3	32.1	33.9
22.7	23.4	22.0	29.4	12.0	18.3	18.1	10.7	22.3	24.2	23.2	20.6	26.2	36.1	33.1	30.5
26.5	27.9	30.6	22.8	19.0	16.0	13.1	20.9	24.5	21.5	26.2	23.6	26.4	34.2	33.7	26.3
25.9	21.7	17.5	26.6	6.0	9.8	14.0	13.5	23.5	20.5	24.7	21.8	26.2	21.6	31.8	30.1
24.5	27.1	18.5	22.7	12.6	14.0	16.2	22.0	24.5	22.3	19.7	33.5	32.4	43.4	32.4	35.8
21.8	21.3	15.5	22.5	17.5	12.6	14.8	13.8	18.2	15.7	19.9	22.1	28.1	32.0	25.0	34.0
18.4	17.4	22.1	21.9	12.1	17.9	18.2	6.4	14.8	27.4	23.7	20.7	21.9	32.5	28.8	27.8
Bed 1: $\mu_1 = 20$				Bed 2: $\mu_2 = 10$				Bed 3: $\mu_3 = 15$				Bed 4: $\mu_4 = 20$			

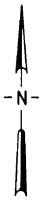


Figure 1.- Model for simulated ghost stratigraphy, showing simulated observations, and means for four beds.

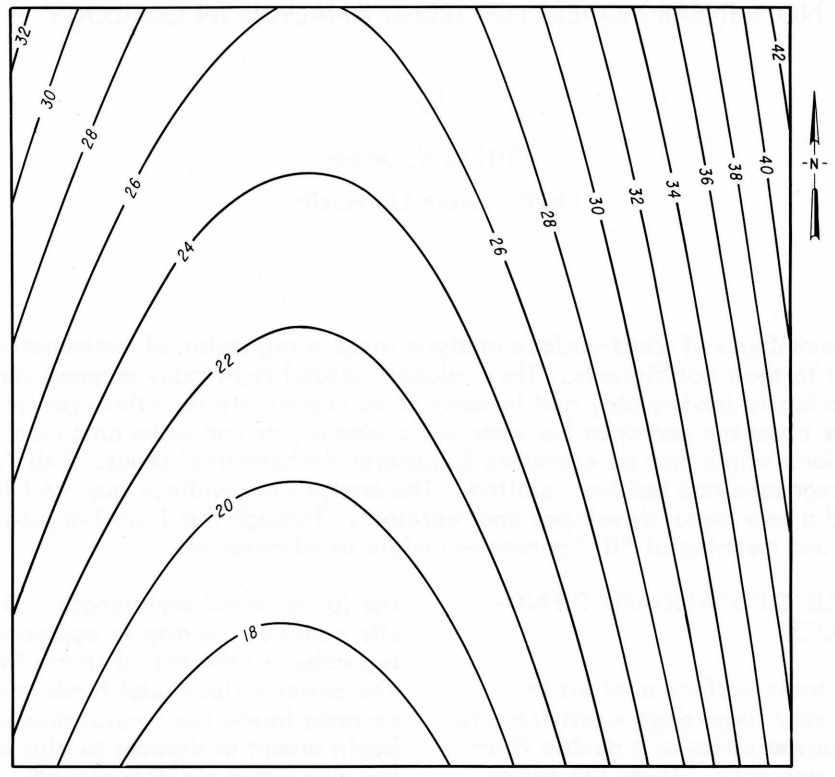


Figure 2.- Quadratic regression surface fitted to simulated ghost stratigraphy data.

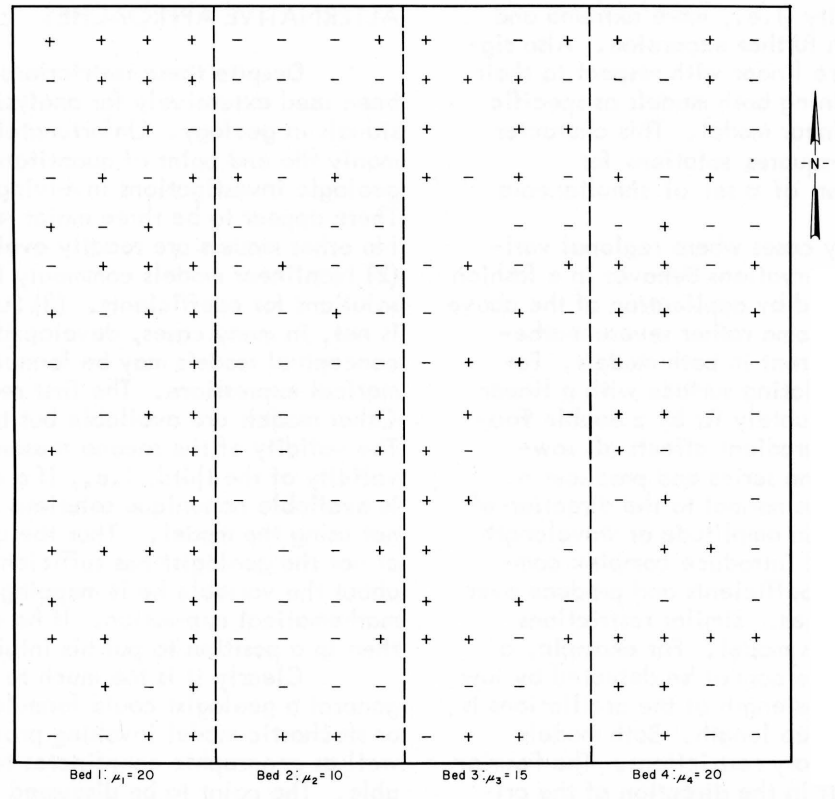


Figure 3.- Positive and negative residuals from quadratic regression surface.

NONLINEAR MODELS FOR TREND ANALYSIS IN GEOLOGY

by

William R. James

Northwestern University

ABSTRACT

The standard techniques of trend-surface analysis involve regression of mathematical functions which are linear with respect to their coefficients. The regional variability of many mapped variables is such that standard models are basically inadequate, and in many cases complicate the otherwise simple regional picture. In many of these cases the geologist has some basic idea of, or can make an educated guess at, some underlying geometric form which may be expressed in general mathematical terms. This "substantive" geometric model may be regressed and residuals plotted. The analysis of residuals may well lead to a revision of the original model and a new model developed and regressed. Through this iterative procedure a satisfactory "trend" may be found and meaningful "fit" parameters might be discovered.

REVIEW AND CRITIQUE OF STANDARD TREND-ANALYSIS TECHNIQUES

The method of trend-surface analysis in geology has been to a very large degree restricted to regression of either a power series or a double Fourier series to observed map data. These two series (or models) have much in common, including some distinct disadvantages. They both represent finite expansions of infinite series. They both possess terms of increasing complexity (i.e., more extrema and inflection points) upon further expansion. Also significant is that both are linear with respect to their coefficients, thus defining both models as specific cases of the general linear model. This character enables unique least-squares solutions for coefficients by solution of a set of simultaneous linear equations.

There are many cases where regional variability of mappable observations behaves in a fashion which cannot be deduced by application of the above models. This is due to some rather severe mathematical restrictions inherent in both models. For example, a simple oscillating surface with a linear gradient cannot be adequately fit by a double Fourier series. The linear gradient affects all row-column coefficients of the series and produces a wavy surface more or less normal to the direction of dip. A regular change in amplitude or wavelength in a given direction will introduce complex components to the Fourier coefficients and produce over complicated trend surfaces. Similar restrictions apply to the power series model. For example, a simple oscillating surface cannot be detected by low degree trends if the wavelength of the oscillations is much smaller than the map length. Both models suffer from severe boundary restrictions. The Fourier series must repeat itself in the direction of the orientation of the map grid over every interval equal to the fundamental wavelength. This forces the opposite edges of the map to approach the same profile regardless of the actual trends in the observations. The power series model tends to use all available extrema inside the control area and is forced to begin ascent or descent to plus or minus infinity as the map edges are approached. These aspects of the two models are discussed and illustrated by Krumbein (1966).

ALTERNATIVE APPROACHES

Despite these restrictions both models have been used extensively for analysis of trend and residuals in geology. Unfortunately this is also commonly the end point of quantitative procedures in geologic investigations involving map analysis. There appear to be three major reasons for this. (1) No other models are readily available to geologists. (2) Nonlinear models commonly involve nonunique solutions for coefficients. (3) Substantive reasoning is not, in many cases, developed to the point where conceptual models may be formulated into mathematical expressions. The first reason is not valid. Other models are available but they are nonlinear. The validity of the second reason depends upon the validity of the third, i.e., if a deterministic model is available nonunique solutions are not reasons for not using the model. Thus the question is whether or not the geologist has sufficiently clear intuition about the variable he is mapping to formulate a mathematical expression. If he can do this, he is then in a position to put his intuition to a test.

Clearly it is too much to expect that in general a geologist could formulate a deterministic or stochastic model invoking process elements as well as geographic coordinates to apply to his variable. The point to be discussed in this paper is that there is an alternative. The geologist may not

understand the complex interplay of process elements which produced the geometry he observes when he maps a variable. Yet he is nearly always capable of generalizing or simplifying that geometry in his own mind. He is also capable of assessing certain boundary conditions to that geometry. For example, a geologist studying sand texture parameters in a nearshore and beach environment is aware of the likelihood of a discontinuity occurring in his map variable along the plunge zone. A fault would produce similar effects on variables measured on a structurally deformed terrain. The development of a simplified conceptual model for the observed geometry is a substantive reasoning process. It does not, however, involve the development of a deterministic model. The substantive geometric model lies somewhere between the arbitrary linear map models and the deterministic model.

It appears that in many cases the geologist is in a position to at least attempt to apply a substantive rather than an arbitrary geometric model. One example is the interpretation of subsurface structure from boreholes. One could conceive of a sloping corrugated surface. This surface might well be approximated by a simple sine wave on a plane. This concept alone is enough to allow development of a mathematical expression for regression. One might develop this expression so that average depth or elevation with respect to some datum, trend and plunge of fold axis, dip of the (ab) plane, and wavelength and amplitude of fold profile, are the parameters of the fitted surface, hereafter called simply the fit parameters. This expression may be regressed by standard iterative techniques until the "best reasonable" fit is found. Residuals may then be computed and plotted, and the trend contribution to the total sum of squares calculated. These may be used as criteria for a substantive judgment on the value of the initial model.

Other examples of the potential use of the substantive geometric model are abundant. For example there has been a good deal of discussion in the literature concerning the formation of beach cusps. One would not find it difficult to make accurate measurements of the geometry of beach cusps and formulate a general equation for regression. When a satisfactory model is found, the fit parameters could be computed for a variety of beach cusps under varying sea conditions. The fit parameters are likely to be meaningful numbers to use in correlation of shore processes and cusp geometry. O.F. Evans (1940) measured several profiles across the "low and ball" structures off the east coast of Lake Michigan. His profiles seem to indicate an oscillating surface with increasing wavelength with distance offshore. These structures are parallel to the shoreline and continuous for many miles. In the study of the origin of these structures one would desire meaningful parameters describing the geometry of these features. These numbers are not likely to be obtained through application of an arbitrary regression function. They may well be obtained from an equation developed by substantive reasoning on the part of the geologist.

In short, the use of arbitrary functions within the framework of the general linear model should not, in many cases, be the end point in quantitative analysis of trend.

Time and space limitations of this report preclude the working out of specific examples to support the foregoing general statements. However, one example is discussed below which displays in some detail, the various procedures which might be used.

EXAMPLE

One of the simpler examples mentioned above is the underlying geometry associated with the "low and ball" structures. The term "low and ball" refers to a parallel set of offshore bars and troughs commonly developed along relatively stable shorelines. O.F. Evans (1940) measured a series of profiles normal to the eastern shoreline of Lake Michigan. Figure 1 shows his sounding profiles. Although Evans did not show the locations of his profiles one could conceive of them as being spaced side by side and thus representing sample observations of a surface.

The geometry appears to be quite regular. As a first approximation one might postulate a simple sine wave form having a linear increase in wavelength with increasing distance offshore, constant amplitude, and origin at the shoreline. The oscillations could be allowed to be about a plane dipping normal to the shoreline with an intercept at the shoreline. The general equation for such a surface is:

$$X_c = a_1 V + a_2 \sin \frac{2\pi V}{a_3 + a_4 V}$$

where

- X_c = computed value of depth below water level
- a_1 = average offshore slope
- a_2 = amplitude of low and ball structures
- a_3 = initial wavelength of low and ball structures
- a_4 = rate of change of wavelength with increasing distance from shoreline
- V = distance from shoreline

This equation may then be fitted to the data using the least-squares method. The least-squares function to be minimized may be written as follows:

$$G = \Sigma (X_i - X_c)^2 = \Sigma \left(X_i - a_1 V_i - a_2 \sin \frac{2\pi V_i}{a_3 + a_4 V_i}\right)^2$$

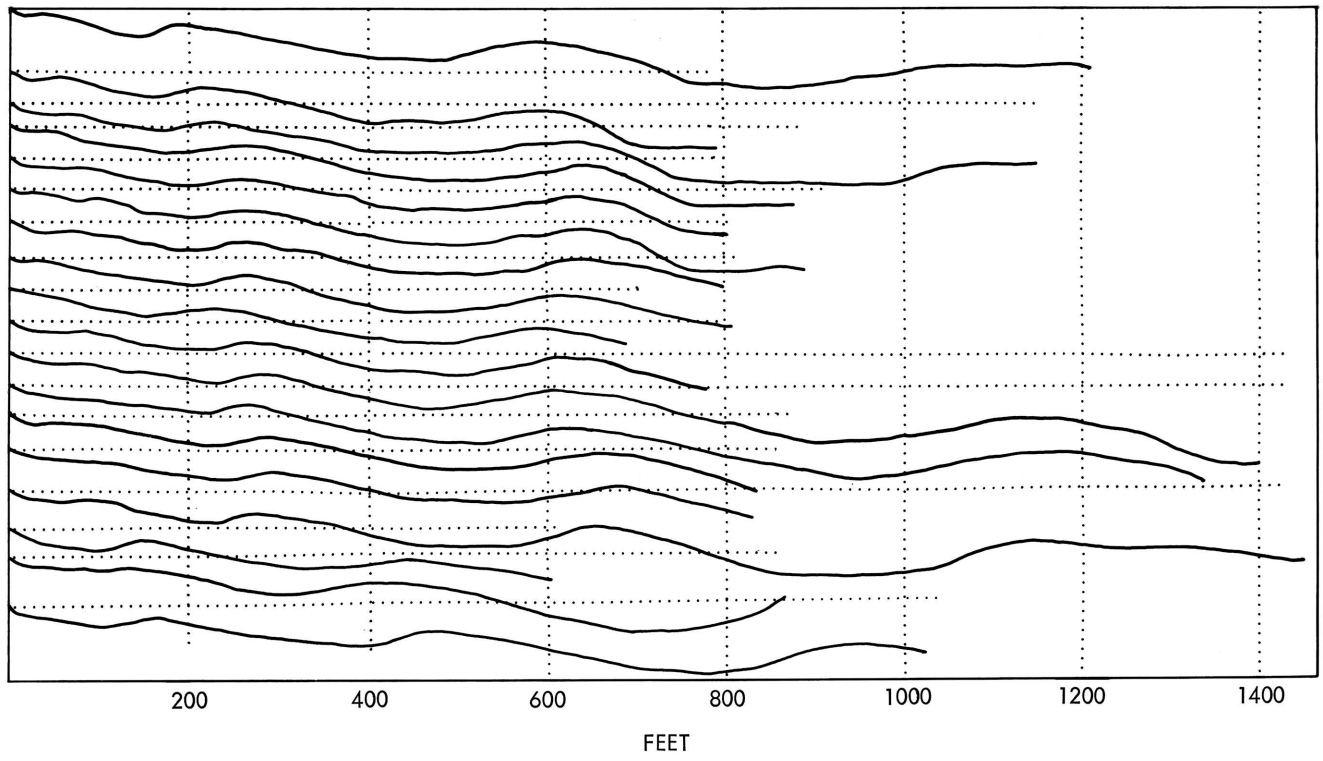


Figure 1.- Low and ball profiles normal to eastern shore of Lake Michigan, after Evans (1940). Essential part of figure is shape of profiles, not affected by loss of clarity of small type.

where

X_i = observed depth at distance V_i

The first partial derivatives with respect to each coefficient may be written as follows:

$$\frac{\partial G}{\partial a_1} = -2 \sum_i V_i \left(X_i - a_1 V_i - a_2 \sin \frac{2\pi V_i}{a_3 + a_4 V_i}\right)$$

$$\frac{\partial G}{\partial a_2} = -2 \sum_i \sin \frac{2\pi V_i}{a_3 + a_4 V_i} \times \left(X_i - a_1 V_i - a_2 \sin \frac{2\pi V_i}{a_3 + a_4 V_i}\right)$$

$$\frac{\partial G}{\partial a_3} = 4\pi a_2 \sum_i \left(\frac{V_i}{(a_3 + a_4 V_i)^2} \cos \frac{2\pi V_i}{a_3 + a_4 V_i}\right) \times \left(X_i - a_1 V_i - a_2 \sin \frac{2\pi V_i}{a_3 + a_4 V_i}\right)$$

$$\frac{\partial G}{\partial a_4} = 4\pi a_2 \sum_i \left(\frac{V_i^2}{(a_3 + a_4 V_i)^2} \cos \frac{2\pi V_i}{a_3 + a_4 V_i}\right) \times \left(X_i - a_1 V_i - a_2 \sin \frac{2\pi V_i}{a_3 + a_4 V_i}\right)$$

In the case of the general linear model, setting the first partial derivatives equal to zero generates a set of simultaneous linear equations. For nonlinear models this is not true. Thus other methods must be used to solve for the coefficients. One of the faster methods is the method of steepest descent. The procedures are as follows. (For general reference to this technique, see Hadley, 1964.)

1. An initial guess is made for the values of the coefficients.

2. The least-squares function and its set of

partial derivatives are computed for that initial guess.

3. The vector which points in the direction of steepest descent is given by the direction numbers $(-\frac{\partial G}{\partial a_1}, -\frac{\partial G}{\partial a_2}, -\frac{\partial G}{\partial a_3}, -\frac{\partial G}{\partial a_4})$. New estimates of the coefficients are made by moving in the direction of steepest descent.

$$a_{1_i} = a_{1_{i-1}} - \Delta \frac{\partial G}{\partial a_{1_{i-1}}}$$

$$a_{2_i} = a_{2_{i-1}} - \Delta \frac{\partial G}{\partial a_{2_{i-1}}}$$

$$a_{3_i} = a_{3_{i-1}} - \Delta \frac{\partial G}{\partial a_{3_{i-1}}}$$

$$a_{4_i} = a_{4_{i-1}} - \Delta \frac{\partial G}{\partial a_{4_{i-1}}}$$

where

Δ = a multiplier which can be chosen before the computer run or programmed to depend on the rate of descent of the function.

4. The new estimates of the coefficients are then used to recompute the function and its first partial derivatives. The same procedure is followed iteratively until the function converges at its minimum.

The initial guess is fairly important in that the least-squares function for this regression will have several minima. One must be certain that the correct one is approached. In the example being discussed here initial guesses are not difficult to make. The general slope of the nearshore bottom appears to be in the neighborhood of 12 ft. of depth per 700 feet offshore distance. Thus an initial guess of 0.017 would be a reasonable estimate for a_1. The amplitude of the low and ball structures appear to be in the range of two to four feet. Because the variable being measured is depth below lake level and the wave form initially dips, the sign of the amplitude will be positive. Thus a value of +3 would be a reasonable guess for the initial value of a_2. The initial wavelength (a_3 and a_4) may be estimated by a few simple measurements and calculations. Let $\phi = 2\pi V/(a_3 + a_4)$. It is evident from Figure 1 that ϕ must get as high as 7.5π in order to produce the four bars observed. Setting $\phi = 7.5\pi$ will provide a maximum estimate for a_4. When this is done it is seen that $3.75a_3 = V(1-3.75a_4)$. Thus a_4 must be less than 0.267 or the fourth bar would not be allowed to appear in the function. The second bar has its peak at around V = 250 ft. The fourth bar has its peak at around 1150 ft. These estimates may be used to set up two equations in two unknowns.

$$3.5\pi = 2\pi \times 250/(a_3 + a_4 \times 250)$$

$$5.5\pi = 2\pi \times 1150/(a_3 + a_4 \times 1150)$$

These equations give an estimate of $a_3 \approx 97$ ft. and estimate $a_4 \approx 0.182$. These estimates predict other peaks as shown below.

peak	distance offshore
1	84 feet
2	250 feet
3	533 feet
4	1150 feet
5	3403 feet
6	never reached (ϕ converges on a constant)

These appear to be close enough to accept the estimates of a_3 and a_4 as satisfactory first guesses. It is interesting to note that the final estimates of a_3 and a_4 will predict a specific number (perhaps zero) of bars farther offshore. If the final fit is judged to be good, this equation may be successful in prediction of offshore bars by measurement of the relative spacing of a few nearshore bars. It is quite evident that the use of arbitrary linear mapping models will not be successful in this respect.

After the fit has been made and residuals plotted, regular deviations might be observed that would suggest modification of the original model. Several such modifications are possible. Amplitude could be allowed to vary with distance offshore. Wavelength could be given a different functional dependence on distance offshore. A linear trend of bar crests not parallel to shore could be allowed. The variety is infinite. By building successively onto simple models a satisfactory model is likely to be discovered.

CONCLUSIONS

Dr. J.W. Tukey of Princeton University has long been the major proponent of the iterative "fit and expose" approach to data analysis. It is thus appropriate to close with references to his ideas on this subject. With regard to the present state of the art of data analysis he states (Tukey, 1965), "Inadequate attention to the objectives and science of data analysis has left us hampered by fragmentary understanding. The technology of data analysis is today far more awkward and inadequate than is necessary. Recognition and use of modular components of diverse forms is now essential ...."

Referring to the objectivity which most people associate with arbitrary statistical models Tukey states, "Data analysis cannot do more than bring to our attention a combination of the content of the data with knowledge and insight about its background. Validity and objectivity in data analysis are dangerous myths." Speaking on the strategy of data analysis Tukey argues, "The twin objectives of data analysis, summarizing and exposing, go hand in hand. Once something has been summarized, we gain by exposing only what has not been described. From such exposition we hope to learn how to summarize more extensively or more precisely in the next cycle. The single process of fitting and forming residuals typifies those important processes that serve both summarizing and exposing at the same time. Fitting can be effective when what is fitted is neither close nor believed in. There is no substitute for looking at residuals, and doing this in more than one way."

Acknowledgments.- I am indebted to Dr. H.C. Helgeson of the Geology Department at Northwestern University for introducing me to nonlinear programming techniques, and to Dr. W. C. Krumbein for providing encouragement and aid in the development of the example in the text.

REFERENCES

Evans, O.F., 1940, The low and ball of the eastern shore of Lake Michigan: Jour. Geology, v. 48, p. 476-511.

Hadley, G., 1964, Nonlinear and dynamic programming: Addison-Wesley Publ., Inc., Reading, Massachusetts, 477 p.

Krumbein, W.C., 1966, A comparison of polynomial and Fourier models in map analysis: Office of Naval Research, Geography Branch, Tech. Report No. 2, ONR Task No. 388-078, 45 p.

Tukey, J.W., and Wilks, M.B., 1965, Souvenir sheet for data analysis and statistics; principles and practice: Talk delivered at the Symposium on Information Processing in Sight Sensory Systems, at California Institute of Technology on November 1.

CORRELATION BETWEEN SURFACES BY SPECTRAL METHODS

by

John N. Rayner

Ohio State University

INTRODUCTION

A large number of disciplines are interested in the two-dimensional spatial pattern of variables. In particular, biology, geography, and geology are frequently concerned with the variation of phenomena at or close to the surface of the earth. In any science one of the initial steps in analysis is the description of the phenomena, and in these areas of spatial study this step is in part fulfilled by the fitting of a theoretical surface. It serves to generalize objectively some of the features of the empirical distribution. However, the fitted surface is only a tool used for describing one individual distribution, and further analysis virtually always calls for the study of joint distributions, the analysis of the associations between phenomena in two-dimensional space. Just as the variance measures the variability in the single variate, the covariance measures the joint variability in two. Suitably normalized by the square roots of the variances of the separate data sets, the covariance becomes the correlation coefficient. It is to the problem of estimating the covariance (and the correlation) between two surfaces that this paper is addressed.

ONE-DIMENSIONAL CORRELATION

Before consideration is given to the two-dimensional case it is instructive to look at the one-dimensional situation. Figure 1a shows the variation of two series along a line. For example, these might be plots of height and density of vegetation along a traverse line. The correlation coefficient estimated in the normal way,

$$r_{xy} = \frac{\mathrm{cov}(xy)}{(\mathrm{var}(x)\mathrm{var}(y))^{1/2}} = \frac{\Sigma xy/n}{(\Sigma x^2/n\ \Sigma y^2/n)^{1/2}}, \quad (1)$$

where x and y are deviations from their mean value, is calculated as 0.5. This is not particularly intriguing. The relationship is positive as a scatter diagram would suggest (Fig. 1b), but it is by no means definite.

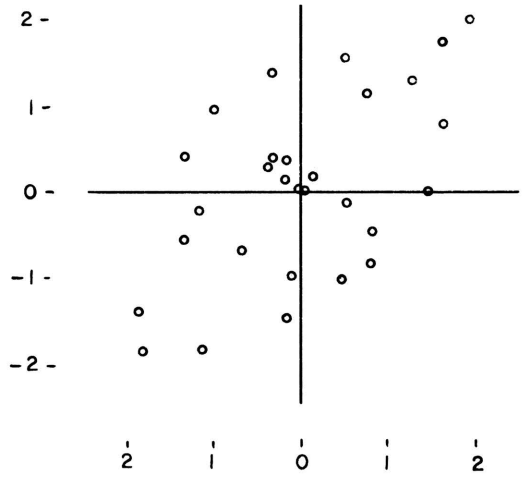


Figure 1b.-Scatter diagram of equally spaced data in Figure 1a. Height is plotted along the ordinate and density along the abscissa.

Various courses of action now lie open to the researcher. If he feels he has a well based hypothesis that there should be a linear relationship between the height and density, then he will look for methods of improving the testing procedure. Obviously, one suggestion he will make is that other factors are working independently on height and density. He might measure these and add to the regression equation. Alternatively, he might hypothesize that the independent factors operate in different areas and that their effect will show up at different intervals (wavelengths) along the traverse. If he takes this point of view he will proceed by isolating the frequency

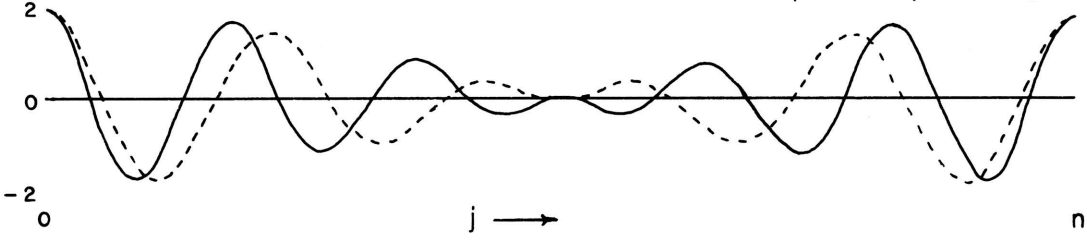


Figure 1a.-Theoretical plots of height (dashed) and density (continuous) of vegetation along traverse line.

components in each series.

This is easily accomplished if the untenable but useful assumption is made that the series is periodic and repeats itself indefinitely both forward from the end and backward from the start of the traverse. Classical Fourier analysis may now be applied. Figure 2 shows such a breakdown of frequencies (or wavelengths) for the series in Figure 1a. In this idealized example height of vegetation has two frequency components, 4 and 5 cycles per traverse length, and density also has two, 5 and 6 cycles per traverse length. If the usual correlation coefficient is now calculated for the separate frequencies it will be seen that there is no correlation for 4 and 6 cycles but a correlation of 1.0 at 5 cycles per traverse length. The real world is never this simple, but the exercise has demonstrated what many have stated before, that scale is important in correlation and unless scales are separated they tend to blur analyses (see Casetti, 1966).

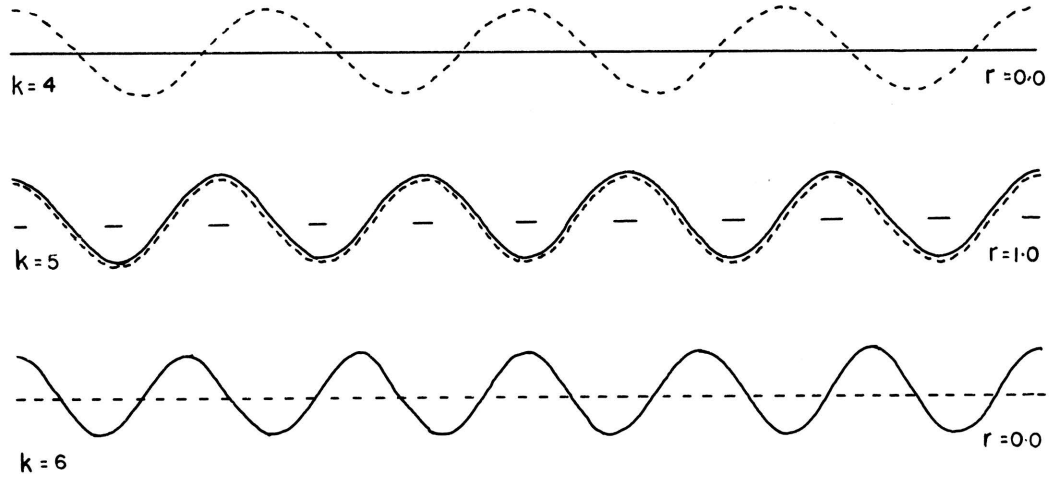


Figure 2.- Frequency breakdown of data in Figure 1a.

The same study reveals other limitations in the use of the simple correlation coefficient. In order to discuss these it will be useful to write x and y in terms of the sinusoidal functions. Thus, the magnitude of the kth frequency component of at point j may be written

$$x_k[j] = A_x[k]\cos\left(\frac{2\pi kj}{n} - \phi_x[k]\right) = a_x[k]\cos\left(\frac{2\pi kj}{n}\right) + b_x[k]\sin\left(\frac{2\pi kj}{n}\right) \quad (2)$$

where

$A_x[k] = (a_x^2[k] + b_x^2[k])^{1/2}$ = amplitude of kth frequency in the x series,

n = number of points along the traverse,

and,

$\phi_x[k] = \tan^{-1}\left(\frac{b_x[k]}{a_x[k]}\right)$ = the phase angle, a constant for the kth frequency in the x series, (Barber, 1961, 1966; Lanczos, 1956; and Harbaugh and Preston, 1965). The squaring, division by n, and integration of equation (2) gives the variance of x,

$$\text{var}(x) = \frac{A_x^2[k]}{2} = \frac{a_x^2[k] + b_x^2[k]}{2}. \quad (3)$$

By defining $y_k[j]$ in the same way the variance of y may similarly be obtained. Also the covariance at the kth frequency may be found in this way,

$$\text{cov}(xy) = c_{xy}[k] = \frac{A_x[k]A_y[k]}{2} \times \cos(\phi_y[k] - \phi_x[k]) \quad (4)$$

$$= \frac{a_x[k]a_y[k] + b_x[k]b_y[k]}{2}. \quad (5)$$

With the substitution from (3) and (4), (1) may now be rewritten for r at the kth frequency

$$r_{xy}[k] = \cos(\phi_y[k] - \phi_x[k]). \quad (6)$$

In other words, the correlation coefficient for two periodic functions at the same frequency is dependent only upon the phase difference between the two series. The amplitudes may be any magnitude greater than zero. This may at first seem reasonable. Completely in-phase series ($\phi_y[k] - \phi_x[k] = 0$) vary together and have a correlation of 1. Completely out-of-phase series ($\phi_y[k] - \phi_x[k] = \pi$) vary inversely and therefore have a correlation of -1. The question now arises as to whether two series with a phase difference of $\frac{\pi}{2}$ or $\frac{3\pi}{2}$ should have an r of zero, as this is what the algebra gives us. Zero would also be obtained from the correlation of the two series at the same frequency when one or both have zero amplitudes. Obviously there is a fundamental difference between these two cases.

One way of solving this problem is to define a new correlation term which ignores the phase and then, in any discussion, to state both this new term and the phase. The covariance discussed so far is the in-phase covariance: that is, the usual covariance. The set of covariance, $c_{xy}[k]$, as a function k is usually known as the cospectrum.

Another covariance, not yet introduced, is the out-of-phase covariance which may be obtained by displacing y, $\frac{\pi}{2}$ radians. This produces an equation equivalent to (4)

$$q_{xy}[k] = \frac{A_x[k]\,A_y[k]}{2} \times \sin(\phi_y[k] - \phi_x[k]) \tag{7}$$

$$= \frac{a_x[k]\,b_y[k] - a_y[k]\,b_x[k]}{2}. \tag{8}$$

A set of the out-of-phase covariances is known as the quadrature spectrum.

Together the cospectrum and quadrature spectrum provide a measure of the cross spectrum, $1/2\,A_{xy}[k]$,

$$1/2\,A_{xy}[k] = (c_{xy}^2[k] + q_{xy}^2[k])^{1/2}, \tag{9}$$

and the phase difference, the displacement of y forward of x,

$$\phi_y[k] - \phi_x[k] = \phi_{xy}[k] = \tan^{-1} \times (q_{xy}[k] / c_{xy}[k]). \tag{10}$$

The new correlation term may now be defined as the cross covariance divided by the standard deviations

$$W_{xy}[k] = A_{xy}[k]/(A_x[k]\,A_y[k]). \tag{11}$$

The square of this term, $W_{xy}^2[k]$, may be thought of as the ratio of the variance of the product series, when phase is ignored, to the product of the variances of the individual series. A little algebra will show that, for an integer frequency of a strictly periodic function, $W_{xy}[k]$ is always 1, except where a particular frequency is missing.

This may seem an insignificant result and it is for periodic functions. However, these equations have important uses in section 5 where the data are more realistic: that is, are nonperiodic. Then, $W_{xy}[k]$ may vary between zero and 1.

TWO-DIMENSIONAL SPECTRAL RELATIONSHIPS

All above equations may be reproduced for the two-dimensional case. From arrays of $x[j1, j2]$ and $y[j1, j2]$, the amplitudes, $a_x[k1, k2]$, $a_y[k1, k2]$, $b_x[k1, k2]$ and $b_y[k1, k2]$, may be calculated where k1 refers to the ordinate and k2 to the abscissa. This is a slight variation on the work of Harbaugh and Preston (1965) in that only two amplitudes for each series are obtained. As before the a's are the cosine amplitudes and the b's are the sine amplitudes. The meaning of any particular amplitude may be easily visualized if its position in the array is considered in terms of polar coordinates. The particular amplitude is the amplitude of a series of parallel waves, which are at right angles to the position vector and which have a frequency proportional to the length of the vector. The vector direction, θ, measured from the abscissa, is

$$\theta = \tan^{-1}(k1/k2), \tag{12}$$

and frequency, k3

$$k3 = (k1^2 + k2^2)^{1/2} \tag{13}$$

It should be noted that k's are in cycles per length of side of array [n1, n2]. These units are most useful since the k's remain integers and therefore may be used as subscripts. However, for equations (12) and (13), they must be in the same dimensional units. In cases where the array is not square ($n1 \neq n2$) and/or the data spacings are not the same ($\Delta j1 \neq \Delta j2$) the k's must be considered in cycles per unit length. That is, k1 and k2 in (12) and (13) should be replaced by $k1/(n1\,\Delta j1)$ and $k1/(n2\,\Delta j2)$ respectively.

An example of a surface produced by a

single amplitude in the array, a[4,2], is shown in Figure 3. These waves have one crest at 0,0. If the first crest were elsewhere the phase angle would be given by a magnitude in the b array at [4, 2]. If only two amplitudes are present the two sets of waves will interact. If the two sets are at right angles they will produce symmetrical depressions and elevations as shown in Figure 4. The addition of more frequencies will produce further variations.

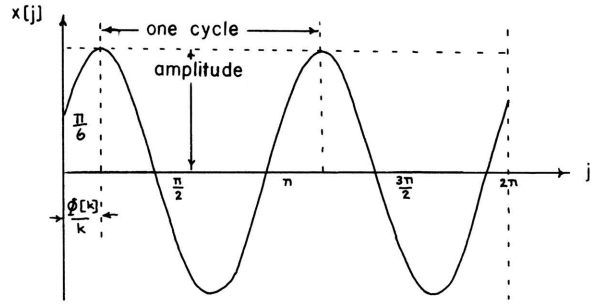


Figure 3.-An example of an x array which produces one amplitude at [4, 2].

The equations remain similar, [j] is replaced by [j1, j2] and k by [k1, k2] and the summations become double summations. Thus the equation for a [k1, k2] may be written

$$a_x[k1, k2] = \frac{2}{n1\, n2} \sum_{j1=0}^{n1-1} \sum_{j2=0}^{n2-1} x_{k1\, k2}[j1, j2] \times \cos\left(\frac{2\pi k1 j1}{n1} + \frac{2\pi k2 j2}{n2}\right), \qquad (14)$$

and,

if cosine is replaced by sine, (14) becomes the equation for $b_x[k1, k2]$. It provides all the information for the calculation of

$$r_{xy}[k1, k2],\ c_{xy}[k1, k2],\ q_{xy}[k1, k2],$$

$$\phi_{xy}[k1, k2] \text{ and } W_{xy}[k1, k2]$$

because each is a function of the a's and b's.

CALCULATION OF THE AMPLITUDES - THE FAST FOURIER TRANSFORM

In the past, most computer programs have used the Goertzel method (Hamming, 1962) for calculating the Fourier amplitudes. Recently a new computational algorithm has become available, and has been adapted for use by Cooley and Tukey (1965). Gentleman and Sande (1966) have further modified some of the procedures and Sande has produced excellent Fourier Transform programs. Readers are referred to the quoted papers for details. Essentially, the algorithm reduces the number of computer operations from n^2 to nlogn where n is the number of data points. This is a considerable reduction, particularly where large arrays are involved.

NONPERIODIC DATA

Excepting for the variation of elements around a latitude circle, a fact used extensively by meterologists in studying the energy conversions in the atmosphere (Saltzman, 1957; Wiin-Nielset and others, 1963, 1964), very little naturally occurring data are truly periodic. Consequently the unmodified results of the above equations have limited value. All the amplitudes are for integer frequencies in the range 0 to $\frac{n}{2}$ for one dimensional series, and 0,0 to $\frac{n1}{2}$, $\frac{n2}{2}$, for the two-dimensional case. If more data are added or some removed the new set of frequencies will not correspond to the old. In other words, a whole set of amplitudes are realizable between a particular pair of frequencies. The effect of this is clearly demonstrated in statistical terms by the fact that the number of degrees of freedom associated with each line amplitude, $A_x[k]$, is only 2.

There are two alternatives for obtaining reliable estimates of the frequency components in nonperiodic data. Presently, the most widely known and used technique is that developed by Tukey (1949) and Blackman and Tukey (1958) from a theorem by Wiener (1930). For a single set of data the autocovariances are first calculated and then these are subject to a cosine transform. Final adjustments must be made by applying a spectral window function to allow for the fact that the autocovariances came from a finite sample of data. Also, some initial modification, called prewhitening, may be made so that large peaks or valleys at one point in the final spectrum do not affect other frequencies. Resulting figures are estimates of the variance in frequency bands rather than in lines. Alternatively they may be in the form of variance density.

For two sets of data the average lagged products are subject to cosine and sine transformations to give estimates of the cospectrum and the quadrature spectrum (Goodman, 1957; Jenkins, 1961, 1962, 1963, 1965). From these spectra the other statistics may be obtained through the use of equations (9) to (11). The $W_{xy}[f]$, where f refers to the central frequency of the band, is now called the coherence. The name coherence or coherency is also sometimes applied to the square of $W_{xy}[f]$. It varies between 0 and 1 and is a measure of how well the two series are related in that band. This

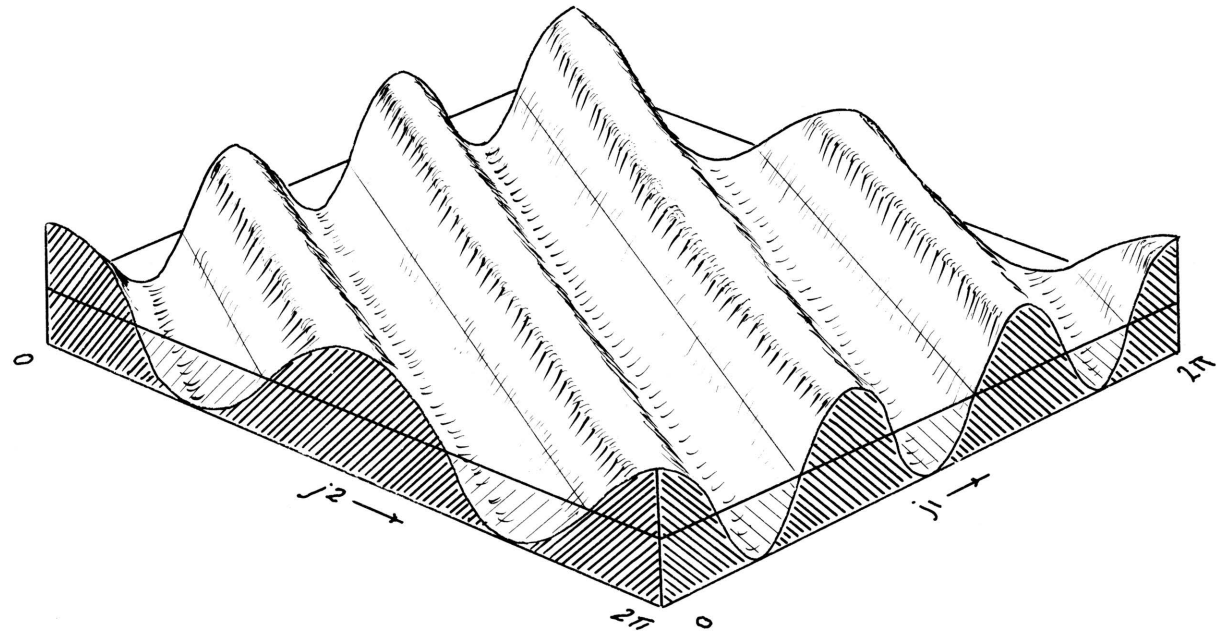


Figure 4.-An example of part of an x array which produces two amplitudes; one at [k1, k2] and another at [-k1, k2], where |k1| = |k2| .

is a most useful measure of association between two sets of data. It supplies information on the degree of the association between data with respect to scale, something which is not available with the simple correlation coefficient. Furthermore, the phase difference provides estimates of the average direction of the relationship; whether the series in that band of frequencies vary together or vary in opposition, etc. A number of examples of the application of this method are given in Lumley and Panofsky (1964).

What has been said above about the one-dimensional case also applies to the two-dimensional case. In geology the equations have recently been given by Preston (1966). Explicit papers on the application of this form of two-dimensional analysis are by Pierson (1960) and Leese and Epstein (1963).

The second method was suggested by Blackman and Tukey in 1958 (p. 90-95). This was to "calculate many values of [the line spectrum] and then average these results over moderately wide frequency intervals. Again our estimates will be estimates of considerably smoothed spectral densities; again our estimates will be moderately stable." Now that the Fast Fourier Transform is available, Tukey (1966) advocates this technique for obtaining spectral estimates.

For the one dimensional case, the estimates for a band 2m+1 wide centred at f are given by

$$A_x^2[f] = \frac{1}{2m+1} \sum_{k=f-m}^{f+m} (a_x^2[k] + b_x^2[k]) \qquad (15)$$

$$C_{xy}[f] = \frac{1}{2m+1} \sum_{k=f-m}^{f+m} 1/2 \quad \times$$

$$(a_x^2[k]\, a_y[k] + b_x[k]\, b_y[k]) \qquad (16)$$

$$q_{xy}[f] = \frac{1}{2m+1} \sum_{k=f-m}^{f+m} 1/2 \quad \times$$

$$(a_x[k]\, b_y[k] - a_y[k]\, b_x[k]) \qquad (17)$$

Equations (9), (10) and (11) may now be applied by replacing k by f. For the two-dimensional case k in (15), and (16) and (17) is replaced by k1, k2, and the individual terms are doubly averaged.

As in the first method a spectral window must be applied and some pre-whitening is usually necessary. As Tukey (1966) points out, arithmetic speed has been gained but many of the old problems still remain.

SUMMARY AND CONCLUSION

It has been demonstrated that the degree of association between one- and two-dimensional series is a function of scale. In any analysis of association, then, the isolation of the scale components of the data should be a routine procedure. Furthermore, it has been pointed out that the simple correlation coefficient has limited use at a specific frequency since it is a function only of phase difference. Another measure which might be

used is coherence with phase.

The earlier equations were given for truly periodic data which seldom occur in nature. However, these equations may be modified by averaging so that they may apply to the nonperiodic situation. In the past this method has seen little use but with the introduction of the efficient Fast Fourier Transform it should become a universal technique. If needed, the auto and lag covariances may be calculated by transforming back the final spectral estimates obtained by the direct method.

REFERENCES

Barber, N.F., 1961, Experimental correlograms and Fourier transforms: International tracts in computer science and technology and their appreciation, Pergamon, London, v. 5, 136 p.

Barber, N.F., 1966, Fourier methods in geophysics, in Methods and techniques in geophysics, ed. S.K. Runcorn: John Wiley and Sons, New York, v. 2, p. 123.

Blackman, R.B., and Tukey, J.W., 1958, The measurement of power spectra: Bell System Tech. Jour., v. 37, 190 p.

Casetti, E., 1966, Analysis of spatial association by trigonometric polynomials: Canadian Geographer, v. 10, no. 4, p. 199-204.

Cooley, J.W., and Tukey, J.W., 1965, An algorithm for the machine calculation of complex Fourier series: Mathematics of Computation, v. 19, p. 297-301.

Gentleman, W.M., and Sande, G., 1966, Fast Fourier transforms -- for fun and profit: Proc. 1966 Fall Computer Conf., A.F.I.P.S., 52 p.

Goodman, N.R., 1957, On the joint estimation of the spectra, cospectrum and quadrature spectrum of a two-dimensional stationary quassian process: Scientific Paper no. 10, Engineering Stat. Lab., New York Univ.

Hamming, R.W., 1962, Numerical methods for scientists and engineers: McGraw-Hill, New York, 411 p.

Harbaugh, J.W., and Preston, F.W., 1965, Fourier series analysis in geology: Univ. Arizona, Coll. Mines, v. 1, p. R-1 - R-46.

Jenkins, G.M., 1961, General considerations in the analysis of spectra: Technometrics, v. 3, p. 133.

Jenkins, G.M., 1962, Cross spectral analysis and the estimation of linear open loop transfer functions, in Symposium on time series analysis: Proc., Brown Univ., John Wiley and Sons, New York.

Jenkins, G.M., 1963, An example of the estimation of a linear open loop transfer function: Technometrics, v. 5, p. 227.

Jenkins, G.M., 1965, A survey of spectral analysis: Applied Stat., v. 14, no. 1, p. 2-32.

Lanczos, C., 1956, Applied analysis: Prentice Hall, Inc., Englewood Cliffs, New Jersey, 608 p.

Leese, J.A., and Epstein, E.S., 1963, Application of two-dimensional spectral analysis to the quantification of satellite cloud photographs: Jour. Applied Met., v. 2, no. 5.

Lumley, J.L., and Panofsky, H.A., 1964, The structure of atmospheric turbulence: Interscience monographs and texts in physics and astronomy, 12, John Wiley and Sons, New York, 239 p.

Pierson, W.F., Jr., ed., 1960, The direction of a wind generated sea as determined from data obtained by the stereo wave observation project: Meteorological Papers, v. 2, no. 6, New York Univ., Coll. Engineering, 88 p.

Preston, F.W., 1966, Two-dimensional power spectra for classification of land forms: Kansas Geol. Survey Computer Contr. 7, p. 64-69.

Saltzman, B., 1957, Equations governing the energetics of the large scales of atmospheric turbulence in the domain of wave number: Jour. Meteor., v. 14, p. 513.

Tukey, J.W., 1949, The sampling theory of power spectrum estimates, in Symposium on application of auto-correlation analysis to physical problems: Office of Naval Research, Woods Hole, Mass., p. 47-67.

Tukey, J.W., 1966, Spectrum calculations in the new world of the fast Fourier transform: Paper presented at the Advanced Seminar on Spectral Analysis of time series at Madison, Wisconsin, October.

Wiener, N., 1930, Generalized harmonic analysis: Acta. Math., v. 55, p. 117.

Wiin-Nielsen, A., Brown, J.A., and Drake M., 1963, On atmospheric energy conversions between the zonal flow and the eddies: Tellus, v. 15, p. 261.

Wiin-Nielsen, A., Brown, J.A., and Drake M., 1964, Further studies of energy exchange between the zonal flow and the eddies: Tellus, v. 16, p. 168.

THE GENERAL LINEAR MODEL IN MAP PREPARATION AND ANALYSIS[1/]

by

W.C. Krumbein

Northwestern University

ABSTRACT

The general linear model is the connecting thread that weaves through many aspects of map preparation, map analysis, and map interpretation. In its conventional form the general linear model is the basis for such multivariate maps as the Q-mode factor map and the discriminant function map. In its two-dimensional form the general linear model is the basis for the polynomial and Fourier map-analysis models, both of which are widely used in a variety of geological fields. All variants of the general linear model can be reduced to a simple matrix equation, $S\hat{\beta} = g$, where S is the matrix of uncorrected sums of squares and cross-products of the independent variables, g is the column vector of the sums of the dependent variable and its crossproducts with the independent variables, and $\hat{\beta}$ is the column vector of estimated linear coefficients.

Coefficient space, developed in the earlier part of this report, is further examined here, and the extension of map analysis to the estimation of "trends" in coefficient matrices is introduced with an example.

INTRODUCTION

Analysis of contour-type maps by computer has undergone several stages of evolution since its introduction into geology a dozen years ago. The search for the underlying "true trend" in a given set of map data has been broadened by recognition that low-order fitted surfaces, such as the linear and quadratic, also give important insight into the structure of the mapped data. Secondary trend components, i.e., systematic effects of relatively high order that seem to linger in the residuals, have also proved valuable in map interpretation, as have the map residuals themselves. All of these influence selection of prediction models for map interpolation and extrapolation, of techniques for map comparison, and development of process models that account for the observed systematic map patterns.

The classical polynomial model, widely available in computer programs for gridded and nongridded data since the late 1950's, has been supplemented in recent years by the double Fourier series model, also programmed for gridded and nongridded map data. The Fourier model opened the way for study of periodic patterns in maps, in contrast to the nonperiodic surfaces obtained with the polynomial model. The problem of identifying the "true map trend" is, if anything, slightly more complicated now that two ways of structuring map data for trend analysis are available.

The purpose of this paper is to examine map analysis in the larger framework of map preparation, analysis, interpretation, and comparison, as an extension of an earlier report (Krumbein, 1966) in which the structure of coefficient matrices under the polynomial and Fourier models was examined.

Four aspects of map study are summarized in Table 1. Kinds of variables used in facies mapping, and methods of preparing maps, which include topics (1a) and (1b), are discussed and described in Bishop (1960) and Forgotson (1960). Multivariate maps in topic (1c) are illustrated by Q-mode factor (vector) maps as introduced with a computer program by Imbrie (1963). Griffiths and his students (cited in Griffiths, 1966) have used discriminant functions in mapping barren and potentially favorable sands in Pennsylvania. Pelto (1954) applied statistical entropy functions to multivariate mapping. The subject is further treated by Forgotson (1960), and an example is given in Miller and Kahn (1962, p. 427).

[1/] This research was conducted under ONR Task No. 388-078, Contract Nonr-1228(36), Geography Branch of the Office of Naval Research. This paper is an extension of a report on the "Classification of map surfaces based on the structure of polynomial and Fourier coefficient matrices," published in Computer Contribution No. 7 of this Series (1966). These two papers constitute Technical Report No. 5 of the contract specified above. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Table 1.-Topics in the study of areally distributed data.

(1) Map Preparation, i.e., ways of structuring areally distributed data for contour mapping.
  (1a) Raw data used directly for univariate maps (thickness of a stratigraphic unit, isolith maps of particular rock types, elevation of marker beds, etc.).
  (1b) Simple combinations of raw data (percentages, ratios, etc.).
  (1c) Maps based on multivariate attributes (Q-mode factor maps, discriminant function maps, statistical entropy maps).

(2) Map Analysis, i.e., ways of seeking "trends" in mapped data, by fitting surfaces of various orders to maps prepared as in (1).
  (2a) Polynomial analysis (nonperiodic fitted surfaces).
  (2b) Fourier series analysis (periodic or cyclical fitted surfaces).
  (2c) Iterative polynomial or Fourier analysis of coefficient-matrices and Z^2-matrices.

(3) Map Comparison, i.e., ways of evaluating the similarities or differences between maps.
  (3a) Direct comparison of individual coefficients.
  (3b) Comparisons based on orders of fitted surfaces.

(4) Map Interpretation, i.e., ways of deriving generalizations from areally distributed data.
  (4a) Conceptual process models.
  (4b) Deterministic or probabilistic models designed to "explain" the observed patterns of areal variation.
  (4c) Development of predictor models.

Map analysis, shown as topic (2) in Table 1, includes the polynomial and Fourier models as mentioned, and topic (2c) extends these models into the analysis of coefficient-matrices and Z^2-matrices in "coefficient space." This topic is expanded in a later section. Map comparison, topic (3) in Table 1, is based partly on direct comparison of polynomial or Fourier coefficients, or on various orders of surface. The last topic in the Table is that of map interpretation. This is largely substantive, although various statistical techniques can be helpful. In general, the intent of map interpretation is to develop conceptual or other kinds of models to account for trends and residuals, as well as to derive generalizations about sedimentary, tectonic, and other influences in basin development and fill. Potter (1962), for example, developed a "basin model" by analysis and interpretation of maps showing the distribution of Pennsylvanian sandstones in the Illinois Basin.

A striking feature of Table 1 is that topics (1c), all of (2), and part of (3) represent applications of the general linear model to the study of areally distributed data. This model, in various forms, is the connecting thread that weaves through much of map preparation, map analysis, and map comparison, as well as forming the foundation, in conjunction with substantive geological reasoning, for map interpretation. The following section affords a brief review of the particular forms that the linear model may take in this context. These variants are sometimes used sequentially; thus, Q-mode factor analysis may be used to obtain factor maps, which can be analyzed for trends with the polynomial or Fourier models, and compared with other maps by study of common elements in linear coefficient vectors or matrices.

THE GENERAL LINEAR MODEL

The general linear model in its conventional form can be stated as follows:

$$W = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + e \qquad (1)$$

where W is an observable random variable; X_1, X_2, ..., X_k represent observable independent variables measured without error; the β's are unknown parameters; and e is an unobservable random variable with mean zero and variance σ^2.

Factor analysis and two-group discriminant functions can be directly derived from this model, which thus serves its purpose in the preparation of the corresponding multivariate maps. For map analysis the general linear model is expressed in its two-dimensional form, as follows:

$$W = \beta_{00} + \sum_{i=1}^{m} \sum_{j=1}^{n} \beta_{i,j} X_i Y_j + e \qquad (2)$$

Here W is an observable random variable (the mapped variable in this context), X_1, X_2, ..., X_m and Y_1, Y_2, ..., Y_n represent observable independent variables measured without error (these become functions of geographic coordinates in the present context), the β's are unknown parameters (the coefficients of the fitted surfaces), and $\underline{e}$ is an unobservable random variable with mean zero and variance σ^2, representing the residuals on the fitted surface.

The polynomial version of equation (2) involves simply the change of the subscripts on X and Y to superscripts:

$$W = \beta_{00} + \sum_{i=1}^{m} \sum_{j=1}^{n} \beta_{i,j} X^i Y^j + e \qquad (3)$$

Here X^i and Y^j represent successive powers of the map coordinates X and Y.

The Fourier model can be expressed in the general form of equation (2) as follows:

$$W = \beta_{00} + \sum_{i=1}^{M} \sum_{j=1}^{N} F(\beta_{i,j} P_i Q_j) + e \qquad (4)$$

where $P_i = 2\pi iX/M$ and $Q_j = 2\pi jY/N$. Here $M = m + 1$, and $N = n + 1$. The $\beta_{i,j}$ of the general model of equation (2) now become a series of sine and cosine terms, yielding generally four Fourier coefficients for each ij:

$$F(\beta_{i,j} P_i Q_j) = cc_{i,j} \cos P_i \cos Q_j + cs_{i,j} \cos P_i \sin Q_j + sc_{i,j} \sin P_i \cos Q_j + ss_{i,j} \sin P_i \sin Q_j \qquad (5)$$

Thus, instead of having simply β_{23}, for example, the corresponding Fourier coefficients are cc_{23}, cs_{23}, sc_{23}, and ss_{23}. (The function F has value 1.0 in equations 2 and 3.)

The close relations between the conventional and the two-dimensional forms of the general linear model becomes more apparent when equations (1) and (2) are expressed in matrix form, whereupon all variants can be condensed to the relatively simple expression:

$$\underline{S}\hat{\underline{\beta}} = \underline{g} \qquad (6)$$

where $\underline{S}$ is the matrix of uncorrected sums of squares and crossproducts of the X's in equation (1), of the X's and Y's in equations (2) and (3), and of the P's and Q's in equation (4). The column vector of estimated linear coefficients is represented by $\hat{\beta}$, and $\underline{g}$ is the column vector of the sums of W and of its crossproducts with the independent variables in each equation.

The forms of the S-matrix and vector-g differ according to the structure assigned to a particular variable W. For the conventional linear regression case of equation (1), where W is structured as:

$$W = \beta_{00} + \beta_1 X_1 + ... + \beta_k X_k + e \qquad (7)$$

the matrix-S and vector-g have the form shown in Krumbein and Graybill (1965, p. 287). For a discriminant function the S-matrix is more complicated in that it is a combination of two submatrices $\underline{S}_1$ and $\underline{S}_2$, one for each multivariate population, and vector-g is generated by taking the differences between the means of the several variables in the two populations. The factor analysis model can also be reduced to the form of equation (6). Of equal importance is the fact that both the polynomial and Fourier models, representing two-dimensional versions of the general linear model, can also be expressed directly in the same form. The matrix-S and vector-g for the nongridded polynomial model are given in Krumbein and Graybill (1965, p. 330), and for the double Fourier series model in James (1966).

The fact that all versions of the general linear model, as used in the preparation and analysis of multivariate maps, can be reduced to a relatively simple expression, has certain advantages for computer use. Once programs have been written for generating the particular forms of matrix-S and vector-g involved in a given map model, equation (6) can be solved by an operation that consists simply in multiplying both sides of the equation by the inverse of $\underline{S}$, $\underline{S}^{-1}$:

$$\underline{S}^{-1}\underline{S}\hat{\underline{\beta}} = \underline{S}^{-1}\underline{g}$$

Inasmuch as $\underline{S}^{-1}\underline{S} = \underline{I}$, the identity matrix, this solution reduces to:

$$\hat{\underline{\beta}} = \underline{S}^{-1}\underline{g} \qquad (8)$$

An added advantage of the condensed model of equation (6) is that the inverse of the matrix-S, $\underline{S}^{-1}$ is the variance-covariance matrix of the

coefficients, and hence this may be used in setting confidence intervals on individual coefficients, or even on the fitted surfaces, at least those of low order (Krumbein, 1963). Although the factor-analysis model can be reduced to the form of equation (6), it appears that conventional computational methods, as described by Harman (1960), and used by Imbrie (1963) are more convenient than the condensed form for the whole succession of obtaining eigenvalues, extracting the principal components, and rotating them into the final factor output.

Map comparison methods, insofar as they involve the estimated coefficients of the maps being prepared, (Miller, 1964; Merriam and Sneath, 1966), are based on comparisons of $\hat{\beta}$-vectors (or matrices) derived from the polynomial or Fourier models. However, both of the above references describe other techniques of map comparison, and Merriam and Sneath include references to other approaches, based in large part on the simple linear model (use of correlation coefficients, for example) or on the general linear model.

$\hat{\beta}$-VECTORS AND ARRAYS IN COEFFICIENT SPACE

The previous section emphasized the general linear model for nongridded map data. In this section we shift to gridded data for convenience of discussion, although the same general remarks apply in both instances. The vector-$\hat{\beta}$ of equation (6), as it applies to equation (1), becomes a matrix $[\hat{\beta}_{i,j}]$ for the two-dimensional cases of equation (2), although in practice the matrix may be printed out in vector form by computer programs. In the gridded case, when the map is analyzed by the polynomial or Fourier models, the matrix $[\hat{\beta}_{i,j}]$ has as many coefficients and $[Z^2_{i,j}]$ values as there are elements in the data matrix $[W_{i,j}]$. As developed in the earlier part of this report (Krumbein, 1966, p. 12) the coefficients can be considered as occupying (i, j) points on a plane with coordinates that are subscripts of the coefficients. The coefficients themselves (and their associated Z^2 array) become maps in that space, and these new maps can be analyzed in turn for their own "trends."

We illustrate the polynomial case with a "trend map" of the coefficients. Table 2 lists the original data, the polynomial coefficients obtained from the data, and the polynomial coefficients of the coefficients. Figure 1 shows the linear and quadratic surfaces associated with the two coefficient matrices. The left map is fitted to the original map data, which represent the mean grain size of beach sand discussed on page 16 in the earlier part of this report. Several interesting features are shown by the two maps. Thus in terms of the observed beach data, collected on a geographic grid, the surface is a hyperboloid (left map), and the linear plus quadratic components account for 44.1 + 33.0 = 74.1% of the total sum of squares of the observed data. The other map has the linear and quadratic components of the coefficients in (i, j) space. It is an ellipsoid, with some negative values, and the sum of squares accounted for are 32.1 + 32.9 = 65.0% of the sum of squares associated with the observed map coefficients. Thus, the "coefficient trend" map has a different form and a smaller linear component than the "geographic trend" map of the original observations. The total corrected sum of squares of the observed beach data, 0.6176, has dropped to 0.0849 for the coefficients themselves.

This sequential analysis of map data, followed by analysis of the map coefficients, then by analysis of the coefficients of the coefficients, and so on, tends toward a matrix of zeroes after several iterations. The justification for including this topic is to suggest that inasmuch as the coefficient matrix based on the observed data represents a degree of generalization of the original map data, the properties of this generalization may well be worth additional study.

PROPERTIES OF COEFFICIENT SPACE

Partitioning of coefficient space according to diagonals, blocks, etc., covered in the earlier part of this report, led to some applications, as in map screening, and to several interesting speculations. One concerned the nature of polynomial maps based on noninteger values of the coefficients. An example presented orally with the first report is included here mainly as a matter of record. It is based on equation (3), in which X^i and Y^j have powers equal to the subscripts in $\beta_{i,j}$. If we set $i = j = 1/2$, and for simplicity let $\beta_{00} = \beta_{i,j} = 1$, then a "linear map" with $i + j = 1$ can be developed with the model:

$$W = \beta_{1/2\ 1/2} X^{1/2} Y^{1/2} \qquad (9)$$

Figure 2 shows this map, and its linear property is the equidistant spacing of the intercepts of all rays from the origin as they cut across the curved contours of the surface itself. It was suggested that such "linear" maps may have potential value in studying the curvature of a linear surface around the edge of an elongate sedimentary of structural basin.

An additional potentially important aspect of noninteger (i, j) positions in coefficient space is related to the equestion of the orthogonalization of a regression function for nongirdded map data.

Table 2.-5 x 5 grid of beach geometric means Lake Michigan, Evanston, Illinois (iterative analysis by polynomial model).

		Data by Rows		
0.210	0.197	0.205	0.200	0.200
0.322	0.230	0.239	0.224	0.200
0.224	0.211	0.203	0.213	0.198
0.210	0.233	0.223	0.214	0.236
0.350	0.584	0.521	0.703	0.726
		Coefficients by Rows		
0.2910	0.0117	-0.0006	0.0009	-0.0023
0.0729	0.0206	-0.0035	0.0048	-0.0031
0.0480	0.0145	-0.0024	0.0025	-0.0018
0.0414	0.0032	0.0002	-0.0019	-0.0010
0.0025	0.0020	-0.0005	0.0002	-0.0004
		Coefficients of Coefficients by Rows		
0.0200	-0.0195	0.0121	-0.0075	0.0005
-0.0129	0.0126	-0.0083	0.0056	-0.0006
0.0051	-0.0052	0.0040	-0.0030	0.0005
-0.0039	0.0043	-0.0034	0.0025	-0.0006
0.0004	-0.0004	0.0002	-0.0002	0.0000

W.R. James has examined some aspects of this question with the Fourier model at Northwestern University, and the following as yet unpublished material is included with his permission.

James started with gridded data, which yields independent integer-valued coefficients of the Fourier surface. The cosine-cosine coefficients were then examined as frequencies over the coefficient plane, to establish the following properties of coefficient space:

(1) The space is continuous.

(2) Mirror images of contoured coefficient space occur over every interval equal to the normal highest harmonic frequency. Thus all information is contained in a finite area.

(3) Relative strengths of trends in grid-parallel directions are maintained at least approximately.

(4) Directional properties and the scale of variability are retained in the Fourier transform.

There is no reason to anticipate that these properties do not also apply to coefficient space when the map data are nongridded. However, in nongridded data the integer-valued coefficients are not independent, but it appears at least intuitively sound to suggest that such independent coefficients may occupy noninteger positions on the coefficient plane. Still, the search for uncorrelated frequencies (wavelengths) proved to be unusually difficult, and has thus far been unsuccessful. Nevertheless, the preliminary study has led to techniques for continuous mapping of the (i, j) plane in coefficient space in terms at least of the cosine-cosine Fourier coefficients.

CONCLUDING REMARKS

This paper shifted its emphasis more than once during preparation, owing in part to the influx of additional concepts and questions as the work proceeded. Initially it was intended to enlarge upon some topics developed in the earlier part of the study, especially with regard to the use of combinatorial models in map analysis. A typical example of this is the use of the polynomial model to extract one or more low-order surfaces from the map data, followed by examination of the residuals by Fourier analysis. It seemed more appropriate within the intent of this Colloquium to bring out the pervasiveness of the general linear model in trend analysis as well as in the development of multivariate mapping techniques.

It is also appropriate to remind the reader that trend analysis is in practice largely a search procedure, in which fitted surfaces are obtained that reflect to some "reasonable" degree the attributes of the real-world phenomenon under study. Trend-surface analysis has amply demonstrated its worth in the development of models for exploration of natural resources, as well as for the building of process models for better understanding of underlying physical processes that produce the trend surfaces as natural responses. In the long run, however, this search procedure will be supplemented by an approach to trend analysis that starts with a conceptual, probabilistic, or deterministic model that predicts certain kinds of surfaces, and then uses observational data to test the models. In this

extension it is likely that nonlinear models may prove equally as valuable as the general linear model. James in this Colloquium (1967) develops some preliminary aspects of this approach.

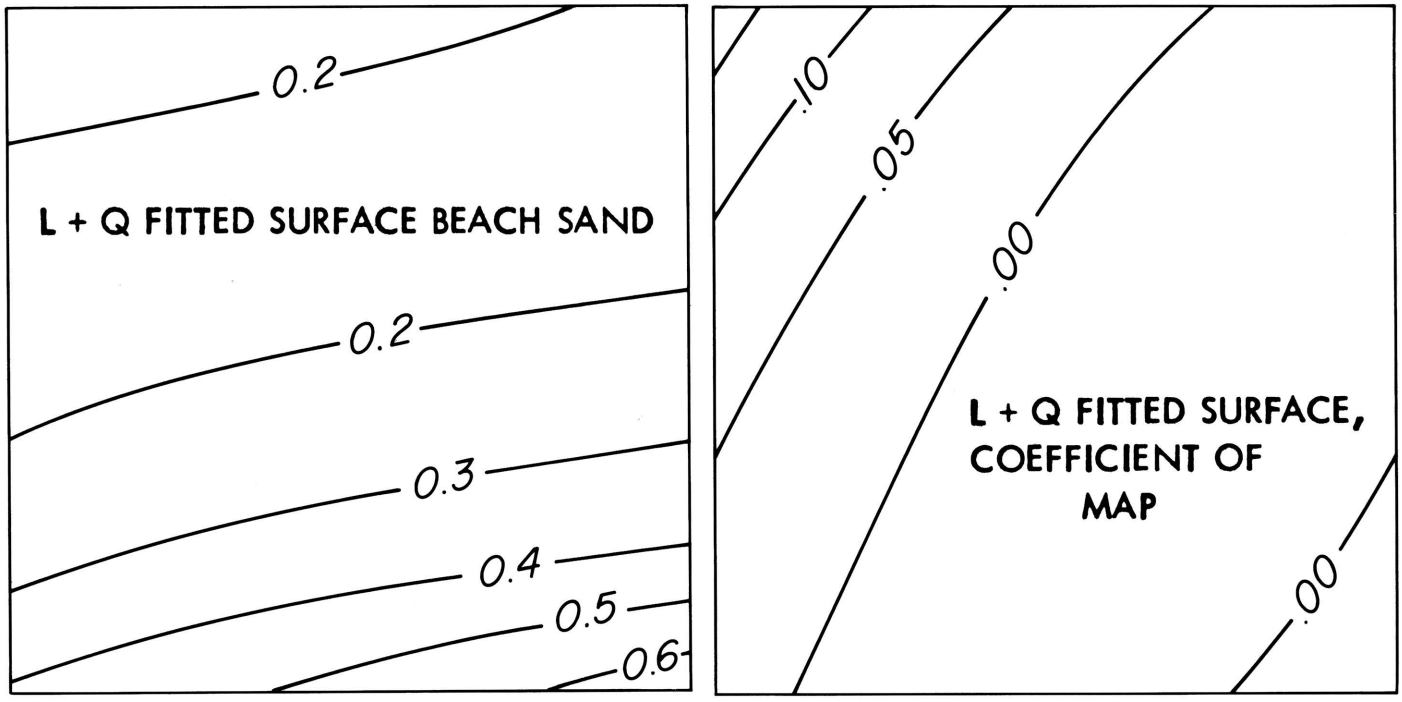


Figure 1.- Linear + quadratic fitted surfaces on (left map) geometric mean diameter in mm of beach sand; and (other map) on polynomial coefficients of observed beach data (see Krumbein, 1966, p. 16).

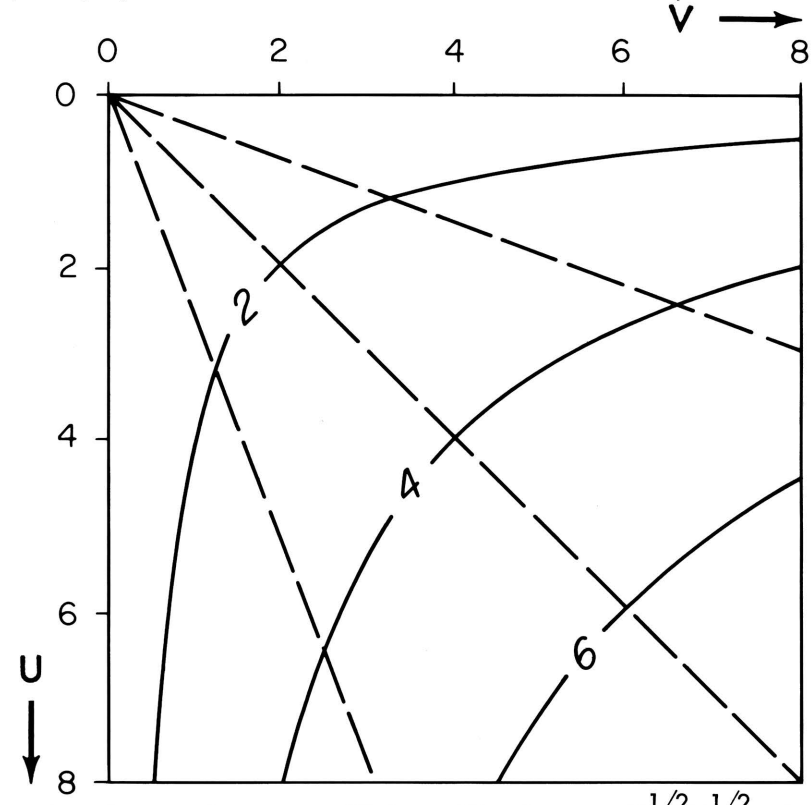


Figure 2.-"Linear" map based on noninteger coefficient, $W = \beta_{1/2\ 1/2} X^{1/2} Y^{1/2}$, showing several contours of surface (solid lines) and rays from origin (dashed lines). X and Y coordinates are expressed as (U, V), with origin at upper left. See text for details.

REFERENCES

Bishop, M.S., 1960, Subsurface mapping: John Wiley and Sons, Inc., New York, 198 p.

Forgotson, J.M., Jr., 1960, Review and classification of quantitative mapping techniques: Am. Assoc. Petroleum Geologists Bull., v. 44, p. 83-100.

Griffiths, J.C., 1966, Application of discriminant functions as a classification tool in the geosciences, in Computer applications in the earth sciences: Colloquium on classification procedures, D.F. Merriam, ed.: Kansas Geol. Survey Computer Contr. 7, p. 48-52.

Harman, H.H., 1960, Modern factor analysis: Univ. Chicago Press, Chicago, p. 445-462.

Imbrie, J., 1963, Factor and vector analysis programs for analyzing geologic data: Office of Naval Research, Geography Branch, Tech. Report No. 6, ONR Task No. 389-135, 83 p.

James, W.R., 1966, The Fourier series model in map analysis: Office of Naval Research, Geography Branch, Tech. Report No. 1, ONR Task No. 388-078, 37 p.

James, W.R., 1967, Nonlinear models for trend analysis in geology, in Computer applications in the earth sciences: Colloquium on trend analysis, D.F. Merriam and N. C. Cocke, eds.: Kansas Geol. Survey Computer Contr. 12, p. 26-30.

Krumbein, W.C., 1963, Confidence intervals on low-order polynomial trend surfaces: Jour. Geophysical Res., v. 68, p. 5869-5878.

Krumbein, W.C., 1966, Classification of map surfaces based on the structure of polynomial and Fourier coefficient matrices, in Computer applications in the earth sciences: Colloquium on classification procedures, D.F. Merriam, ed.: Kansas Geol. Survey Computer Contr. 7, p. 12-18.

Krumbein, W.C., and Graybill, F.A., 1965, An introduction to statistical models in geology: McGraw-Hill, New York, 475 p.

Merriam, D.F., and Sneath, P.H.A., 1966, Quantitative comparison of contour maps: Jour. Geophysical Res., v. 71, p. 1105-1115.

Miller, R.L., 1964, Comparison-analysis of trend maps: Computers in the Mineral Industries, Stanford Univ. Publ., Geol. Sci., v. 9, p. 669-685.

Miller, R.L., and Kahn, J.S., 1962, Statistical analysis in the geological sciences: John Wiley and Sons, Inc., New York, 483 p.

Pelto, C.R., 1954, Mapping of multicomponent systems: Jour. Geol., v. 62, p. 501-511.

Potter, P.E., 1962, Regional distribution patterns of Pennsylvanian sandstones in Illinois Basin: Am. Assoc. Petroleum Geologists Bull., v. 46, p. 1890-1911.

TREND-SURFACE ANALYSIS OF NOISY DATA

by

Donald B. McIntyre

Pomona College

INTRODUCTION

In a normal day's trading, the stock of a single company varies erratically in price. While it is true that this variation is sometimes obviously related to announcements of financial, political, or military actions, the short term fluctuations are usually dependent on a multitude of minor causes (in themselves of no lasting importance) that obscure the general pattern of the market. If a smooth curve is drawn to approximate the data, without necessarily passing through any individual point, the background noise is filtered out and the trend can be seen more clearly: price is then considered as a simple function of time alone.

In the more complex models of multivariate statistics, the dependent variable is treated as a function of more than one independent variable; and the case we are to consider is the mapping problem in which a measured quantity (y) varies with map coordinates (u, v), and possibly with elevation (w). First applications were to geophysical (gravity) maps of areas where the regional effect could be assumed to be given by

$$y' = a + b.u + c.v$$

This is the equation of a plane; and, because no powers higher than unity are involved, it is a first-degree trend surface. If y_i is an observed value and y_i' is the predicted value at the same locality, then the residual $(y_i - y_i')$ estimates the local effect; and this may be what is of interest. Thus, trend-surface analysis can separate two factors when it is known that one (the regional trend) is simple and the other is superimposed upon it. On the other hand, if the local effects are considered to be noise and hence of no immediate interest, then the trend surface -- like the economic trend -- is an expression of the meaningful part of the raw data. This is the aspect of trend-surface analysis that we are to consider here.

If in this sense the computed surface does represent the data, the map has been expressed as an equation; and this is of great importance as a partial solution to the problem of storing maps in digital form. Whether this is possible in practice depends on the complexity of the pattern, scale considered, noisiness and distribution of the data, and computing power available. It should be noted that the word trend implies a simple pattern, and one would expect a trend surface to display a general tendency for the values to change in particular directions. If the surface is a complicated one it loses this property; but the principle of fitting a smooth surface (not necessarily passing through the individual data points) is unchanged if the surface is made more complex, and the term trend surface is used even when no general tendency is evident.

LEAST-SQUARES CRITERION

The fitting of a trend surface is an extension of regression procedures from two (y, x) to three (y, u, v) or four (y, u, v, w) dimensions, and polynomial equations in the independent variables are normally used. The highest power to which u and v are raised is the degree of the polynomial; and, unless specifically stated otherwise it is assumed that all cross products are included. As in standard methods of interpolation, no theoretical meaning is sought in the form of the equation or in its individual terms. The justification is the practical utility of the results, and the preference for polynomials is based on the extension of the mean and the simple linear model, and on the straightforward algebra involved. Other forms can be developed for special cases, but it is doubtful whether an ad hoc approach is worth the trouble and computing costs that would normally be involved. If, as with chemical analyses, the data must lie within a given range, it is usually satisfactory to prescribe limits and, where the trend surface goes out of bounds, admit that the data are inadequate.

The criterion for best fit must be decided upon before the coefficients can be computed, and a common approach is as follows: consider each value (y_i) as consisting of a meaningful part $(y_i'$ -- the value on the trend surface) plus noise $(y_i - y_i' = \epsilon_i)$; and suppose that the noise is compounded of many small independent parts each of which has an equal probability of being positive or negative. Then the probability of an error between ϵ_i and $\epsilon_i + d.y$ is

$$P_i = \frac{k}{\sqrt{\pi}} \; e^{-k \; \epsilon_i^2} \; . \; dy$$

where

$$k^2 = 1/2 \; \sigma^2$$

Different surfaces give different predictions (y_i'), and it seems reasonable to accept as the best surface that one making the observed values (y_i), most probable.

The probability of η independent events all occurring together is the product of the individual probabilities; hence

$$P_{total} = \text{constant } e^{-k^2 \sum_{i}^{\eta} \epsilon_i^2}$$

which is a maximum when $\sum_{i}^{\eta}(y_i - y_i')^2$ is a minimum.

This is the least-squares criterion, and for its justification in terms of maximum likelihood, it is necessary that the data points be independent and that the residuals (noise) be normally distributed about the surface. But, if these conditions are insisted upon, we must exclude the attempt to separate regional and local effects; for, if the latter are meaningful, they will not be normally distributed. And, if the locations of data points are independent, they will be clustered, so that some important areas will be omitted from the sample. Moreover, this clustering may involve serial correlation (Southworth, 1960; Agterberg, 1965), and hence lack of independence, of the values measured. In my experience, it is necessary to supplement a random sample by points as strategic locations; for a bad distribution is likely to be a more serious consideration than is defense of the applicability of the principle of maximum likelihood. It should also be remembered that Gauss himself justified the least-squares method in a different way and without any assumptions concerning normality: namely, the coefficients are to be those unbiased estimates whose sampling variance is minimal (Plackett, 1949).

COMPUTATIONAL PROBLEMS

The steps involved in the computer program are as follows: (i) accumulation of the sums of powers of the coordinates, and of their crossproducts; (ii) arranging these values into a matrix; (iii) inverting the matrix; and (iv) solving for the coefficients. The first and last of these are trivial, but because the matrices are large intermediate steps are not. If all crossproducts are included, the number of coefficients is related to the degree of the polynomial surface as follows:

$$\eta_{u,v} = 1 + 3/2\, d + 1/2\, d^2$$

$$\eta_{u,v,w} = 1 + 11/6\, d + d^2 + 1/6\, d^3$$

The corresponding matrix has η^2 terms, and in double precision this means $2\eta^2$ computer words.

The instructions for building a large matrix may occupy more space in the memory of the computer than the matrix itself does; and the execution time may also be appreciable. The reason is that each element has to be individually placed; except that the matrix is symmetric, the pattern is not a simple one. Moreover, high-level languages, such as FORTRAN, are not designed for this purpose, and lack the ability to look far enough ahead to do an efficient job. For instance the value to be stored may already be in the accumulator, but the compiler does not take advantage of this, and a redundant instruction to load the accumulator is inserted.

The only way of solving this problem is to write an assembly language program that makes use of the features of the particular computing system. But such programs are very tedious to write. I have done this in practice by writing a preliminary program that computes all necessary addresses and generates the required source program. Unfortunately, such programs are dependent not only on the particular machine but on the particular operating system that controls it. Thus, on the IBM 7094 at Western Data Processing Center, I had to change from an older system in which arrays were stored downwards to a newer one in which they were stored upwards; and storage for double precision numbers changed from separate arrays for the more and for the less significant words to an interleaved mode that permitted use of double-load and double-store instructions. But the gain is great: e.g., a FORTRAN subroutine, which required 8371 36-bit words for the instructions to build a 45 by 45 double-precision matrix, was replaced by a FAP subroutine of only 1231 words. I am now using a general routine that writes programs for building any matrix in powers and cross products of u, v, w (in either single or double precision) for IBM System 360.

The other programming problem is the maintenance of significant figures during the inversion of a large matrix; and again the routines should be written in assembly language. I use a very efficient double-precision routine that occupies fewer than 500 bytes (125 words). But most important of all is to insure that the raw data are well scaled (for high powers will be generated), and that the matrix is further scaled by making use of the property that its elements tend to increase in size towards the bottom right-hand corner (Mandelbaum, 1963; McIntyre, 1963). When high-order surfaces are involved, it is difficult to maintain significant figures when short cuts of the form

$$\Sigma y^2 - a \Sigma x - b \Sigma xy$$

are used to compute the residual sum of squares. It is safer to compute this quantity from the definitional form; and this procedure also has the advantage that the individual residuals can be stored on tape for subsequent plotting or for testing the assumption that they form a normal distribution.

EVALUATION OF NOISE

The most common sources of noise in quantitative mapping arise from lack of precision in the

analytical or measurement technique and form defective sampling procedures. Sometimes, as at sea, the location is in error; but it can usually be assumed that the coordinates u, v, w are independent variables without appreciable error. Composition of ground water gives an example of dependent variables that may change rapidly with time; so that, if the survey of a basin takes several months or years to complete, the conditions will not have been constant, and a high noise level results. In this case it may be essential to plan the work so that the area is covered in a shorter time, even if this is at the expense of the number of sampled points.

Usually there is a hierarchy of nested variances that control the precision, and it is essential that the variances be analyzed so that it is known where the largest contributions arise (for examples see Baird, MacColl, and McIntyre, 1962; McIntyre, Welday, and Baird, 1965; Baird, McIntyre, and Welday, 1967). Because of the peculiar arithmetic involved by additivity of variances, one large standard deviation will dominate many smaller ones. The common dichotomy between laboratory and field work should be minimized; for noise can be generated at any level, and the anthropomorphic term error can be misleading when the cause is a natural variation in the material sampled.

If each station is represented by at least two samples, and it can be assumed that the variance on the scale of a collecting station is approximately constant, a pooled estimate of total variance within-stations can be determined. This can provide the standard necessary for judging the significance of variance between-stations. However, when we contour a map, we expect each data point to be more or less representative of an area much larger than a single collecting locality. It is as if each locality could be expanded until the whole is subdivided into polygonal domains with a single point in each. But we have no degrees of freedom for estimating the variance within these domains. A possible solution to this problem is offered if the sum of squares of the residuals appears to approach a constant value as the degree of the surface is increased; for it seems reasonable to suppose that most of the meaningful part of the data has then been incorporated in the trend surface, so that inclusion of further terms in the equation does little to improve the fit.

It is very likely that some of the terms in a high-order polynomial contribute little to the percentage sum of squares accounted for, and it is perhaps appropriate to examine the contributions of various coefficients. However, no coefficient can be deleted without an effect on all the others; and the computational time required for the permutations of a complete stepwise regression seems unjustified -- especially because we are not seeking physical significance in the individual coefficients (see, however Mandelbaum, 1963). On the other hand, if we carry redundant coefficients, we cut down the number of degrees of freedom; for even when a coefficient is known a priori to be zero, if an estimate is computed from the data then the number of degrees of freedom must be reduced. Perhaps this helps to offset an overestimate of degrees of freedom resulting from a nonideal distribution of data points. As an extreme example, if 100 coincident or collinear points are used to fit a plane, we will not have 97 degrees of freedom for its attitude.

Although the least-squares criterion does not require any assumption of normality, tests of significance usually do. Fortunately it seems that most tests are robust, i.e. departure from normality is not critical. What may be more important is possible correlation where independence is assumed, and it is difficult to test this. Formal evaluation of trend surfaces invites a variety of approaches, and this is an area of current research interest (see for example, Krumbein, 1963; Krumbein and Graybill, 1965; Merriam and Sneath, 1966a, 1966b; Mandelbaum, 1966; Baird, McIntyre, and Welday, 1967). However, even in the ideal case -- where the errors all have zero mean, are uncorrelated, and are normally distributed -- it is easy to show that a high-degree surface can give a significant improvement when lower degree surfaces do not; the plot of residual variance against complexity of surface is not necessarily a smooth curve.

AUTOMATIC CONTOURING

Closely connected with the computation of trend surfaces is the problem of automatic contouring -- particularly when the data are randomly distributed and noisy. In approaching this problem, it is important to recognize that a plotter can only be directed to draw straight lines, and consequently the map must be divided into domains in which the surface is taken as planar. I do this by constructing a square grid and dividing each square into four triangles. The contours within any triangle can then be determined as straight lines, and the pattern of contours will be consistent over the whole area. Rectangular, hexagonal, or curved grids can be used, but I have found no occasion to prefer these patterns to the square grid, which is computationally convenient.

If a trend surface is to be contoured, the grid points can be computed from the equation of the surface; otherwise the first problem is to construct the grid. Obviously, a grid point should be determined from the data points nearest to it, and these can be found by drawing a circle around the grid point and taking all data points that fall within it. But this step must be carefully programmed, or else the time required for execution will be excessive. Some preliminary ordering of the data is advisable. If too few points are found within the circle, either the circle must be enlarged or the attempt to assign a value to

that grid point must be abandoned. So there are many options to choose from, and few theoretical principles to guide the choice.

There is some minimum number of data points on which any grid point should be based, and it seems reasonable to use a least-squares fit to these data points. If a second-degree surface is fitted, that minimum number is 6; but in practice, because an exact fit is not desirable when data are noisy or badly distributed, a greater number must be insisted upon. I have found 8 to be a satisfactory number in hundreds of maps, mainly of gravity anomalies, water quality determinations, and rock chemistry. If the minimum number is made too large, then the area around the grid point can often no longer be represented by such a simple surface, and consequently results deteriorate. The optimum number depends on the density of data relative to complexity of the surface. In a gravity study of part of the Los Angeles basin, a contoured map based on 18 stations was almost the same as one of the same area based on 282 stations; in this case the pattern was simple and a small number of points sufficed to define it. This situation is exceptional, and in most of my work I find the data to be noisy and badly distributed. Alignment and clustering of data points are sometimes very serious, particularly with oceanographic and mining data, and it should be kept in mind that a high density of data along one line cannot make up for absence of data off the line. For instance, one oil company complained that the trend-surface program would not work; but on investigation it emerged that the data used consisted of collinear points.

It is very wasteful to permit highly discrepant grid points to be used for contouring, and yet it is expensive to test the selected data points for orientation. If a value calculated at a grid point lies inside the range of the data, or not more than $\pm$ 20% beyond that range, I include it; otherwise it is arbitrarily rejected. This criterion works well in practice. Data near the grid point should count for more in estimating it than data farther away, and consequently I weight the data as a function of their distances from the grid point. The weighting factor $1/d^2$ works well but I have tried $1/d$ and $1/d^3$ and the results are not very different. This of course, depends upon the data.

In order to test the procedures that have been described in the preceding paragraphs, a contour map was drawn freehand (Fig. 1) and attempts were made to reconstruct this map from sample data drawn from it. Obviously the relative success of the tests is considerably dependent on the nature of the surface to be sampled; and, if the surface is made more complex, a greater density of sample points would be required. To define 100 random points on the map, 200 random numbers were used and data values were assigned to these points by visual inspection and interpolation. The contour map (Fig. 2) generated from these data seems to be a very satisfactory representation of the original surface. For a second test, the first 50 and the second 50 of these random samples were used separately (Fig. 3, 4). As would be expected, the resolution is poorer. But the closure of contours on the west side of the map is remarkable when the distribution of the sample points is taken into account. The eighth-degree surface based on the 100 random points is given in Figure 5. To minimize marginal effects, the map was extended beyond its original bounds and new data, based on a grid (squares of 1% area), were used for contouring (Fig. 6) and for computation of an eighth-degree trend surface (Fig. 7). Such exceptionally good data would justify a surface of even higher degree.

The polygonization that can be detected on these maps is due to the use of an open grid, in this case with squares of area 1% of the whole. If this feature is undesirable, or if the data points are to be more clearly honored, a finer grid can be constructed; but this will cost more in computing time, and spurious wiggles will be introduced. An alternative is to use an analog rather than a digital plotter. Filtering of the data is possible when the magnetic tape with plotting information is being read, and the momentum of the plotter arm can be used to advantage. I have experimented with this technique, but I normally use a digital plotter because of availability and cheaper price. The polygonization is likely to be no more than a psychological disadvantage if data are noisy.

If an analog plotter is to be used to smooth the contours, it is necessary that the line segments be ordered so that one contour will be continuously followed across the map. And, even with a digital plotter such optimization is an important factor, because it costs as much to move the plotting head with the pen up as with the pen down. In planning an economic system for automatic plotting, it is important to consider the relative costs of computer and plotter time when balancing the sizes of the various buffers that are required. I have used both off-line and on-line plotters, as well as computers at different centers (including teleprocessing); and these different configurations call for very different systems The economic importance of systems analysis is often overlooked.

The method of automatic contouring described here results in a map that can be considered as a moving average low-degree trend surface. It has proved to have serious disadvantages when used with badly distributed, noisy data. Ground-water data tends to be of this sort, particularly when a considerable time elapses before the sampling is completed. Neighboring wells may be recorded with very different readings. Contouring by hand then becomes very subjective, and automatic contouring can give plots that are almost useless. But even in such a case, good results can be obtained by contouring a trend surface fitted to the data as a whole. In many cases I have found this to be the only way of extracting meaningful information from noisy data. When one

is looking for a pattern in the residuals from a low-degree surface, the signal to noise ratio is often low and contouring correspondingly difficult. A useful technique is then to fit a high-degree surface to these residuals; and, because the residuals are written on magnetic tape, it is very simple to read them back for this purpose.

It has been suggested (Dodd, Cain, and Bugh, 1965) that apparently significant contour patterns can be demonstrated with random data. My experience is that only highly subjective contouring is possible; the automatic contouring program rejects such data as too noisy to process; and trend surfaces account for only a small percentage of the total sum of squares.

SIMULATION OF SAMPLING

Trend-surface analysis combined with automatic contouring provides a method for testing various sampling schemes by simulation. The geologist can sketch a contour map of his preferred model, preferably one based on a pilot study; a high-degree trend surface can then be fitted and contoured; and this surface becomes the test model. Various sampling schemes are used to extract data from the model, and results are plotted for comparison with the original. Costs of the different schemes can be weighed against the success of results. In the program that I have developed for this purpose, random errors from several normal populations can be introduced at different levels to simulate actual analytical and field variances.

Studies of this sort suggest that a random distribution of sampling localities may be better than grid data under certain conditions. A particularly interesting case was an attempt to digitize terrain as recorded on a topographic map: random sampling gave better results than either a grid or a selection of points made by a skilled cartographer. This is a topic that deserves further investigation.

Acknowledgments.- Some of the programs described here were developed as part of a project supported by NSF grant GA-686, and I have benefited from discussions with A.K. Barid and E.E. Welday throughout the work. David D. Pollard, Roger Smith and Robert M. Viney made valuable contributions while working under NSF Undergraduate Research Participation grants GE-2804, GE-6160 and GE-8187. Computer time on the IBM 7094 was generously provided by Western Data Processing Center, Graduate School of Business Administration, University of California at Los Angeles. The IBM System 360 programming was tested at the Pomona College Computer Center. Wesley Folsom, Roy Graves, and Dianne Willis of IBM made helpful contributions. Some of the plotter time was paid for by a grant from the Shell Assists Fund, Shell Companies Foundation.

REFERENCES

Agterberg, F.P., 1965, The technique of serial correlation applied to continuous series of element concentration values in homogeneous rocks: Jour. Geology, v. 73, p. 142-154.

Baird, A.K., MacColl, R.S., and McIntyre, D.B., 1962, A test of the precision and sources of error in quantitative analysis of light, major elements in granitic rocks by X-ray spectrography: Advances in X-ray analysis, Plenum Press, v. 5, p. 412-422.

Baird, A.K., McIntyre, D.B., and Welday, E.E., 1967, Geochemical and structural studies in batholithic rocks of southern California: Part II. Sampling of the Rattlesnake Mountain pluton for chemical composition, variability, and trend analysis: Geol. Soc. America Bull., v. 78, p. 191-222.

Dodd, J.R., Cain, J.A., and Bugh, J.E., 1965, Apparently significant contour patterns demonstrated with random data: Jour. Geol. Ed., v. 13, p. 109-112.

Krumbein, W.C., 1963, Confidence intervals on low-order polynomial trend surfaces: Jour. Geophysical Res., v. 68, p. 5869-5878.

Krumbein, W.C., and Graybill, F.A., 1965, An introduction to statistical models in geology: McGraw-Hill, Inc., New York, 475 p.

Mandelbaum, H., 1963, Statistical and geological implications of trend mapping with nonorthogonal polynomials: Jour. Geophysical Res., v. 68, p. 505-519.

Mandelbaum, H., 1966, Comments on paper by Daniel F. Merriam and Peter H.A. Sneath, 'Quantitative comparison of contour maps': Jour. Geophysical Res., v. 71, p. 4431-4432.

McIntyre, D.B., 1963, Program for computation of trend surfaces and residuals of degree 1 through 8: Dept. Geology, Pomona College, Tech. Rept. 4, 24 p.

McIntyre, D.B., Welday, E.E., and Baird, A.K., 1965, Geologic application of the air pycnometer: a study of the precision of measurement: Geol. Soc. America Bull., v. 76, p. 1055-1060.

Merriam, D.F., and Sneath, P.H.A., 1966a, Quantitative comparison of contour maps: Jour. Geophysical Res., v. 71, p. 1105-1115.

Merriam, D.F., and Sneath, P.H.A., 1966b, Reply to discussion: Jour. Geophysical Res., v. 71, p. 4433.

Plackett, R.L., 1949, A historical note on the method of least squares: Biometrika, v. 36, p. 458-460.

Southworth, R.W., 1960, Autocorrelation and spectral analysis, in Mathematical methods for digital computers, Ralston, A. and Wilf, H.S., eds.: John Wiley and Sons, Inc., New York, 293 p.

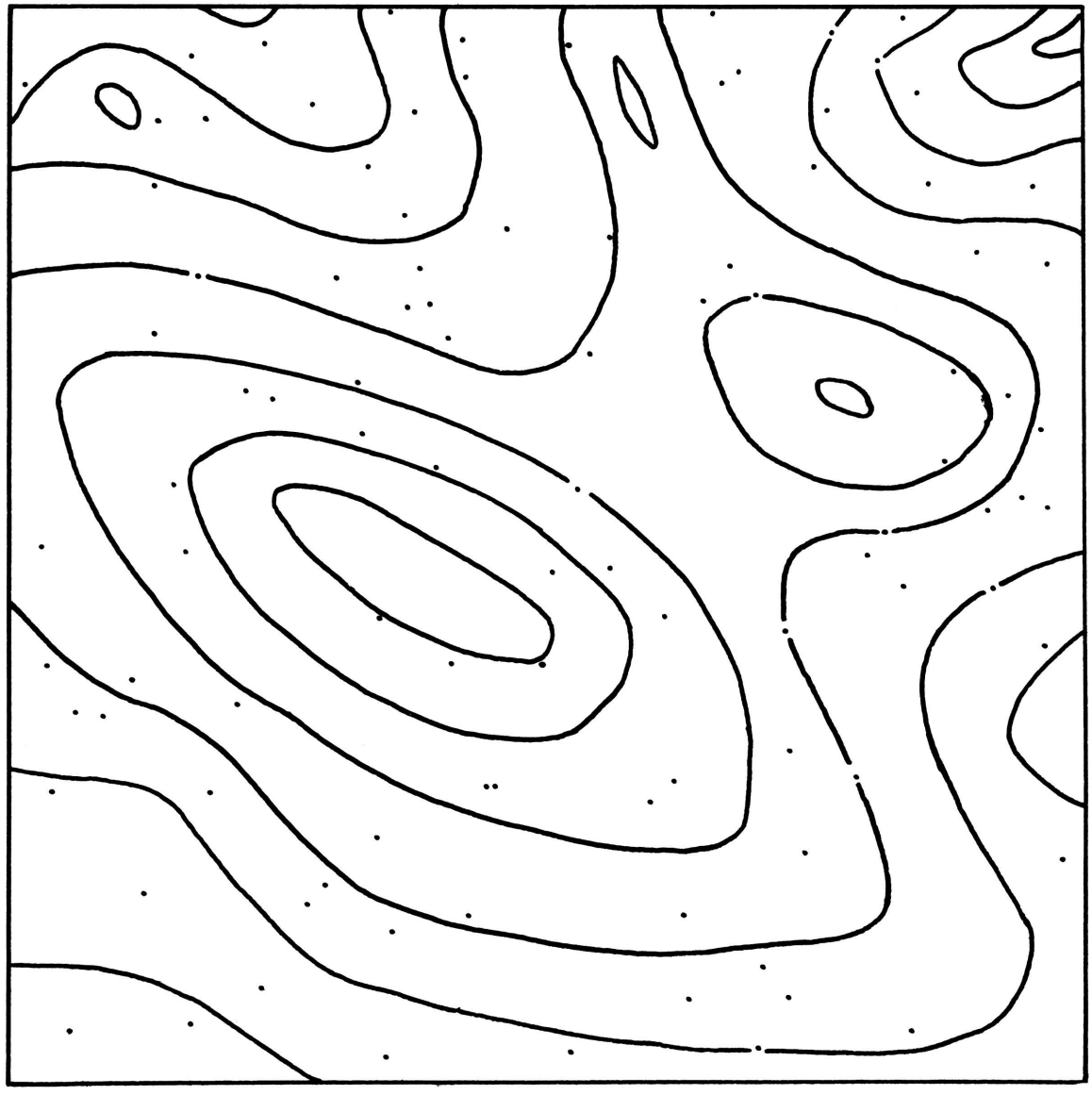


Figure 1.- Test pattern drawn by hand. Locations of 100 random points are shown. Values assigned at these locations were based on visual interpolation from contours.

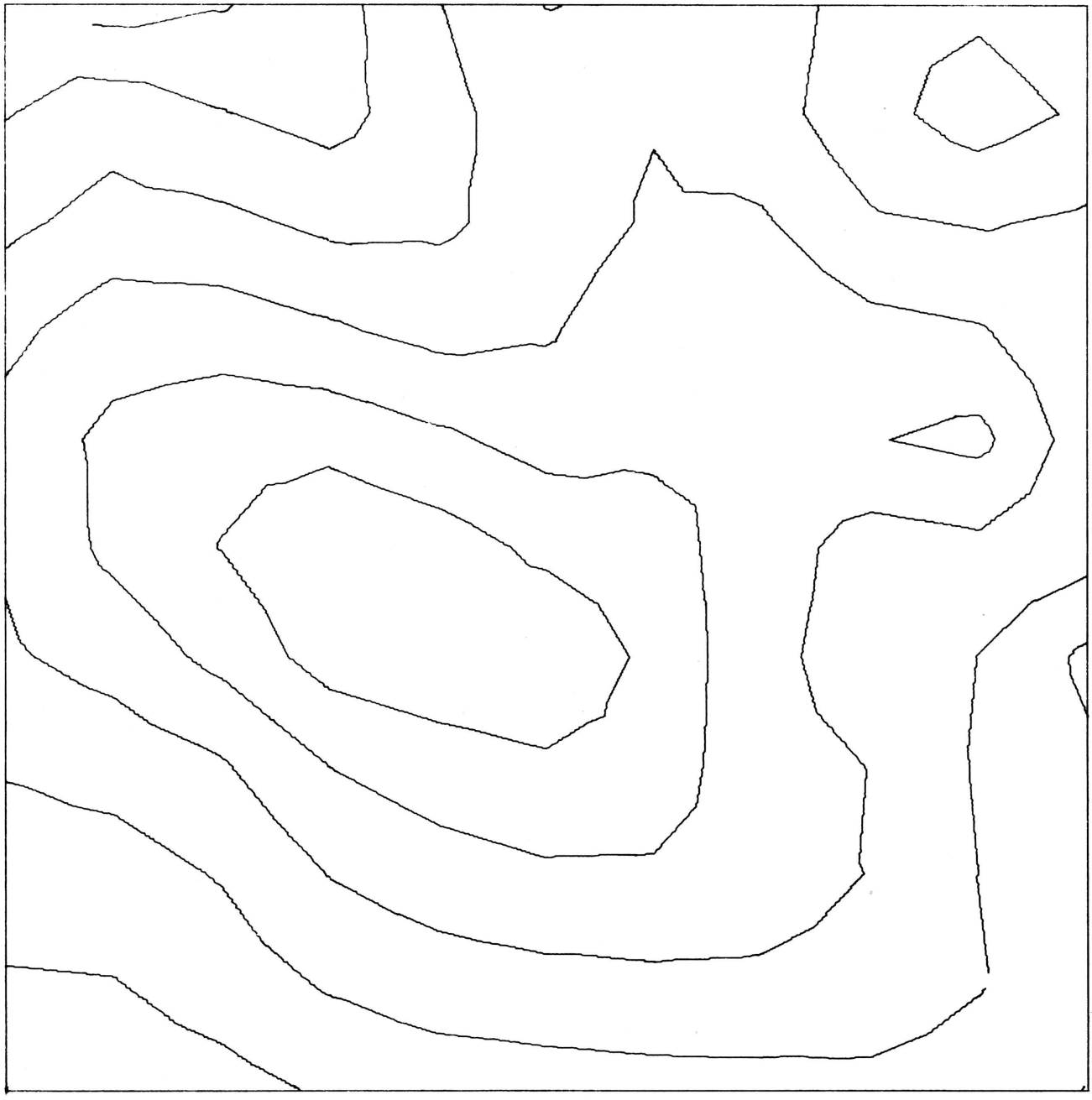


Figure 2. - Automatic contouring of values at 100 random points of Figure 1. Grid squares are 1% of total area, and each grid point is based on a minimum of 8 data points.

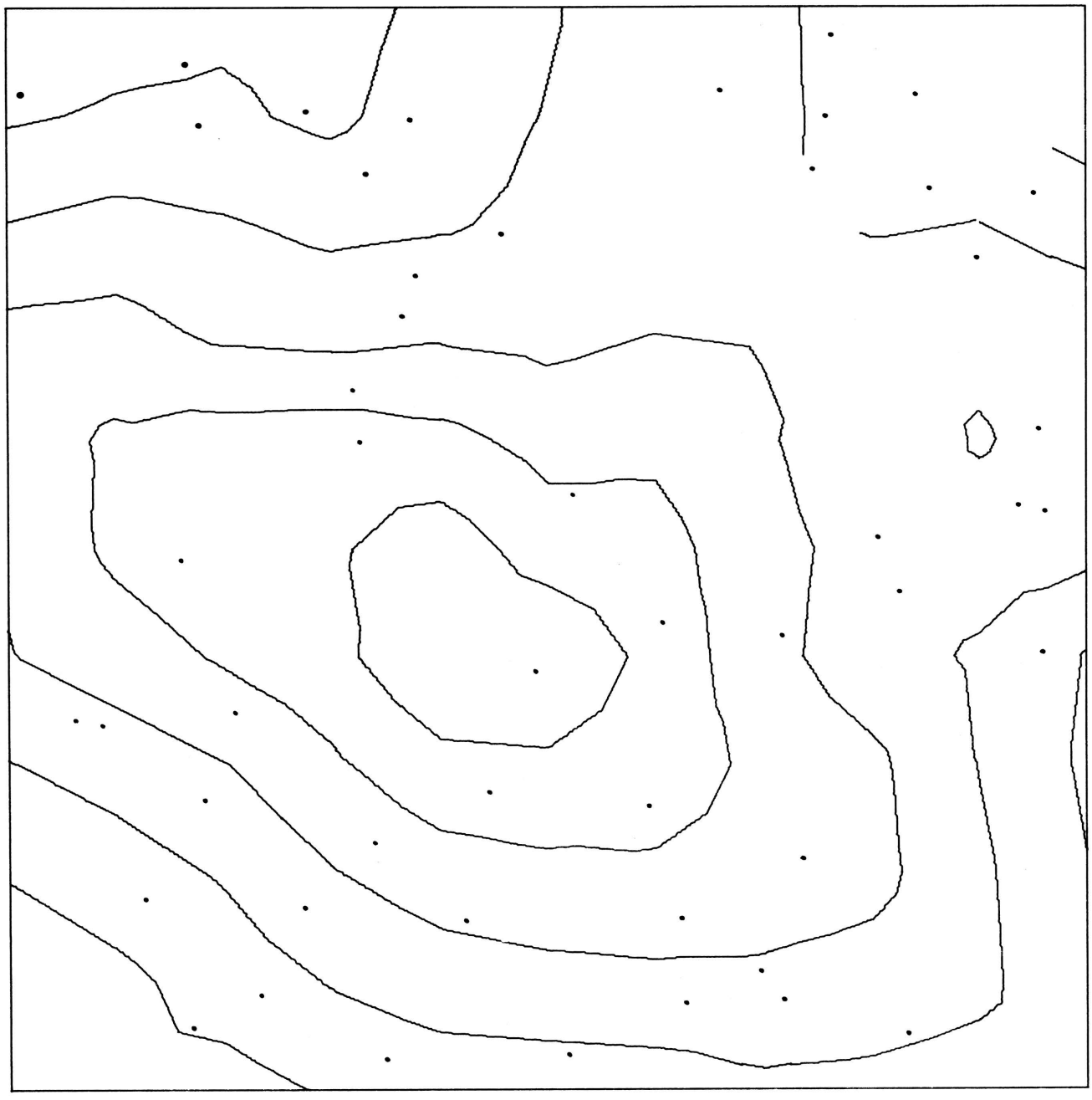


Figure 3.- Automatic contouring of values at first 50 random points of Figure 1. Conditions for constructing grid are same as in Figure 2.

Figure 4.- Automatic contouring of values at second 50 random points of Figure 1.

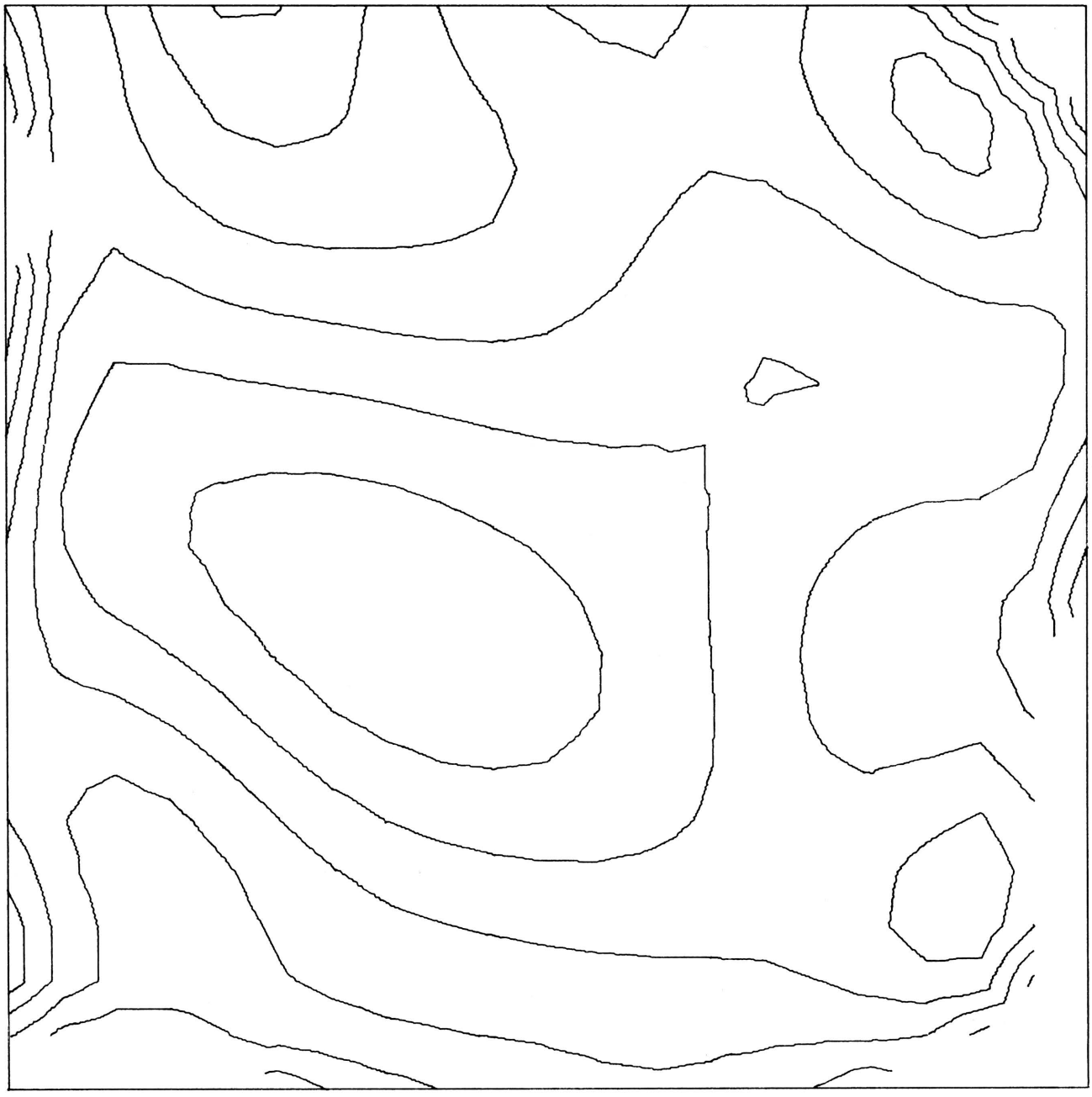


Figure 5.- Eighth-degree trend surface of same data contoured in Figure 2. Percent sum of squares accounted for by surface is 97.7.

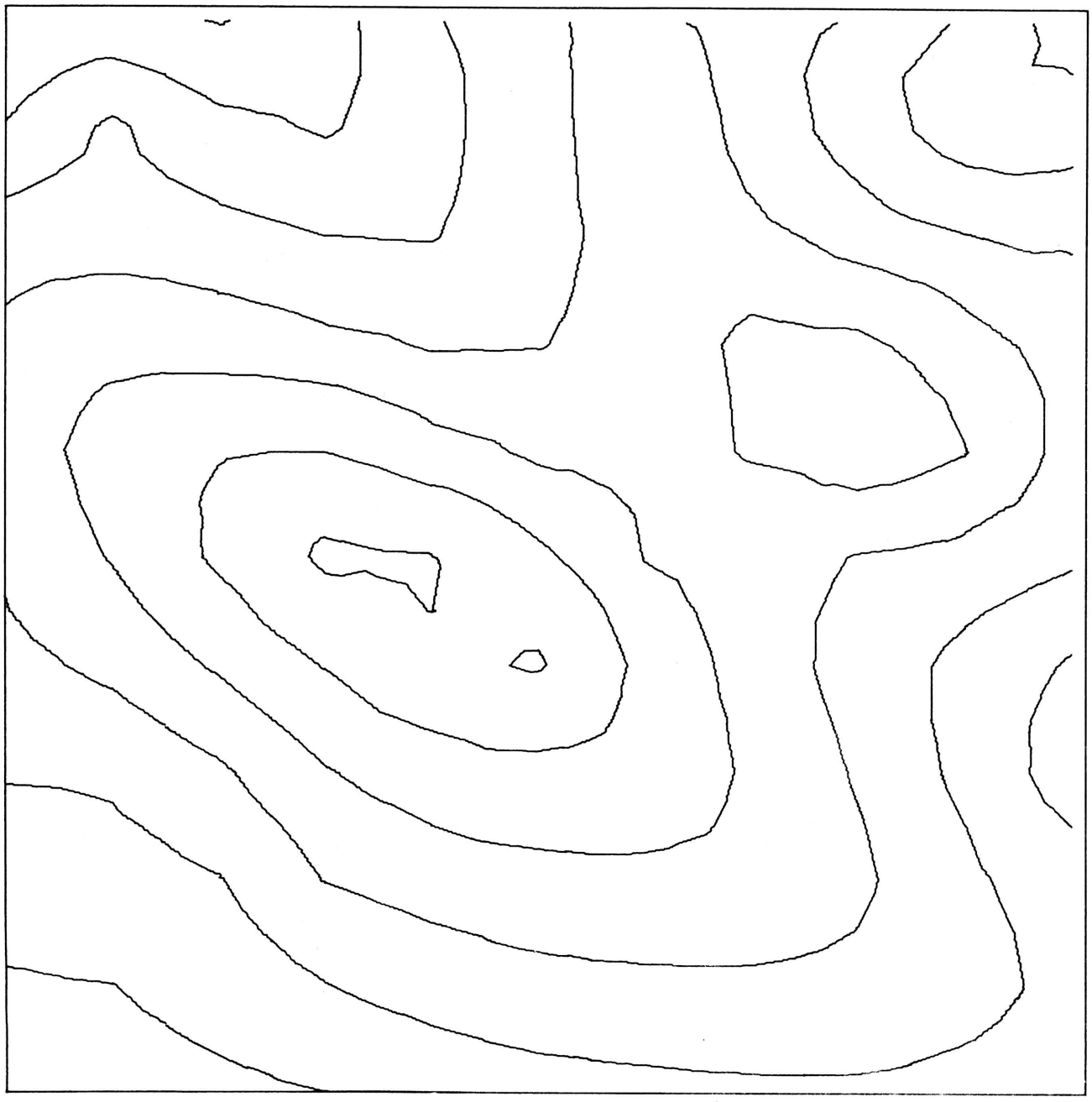


Figure 6.- Pattern of Figure 1 was extended beyond bounds of original, and values were assigned by eye to intersections on grid of same mesh as used in Figure 2. This provides control around the edges; 169 points were automatically contoured, and this map is result for same area as shown in other figures.

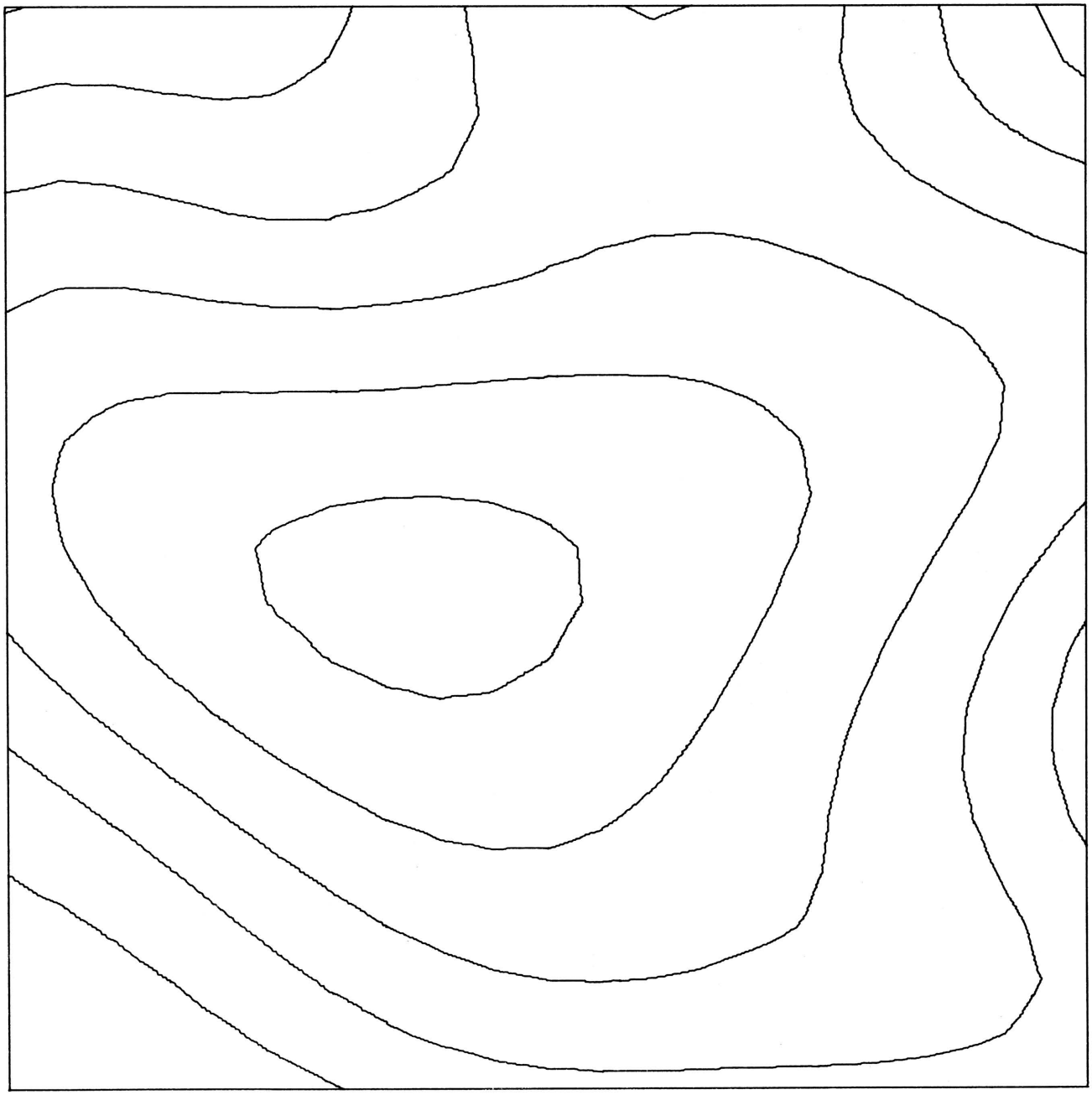


Figure 7.- Eighth-degree trend surface of same data contoured in Figure 5. Percent sum of squares accounted for by surface is 94.5 .

APPLICATION OF RESPONSE-SURFACE ANALYSIS TO SEDIMENTARY PETROLOGY

by

John C. Davis

Kansas Geological Survey

Factors controlling mineral distribution patterns in sedimentary rocks are poorly understood, in part because of a general lack of raw data. Few studies have been made on the petrology of sedimentary units which involve enough samples for adequate descriptions of the units. Stratigraphers and petrographers have been content to characterize units comprising thousands of cubic miles from a small collection of samples gathered in a haphazard manner along a narrow strip of outcrop. This situation is unavoidable to a certain extent. Samples must be taken where there are exposures or drill holes, and there is a limit to the number of specimens that can be justifiably analyzed in any study. Nevertheless, geologists have been notoriously naive about the statistical validity of most of their efforts.

Recent advances in analytical technology have greatly increased rates at which samples may be processed. It is now practical to make detailed three-dimensional studies of rock units, involving analysis of hundreds or even thousands of samples. Unfortunately, this wealth of information is lost in many cases because investigators are unable to impart data efficiently. A common practice for handling sample information from measured sections is to compute average values for each section, plot these values on a map, and contour the averages. Vertical variation, which may be more significant than lateral variation, is completely lost.

In many cases, three-dimensional polynomial response surfaces are effective for producing summary representations of spatially distributed data. They have been applied to studies of igneous plutons by several workers (Peikert, 1962, 1963, 1965; Whitten and Boyer, 1964) and to hydrocarbon distributions in sedimentary rocks (Harbaugh, 1964; Smith and Harbaugh, 1966). However, they have not been widely applied to other stratigraphic problems.

Polynomial response surfaces are power-function regressions in which independent variables are geographic coordinates. Values of the regressed, or predicted, variable form the response surface, which may be expressed as a contoured map in the case of two geographic variables, or a block diagram when three geographic variables are used. In the latter case, a solid enclosed by equal-value surfaces represents the functional approximation of the regressed variable. Harbaugh (1964) refers to these as "hypersurfaces." In my investigations, I have called these isopleth envelopes, an extension of a term first used by Krumbein and Pettijohn (1938).

Use of a polynomial expansion as an approximating function does not imply any belief that sedimentary components are distributed according to a power-function law. The power series is simply the easiest of a number of approximating functions to program and utilize in three dimensions. It seems premature to attempt to fit simulation models to composition data from sedimentary rocks, because not enough is presently known about spatial relations in sedimentary bodies to quantitize controlling factors.

RESPONSE SURFACE ANALYSIS OF THE MOWRY SHALE

Polynomial response surfaces were used to analyze mineral distribution in the Mowry Shale of Wyoming. The Mowry is a black, highly siliceous fine-grained Lower Cretaceous unit. Because of the structural pattern of Wyoming, the formation could be sampled at many localities which are widely distributed geographically. Over 300 specimens were analyzed by x-ray diffraction and other methods for four components: quartz, feldspar, analcite, and cristobalite. Approximately 500 additional analyses of organic carbon from the Mowry were obtained from Gulf Research and Development Company. Clays were not quantitatively measured because the highly silicic shale could not be adequately dispersed. No other constituents are present in significant quantities.

Four problems were involved in the analysis and presentation of this data:

(1) Representing the distribution of a single component through space.

(2) Comparing the distribution pattern of one component with another.

(3) Integrating organic carbon analyses into the study even though carbon samples were collected at localities different from those used for the rest of the investigation.

(4) Measuring the amount of variation within components and computing "goodness" of representation of the distribution.

These problems were solved to varying degrees of satisfaction by utilizing an experimental response-

surface program being developed at the Kansas Geological Survey. This program will compute regressions having up to 35 terms, automatically generate polynomial expansions up to 35 terms from one to seven original independent variables, compute error measures and related statistics, and produce slice-maps or graphs through any specified level. Any coefficient may be deleted at will and the loss in goodness-of-fit computed. Using this feature, a "best" regression equation may be found in a manner similar to the backward elimination procedure suggested by Draper and Smith (1966). The procedure requires the following steps:

(1) The maximum polynomial regression is computed. With our program, the fourth order is the highest that can be computed for three independent variables.

(2) Each variable is successively eliminated from the regression, but restored before elimination of the next variable. The significance level of each loss in regression sum of squares is computed for a fixed value of $F_{1,\ n-m-1}$.

(3) Least significant variables are eliminated and the regression recomputed. If the loss in goodness-of-fit is not significant, additional variables with low significance may also be eliminated.

In practice, step 2 can be simplified by examining the standard partial regression coefficients and considering only those variables having low values. Choosing the "least significant variables" in step 2 requires subjective judgment; in this example, individual variables having significance levels below about 70% would not produce a significant loss of fit in step 3. Variables having significance levels of about 80% produced unacceptable losses in fit when included in the deleted group.

Table 1 is an abbreviated ANOVA for regressions fitted to organic carbon data from the Mowry Shale. Computer-printed lists of deviations from the final regression were examined for trends. None were apparent, indicating that residual variation probably could not be reduced significantly even if larger regression programs were available. It has been suggested that deviations could be regressed in an attempt to determine if significant trends exist within them, but this approach has not been tested.

The response surface corresponding to the "best" regression in the ANOVA table can be shown as an isopleth envelope, or as a series of slice-maps through the sample space (Fig. 3). The latter illustration is useful for showing details, whereas the generalized block diagram shows overall relationships more clearly. For comparison, Figure 2 is an isopleth block diagram of the distribution of quartz in the Mowry Shale and Figure 4 is the corresponding series of slice-maps. Isopleth envelopes on the block diagrams enclose areas having above average concentrations of constituents. The distribution pattern of quartz can be compared visually to organic carbon distribution, even though raw data were not derived from the same sampling localities.

One of the more pleasing aspects of this example is that the response surfaces can be interpreted geologically. The lobate low extending from the northwestern corner of Wyoming into the carbon isopleth diagram corresponds to a tongue of tuffaceous sandstone in the upper Mowry. The generally arcuate patterns of the two isopleths reflect influence of a transgressing clastic unit which overlies the Mowry and is in part laterally equivalent. The main bodies of the two isopleth envelopes coincide; rocks in this area contain abundant fine quartz silt and large numbers of radiolarian tests and algal (?) spores. High carbon and quartz values apparently reflect specialized ecologic conditions that prevailed through this region. In southeastern Wyoming, carbon content increases and quartz decreases because the sediments are increasingly finer grained and richer in clay in this direction.

Polynomial response surfaces have proven useful in this study and provide adequate, easily represented discriptions of distribution patterns. They are less successful for comparison on one distribution to another; true multivariate techniques, which consider many dependent variables simultaneously, are needed. We are experimenting with regressions of variables condensed by principal components analysis, and with covariance analysis, as possible approaches to this problem. Response surfaces provide measures of variability, but as already noted, examination of residuals needs further development.

Acknowledgments.-Organic carbon analyses from the Mowry Shale were provided by Dr. Grover J. Schrayer, Gulf Research and Development Company, Pittsburgh, Pennsylvania. The experimental polynomial regression program is being developed by Paul Smith and Jim Esler of the Kansas Geological Survey.

REFERENCES

Draper, N.R., and Smith, H., 1966, Applied regression analysis: John Wiley and Sons, Inc., New York, 407 p.

Harbaugh, J.W., 1964, A computer method for four-variable trend analysis illustrated by a study of oil-gravity variations in southeastern Kansas: Kansas Geol. Survey Bull. 171, 58 p.

Krumbein, W.C., and Pettijohn, F.J., 1938, Manual of sedimentary petrography: Appleton-Century-Crofts, Inc., New York, 549 p.

Peikert, E.W., 1962, Three-dimensional specific gravity variation in the Glen Alpine stock, Sierra Nevada, California: Geol. Soc. America Bull., v. 73, p. 1437-1442.

Peikert, E.W., 1963, IBM 7090 program for least-squares analysis of three-dimensional geological and geophysical observations: Office of Naval Research, Geography Branch, Tech. Report No. 4, ONR Task No. 389-135, 71 p.

Peikert, E.W., 1965, Model for three-dimensional mineralogical variation in granitic plutons based on the Glen Alpine stock, Sierra Nevada, California: Geol. Soc. America Bull., v. 76, p. 331-348.

Smith, J.W., and Harbaugh, J.W., 1966, Stratigraphic and geographic variation of shale-oil specific gravity from Colorado's Green River Formation: U.S. Bur. Mines Rept. Invest. 6883, 11 p.

Whitten, E.H.T., and Boyer, R.E., 1964, Process-response models based on heavy-mineral content of the San Isabel Granite, Colorado: Geol. Soc. America Bull., v. 75, p. 841-862.

Table 1.-Abbreviated ANOVA table for regressions on organic carbon data from the Mowry Shale.

3rd Order (20 terms)		R = .68
$SS_{regression}$	113.15	F = 16.5**
$SS_{deviation}$	131.18	
4th Order (35 terms)		R = .70
$SS_{regression}$	123.82	F = 11.0**
$SS_{deviation}$	120.51	
Significance of 4th over 3rd order (15 terms)		
$SS_{regression}$	10.67	F = 2.15**
4th order without 11 nonsignificant variables [1] (24 terms)		R = .70
$SS_{regression}$	122.47	F = 16.12**
$SS_{deviation}$	121.86	
Significance of deleted variables (11 terms)		
$SS_{regression}$	1.35	$F = .36^{ns}$
Total Sum of Squares ...	244.33	

[1]Deleted variables are X_3, X_1X_3, $X_1^2X_3$, X_2^3, $X_2X_3^2$, $X_1^3X_3$, $X_1X_2^3$, $X_2X_3^3$, X_3^4, $X_1^2X_2X_3$, $X_2^2X_3^2$, where X_1 is east-west dimension, X_2 is north-south dimension, X_3 is depth below top of formation.

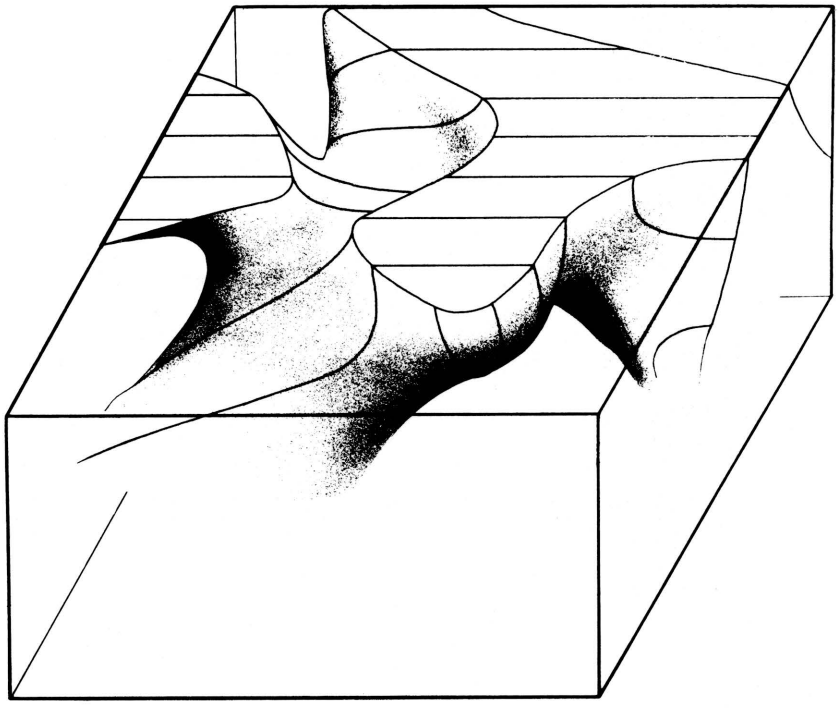


Figure 1.- Isopleth block diagram showing distribution of organic carbon in Mowry Shale. Isopleth envelope encloses areas having > 2% carbon. Thickness of block is approximately 500 feet. Horizontal limits of block coincide with Wyoming state boundaries.

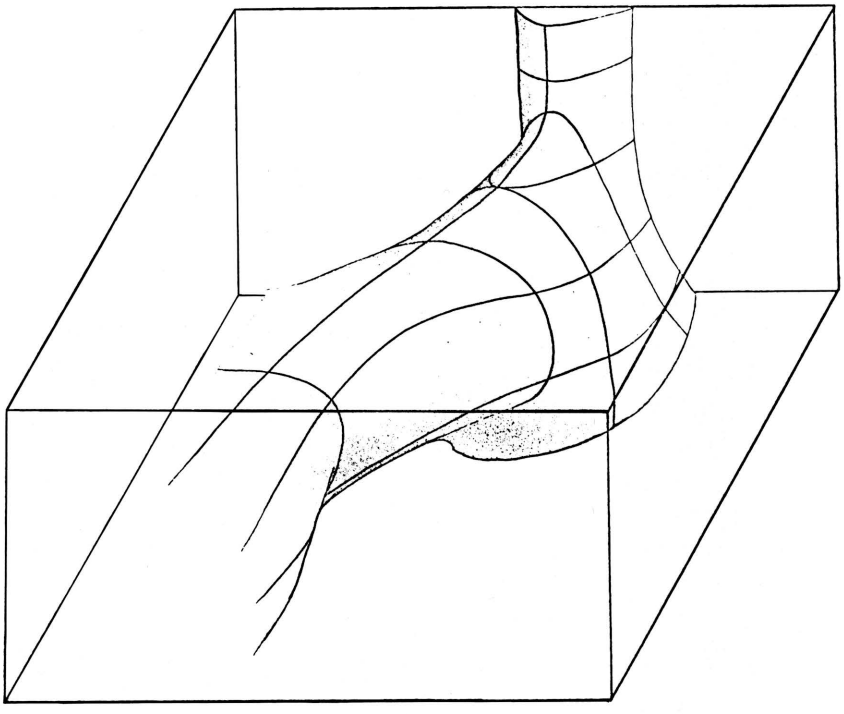


Figure 2.- Isopleth block diagram showing quartz distribution in Mowry Shale of Wyoming. Isopleth envelope encloses areas having > 50% quartz. Dimensions of block are same as Figure 1.

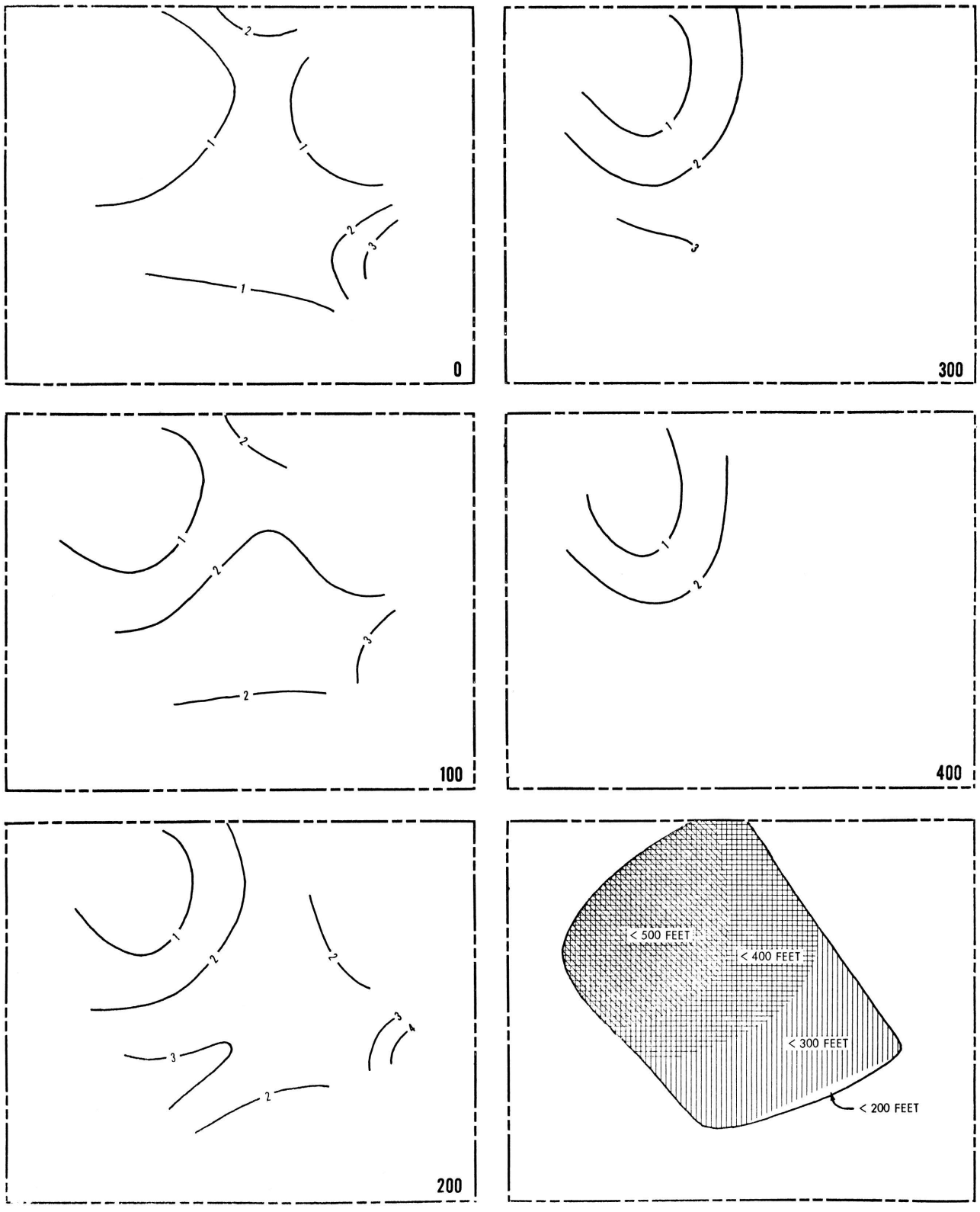


Figure 3.-Slice-maps at successive levels through Mowry Shale, showing organic carbon distribution in percent. Large numbers indicate level of slice-map in feet below top of Mowry. Map at lower right shows limits of control. Margins of maps coincide with Wyoming boundaries.

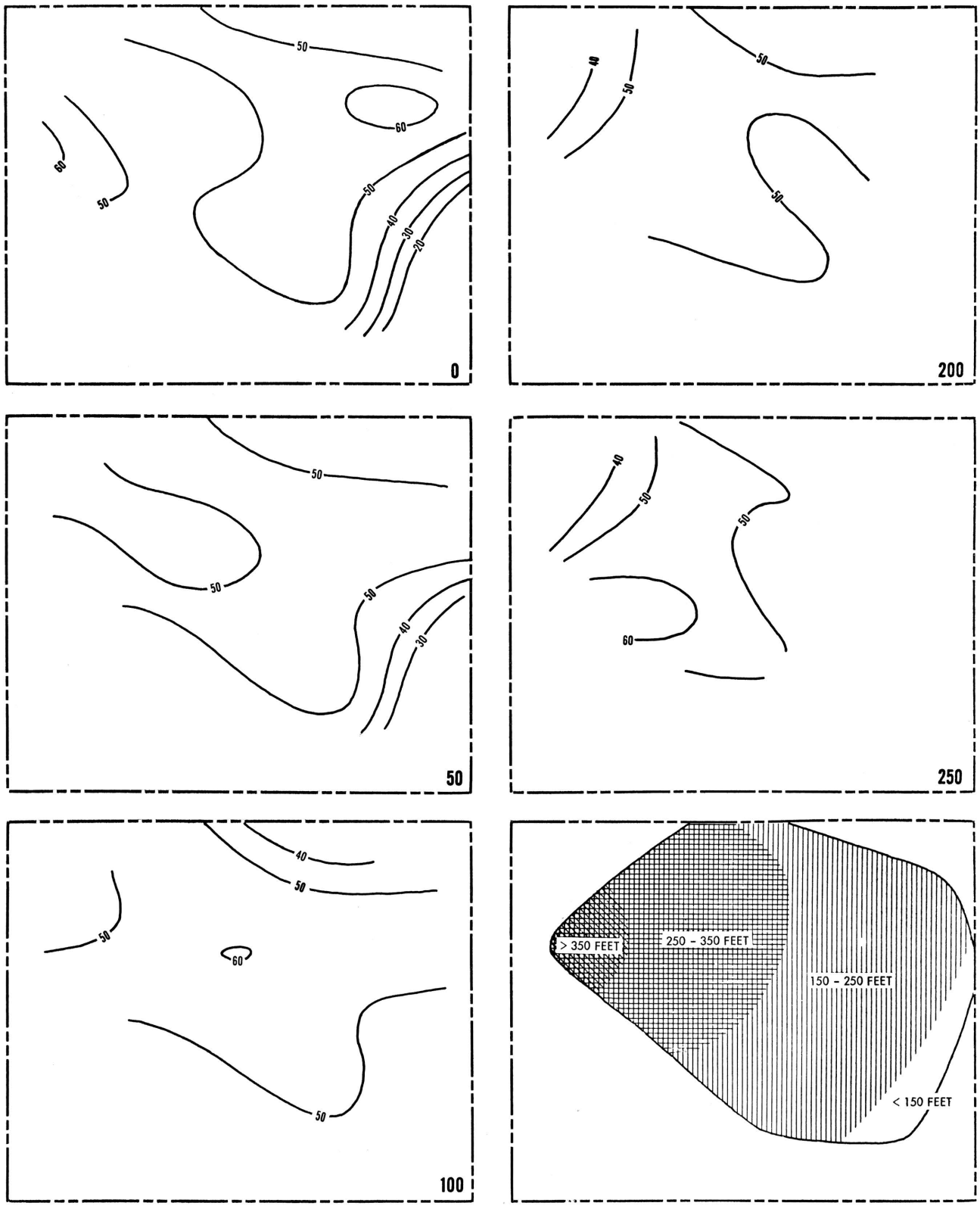


Figure 4.-Slice-maps at successive levels through Mowry Shale, showing quartz distribution in percent. See Figure 3 for explanation of diagrams.

COMPUTER CONTRIBUTIONS

Kansas Geological Survey
University of Kansas
Lawrence, Kansas

Computer Contribution

1. Mathematical simulation of marine sedimentation with IBM 7090/7094 computers, by J.W. Harbaugh, 1966. . . . . $1.00
2. A generalized two-dimensional regression procedure, by J.R. Dempsey, 1966 . . . . . $0.50
3. FORTRAN IV and MAP program for computation and plotting of trend surfaces for degrees 1 through 6, by Mont O'Leary, R.H. Lippert, and O.T. Spitz, 1966 . . . . . $0.75
4. FORTRAN II program for multivariate discriminant analysis using an IBM 1620 computer, by J.C. Davis and R.J. Sampson, 1966 . . . . . $0.50
5. FORTRAN IV program using double Fourier series for surface fitting of irregularly spaced data, by W.R. James, 1966 . . . . . $0.75
6. FORTRAN IV program for estimation of cladistic relationships using the IBM 7040, by R.L. Bartcher, 1966 . . . . . $1.00
7. Computer applications in the earth sciences: Colloquium on classification procedures, edited by D.F. Merriam, 1966 . . . . . $1.00
8. Prediction of the performance of a solution gas drive reservoir by Muskat's Equation, by Apolonio Baca, 1967 . . . . . $1.00
9. FORTRAN IV program for mathematical simulation of marine sedimentation with IBM 7040 or 7094 computers, by J.W. Harbaugh and W.J. Wahlstedt, 1967 . . . . . $1.00
10. Three-dimensional response surface program in FORTRAN II for the IBM 1620 computer, by R.J. Sampson and J.C. Davis, 1967. . . . . $0.75
11. FORTRAN IV program for vector trend analyses of directional data, by W.T. Fox, 1967 . . $1.00
12. Computer applications in the earth sciences: Colloquium on trend analysis, edited by D.F. Merriam and N.C. Cocke, 1967 . . . . . $1.00

Reprints (available upon request)

Finding the ideal cyclothem, by W.C. Pearn (reprinted from Symposium on cyclic sedimentation, D.F. Merriam, editor, Kansas Geological Survey Bulletin 169, v. 2, 1964)

Fourier series characterization of cyclic sediments for stratigraphic correlation, by F.W. Preston and J.H. Henderson (reprinted from Symposium on cyclic sedimentation, D.F. Merriam, editor, Kansas Geological Survey Bulletin 169, v. 2, 1964)

Geology and the computer, by D.F. Merriam (reprinted from New Scientist, v. 26, no. 444, 1965)

Quantitative comparison of contour maps, by D.F. Merriam and P.H.A. Sneath (reprinted from Journal of Geophysical Research, v. 71, no. 4, 1966)

Trend-surface analysis of stratigraphic thickness data from some Namurian rocks east of Sterling, Scotland, by W.A. Read and D.F. Merriam (reprinted from Scottish Journal of Geology, v. 2, pt. 1, 1966)

Geologic model studies using trend-surface analysis, by D.F. Merriam and R.H. Lippert (reprinted from Journal of Geology, v. 74, no. 5, 1966)

Geologic use of the computer, by D.F. Merriam (reprinted from Wyoming Geol. Assoc., 20th Field Conf., 1966)

Computer aids exploration geologists, by D.F. Merriam (reprinted from the Oil and Gas Journal, 1967)

Comparison of cyclic rock sequences using cross-association, by D.F. Merriam and P.H.A. Sneath (reprinted from Essays in Paleontology and Stratigraphy: R.C. Moore commemorative volume, edited by C. Teichert and E. Yochelson, Dept. Geology, Univ. Kansas, 1967)