[AMERICAN JOURNAL OF SCIENCE, VOL. 265, SUMMER 1967, P. 619-625]

TREND-SURFACE FITTING TO RANDOM DATA—AN EXPERIMENTAL TEST

R. J. HOWARTH

Geology Department, Bristol University, Bristol, England*

ABSTRACT. The percentage sum of squares accounted for by linear, quadratic, or cubic trend-surfaces is a commonly used measure of their reliability in extracting the regional trend "inherent" in the variation of a mapped variable with geographic coordinates. Sixty experimental tests of fitting such surfaces to randomly distributed data suggest that if the sum of squares test produces values that fall below 6.0, 12.0, and 16.2 percent for the linear, quadratic, and cubic surfaces respectively, the distribution of data points is not significantly different from random at the 0.05 level.

Trend-surface analysis may be defined as the least-squares best-fitting of a smooth surface to a mapped variable with non-orthogonal geographical (U, V) coordinates to investigate the systematic regional trends "inherent" in the data as opposed to small-scale local fluctuations.

In practice an integer power series in U and V is used as the approximating polynomial, and the coefficients of the first three degrees are listed in table 1. These functions will be referred to hereafter as the "linear", "quadratic", and "cubic" surfaces. Computer programs for fitting such surfaces up to the sixth degree are available (Whitten, 1963a; Harbaugh, 1963; Good, 1964; O'Leary, Lippert, and Spitz, 1966).

It should be emphasized that a power series of this type is only one of a large number that could be used for trend-surface fitting. Orthogonal polynomials such as the Fourier and Chebyshev series could also be used (Spitz, 1966; Harbaugh and Preston, 1965; Preston and Harbaugh, 1965). The analysis presented here pertains only to the conventional power series.

Tests to indicate the reliability of the fit of the trend-surfaces must form an important part in estimating the usefulness of the technique in the investigation of areally distributed data. It is highly desirable to

TABLE 1

Classification of the trend-surface equations in which the terms are grouped according to degree

Surface	Dependent variable	Linear component	Quadratic component	Cubic component
Degree 1	$x =$	$a_0+a_1U+a_2V$		
Degree 2	$x =$	$b_0+b_1U+b_2V$	$+b_3U^2+b_4UV+b_5V^2$	
Degree 3	$x =$	$c_0+c_1U+c_2V$	$+c_3U^2+c_4UV+c_5V^2$	$+c_6U^3+c_7U^2V+c_8UV^2+c_9V^3$

a, b, and c are the coefficients of the equations of degree 1, 2, and 3 respectively.

[*] Present address: Bataafse Internationale Petroleum Maatschappij N.V., The Hague, Netherlands.

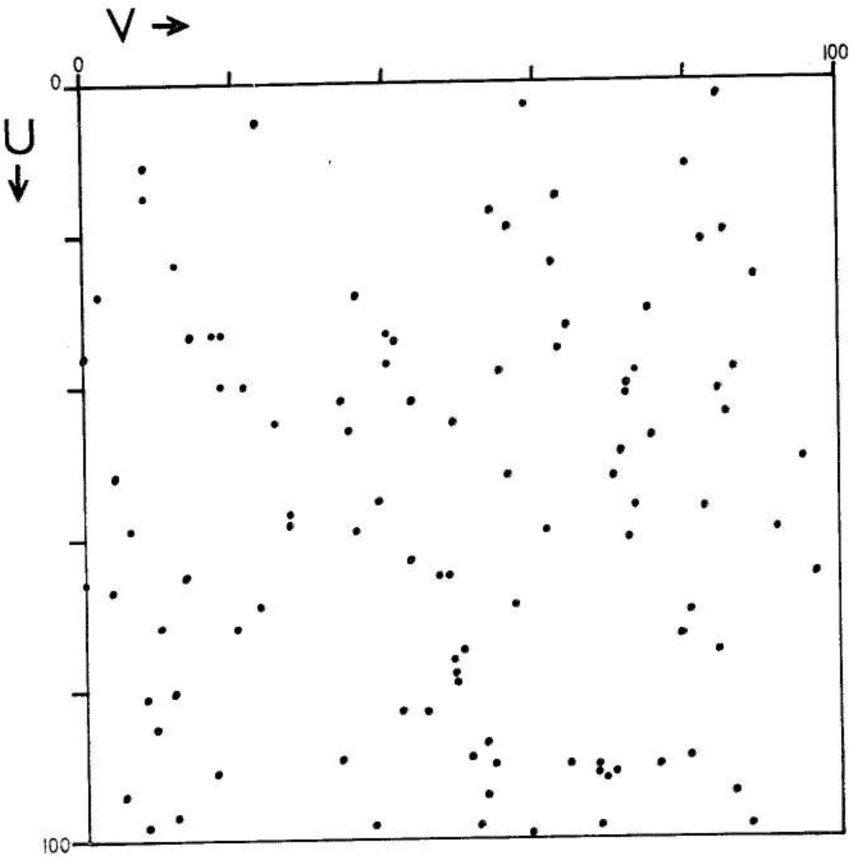

Fig. 1. Typical distribution of data points using randomly generated U and V coordinates.

distinguish between data whose variability overlies a significant geological trend and data distributed in a truly random manner.

Confidence limits can be calculated (Krumbein and Graybill, 1965, p. 340-345), but this is a complex procedure, and the results might be difficult to use in practice.

The most commonly used measure of reliability of the fitted surfaces is the sum of squares test which has been used by Allen and Krumbein (1962), Dawson and Whitten (1962), Whitten (1963a), Whitten and Boyer (1964), Harbaugh (1964a,b), Merriam and Harbaugh (1964), Merriam (1964), Duff and Walton (1964), and Merriam and Sneath (1966), but as yet no limits of reliability have been proposed.

The sum of squares test is based on the percentage of the total sum of squares of a particular variate (X) accounted for by a surface of given degree. For n observations the percent of squares accounted for (p) is given by

$$p = 100 \left[\frac{\sum_{i=1}^{n} X_{\text{computed}}^{2} - \left(\sum_{i=1}^{n} X_{\text{computed}} \right)^{2}/n}{\sum_{i=1}^{n} X_{\text{observed}}^{2} - \left(\sum_{i=1}^{n} X_{\text{observed}} \right)^{2}/n} \right]$$

Whitten (1963b) stated that "the sum of squares and the confidence levels for each surface are at least informative, if not decisive". In some cases, however, there has been doubt about the validity of assuming that such a surface exists where the contribution of that surface to the total sum of squares is small (Chayes and Suzuki, 1963, p. 308).

Accordingly, a computer program was devised using a random number generator to provide artificial data sets of 100 points each in which both the geographic (U, V) coordinates and the variate X varied in a random manner between 0 to 100 and 1 to 9 respectively. A typical data point distribution is shown in figure 1. Trend-surfaces were fitted to 60 such data sets using a program written by the author in Algol 60 for the Elliott 503 computer. In all cases the matrices inverted symmetrically, and the coefficients of the trend-surfaces and the percentage

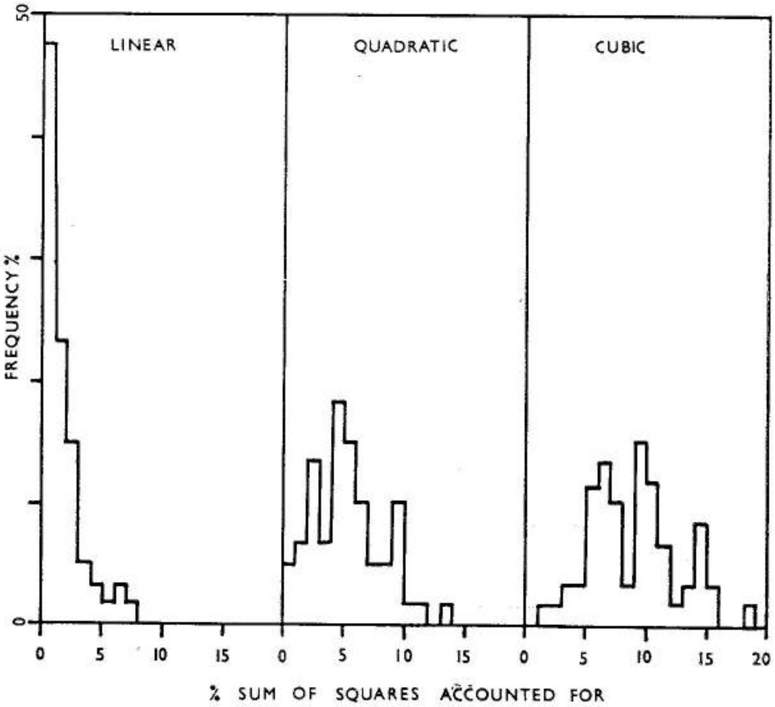

Fig. 2. Histograms of the frequency distribution of the values of the percent sum of squares accounted for by the 60 linear, quadratic, and cubic trend-surfaces (see al o table 2).

sum of squares accounted for by the three surfaces were compared. The frequency distribution, mean, and standard deviation of the percent of squares values for each surface are given in table 2 and are shown as histograms in figure 2. The frequency distributions are all asymmetrical and positively skewed. Maximum values of the percent of squares accounted for by the surfaces were linear, 7.89; quadratic, 13.18; and cubic, 18.12.

The appearance of the trend-surfaces themselves is probably not diagnostic. Many of the linear surfaces were horizontal or of very low angle slope at a level close to that of the mean. The quadratic and

TABLE 2

Percent frequency distribution of the values of the percent of squares accounted for by the linear, quadratic, and cubic trend-surfaces fitted to 60 random data sets. The expected lognormal frequencies (for a sample size of 60) are also shown.

	Percent frequency per class					
	Linear surface		Quadratic surface		Cubic surface	
Percent of squares	observed	lognormal	observed	lognormal	observed	lognormal
0.0- 0.9	46.7	46.7	5.0	0.0	0.0	0.0
1.0- 1.9	23.3	25.0	6.7	8.3	1.7	0.0
2.0- 2.9	15.0	13.3	13.3	13.3	1.7	1.7
3.0- 3.9	5.0	5.0	6.7	18.3	3.3	5.0
4.0- 4.9	3.3	3.3	18.3	13.3	3.3	8.3
5.0- 5.9	1.7	1.7	15.0	13.3	11.7	11.6
6.0- 6.9	3.3	1.7	10.0	8.3	13.3	11.6
7.0- 7.9	1.7	1.7	5.0	6.7	10.0	11.6
8.0- 8.9		1.7	5.0	5.0	3.3	10.0
9.0- 9.9			10.0	3.3	15.0	10.0
10.0-10.9			1.7	3.3	11.7	6.7
11.0-11.9			1.7	1.7	6.7	6.7
12.0-12.9			0.0	1.7	1.7	5.0
13.0-13.9			1.7	1.7	3.3	3.3
14.0-14.9				1.7	8.3	1.7
15.0-15.9					3.3	1.7
16.0-16.9					0.0	1.7
17.0-17.9					0.0	1.7
18.0-18.9					1.7	1.7
19.0-19.9						

	Surface		
	Linear	Quadratic	Cubic
Mean percent of squares (observed)	1.7	5.1	8.9
Median percent of squares (observed)	1.1	4.7	10.0
Std deviation percent of squares (observed)	1.8	2.9	3.4

cubic surfaces were of similarly low amplitude. Three typical results are shown in figure 3.

Whereas it could be said that these surfaces represent real trends hidden by "noise" in the data, it is inevitable that since the data were generated in a random manner, devoid of any underlying (geological) process that could impose some sort of orderly variation upon the system, we may expect any such trends to be purely fortuitous.

The cumulative relative frequency distributions of the percentage sum of squares values appear to be lognormal when plotted on logarithmic-probability paper. The hypothesis that the sample frequency distributions are lognormal may be tested by the nonparametric Kolmogorov-Smirnov statistic. The maximum observed absolute differences between the sample and calculated lognormal distributions (d_n) are linear, 0.017; quadratic, 0.082; and cubic, 0.115 respectively. Since these values are all smaller than the critical value of $d_n = 0.138$ for a sample size of 60, with a 0.20 chance that this value is not exceeded by

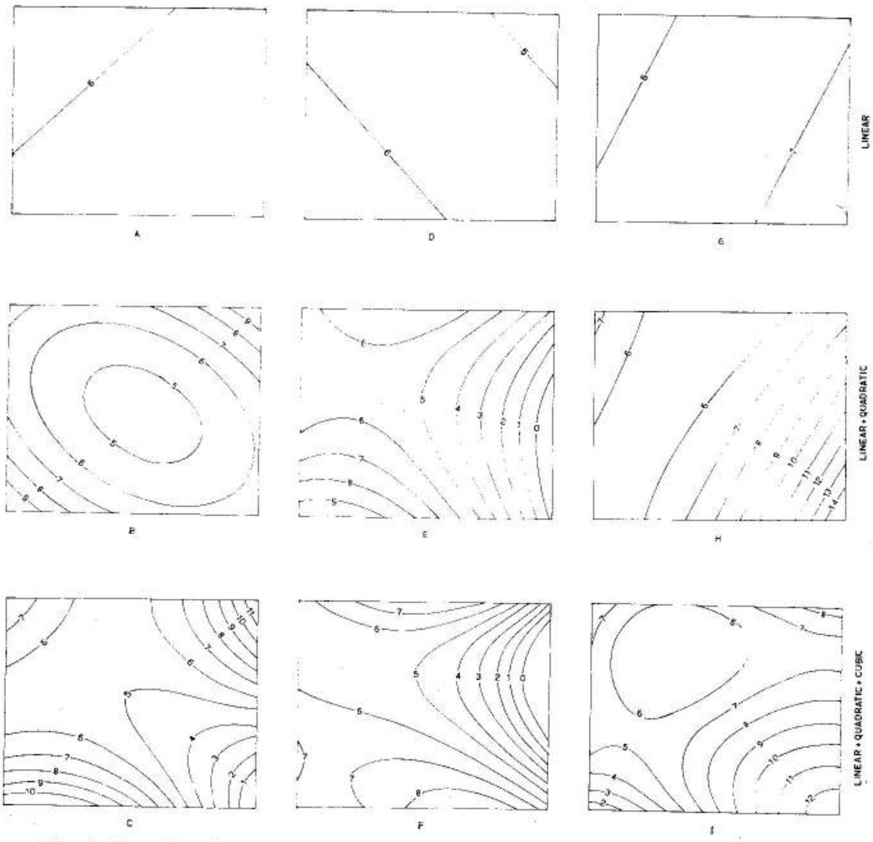

Fig. 3. Trend-surfaces of "regional variation" in three sets of randomly distributed data (A-C, D-F, G-I). Sums of squares accounted for are respectively: A, 0.8; B, 6.9; C, 9.7; D, 1.1; E, 8.1; F, 9.9; G, 1.4; H, 7.1; I, 9.2 percent.

the maximum difference between the distributions of the population and the sample (Miller, 1956), we may accept the hypothesis that the lognormal distribution is a very good fit in all three cases. Accordingly the lognormal distributions have been used to estimate the probable maximum values attained by the percent sum of squares. Ninety-five percent of the observed values of the percent sum of squares are smaller than 6.0, 12.0, and 16.2 for the linear, quadratic, and cubic surfaces respectively. It is suggested that these values be taken as an upper limit to the percent sum of squares accounted for by the random data.

For example, the data for the feldspar ratio and total feldspar percentage surfaces in the granitic complex of Quebec (Dawson and Whitten, 1962) yield such low percentage sums of squares as:

	Feldspar ratio	Total feldspar (modal percent)
Linear surface	9.44	2.43
Quadratic surface	14.21	4.50
Cubic surface	15.16	6.39

The authors commented on this, remarking that "for numerous granite complexes it has been found consistently that less significance can be attached to the total feldspar surfaces than those for any other modal variates". These results suggest that the total feldspar distribution is very probably random and that the feldspar ratio may well be also.

The mean value of the dependent variable in the data will affect the percent of squares accounted for. In order to achieve a consistent measure for determining the reliability it may be desirable to normalize the data prior to computing the percent of squares values. Possible methods include the percentage range transformation, in which each sample is expressed as a proportion of the maximum value or reduction to a zero mean.

It is also possible that the answer may be affected by ill-conditioning of the data. If this is suspected, then it is advisable to check the possibility by rerunning the computation with small perturbations of the data (say ± 1 in the last digit). If this causes large changes in the answer, then it is only worth quoting the figures that remain unchanged. If there are large differences between the apparent trends of surfaces fitted to random sub-samples of the original data set, an unreliable fit is also suggested (Agterberg, 1964).

In conclusion, it is suggested that suspicion must fall on any proposition that the distribution of a mapped variable is significantly different from random when the sum of squares test produces values which fall close to, or below, 6.0, 12.0, and 16.2 percent for the linear, quadratic, and cubic surfaces respectively.

ACKNOWLEDGMENTS

This study was carried out while the writer held a Shell International Petroleum Company research award at the University of Bristol. Dr. B. E. Leake supervised the project of which this forms a part, and grateful acknowledgments are made to the staff of the Computer Unit, Bristol University, and to Professor J. W. Harbaugh and Mr. D. Hamilton for reading the manuscript.

REFERENCES

Agterberg, F. P., 1964, Methods of trend-surface analysis, *in* International Symposium; Applications of statistics, operations research, and computers in the mineral industry: Colorado School Mines Quart., v. 59, no. 4, pt. A, p. 111-130.

Allen, Percival, and Krumbein, W. C., 1962, Secondary trend components in the Top Ashdown Pebble Bed: a case history: Jour. Geology, v. 70, p. 507-538.

Chayes, Felix, and Suzuki, Y., 1963, Geological contours and trend-surfaces. A Discussion: Jour. Petrology, v. 4, p. 307-312.

Dawson, K. R., and Whitten, E. H. T., 1962, The quantitative mineralogical composition and variation of the Lacorne, La Motte, and Preissac granitic complex, Quebec, Canada: Jour. Petrology, v. 3, p. 1-37.

Duff, P. McL. D., and Walton, E. K., 1964, Trend surface analysis of sedimentary features of the Modiolaris zone, East Pennine Coalfield, England, *in* Developments in sedimentology, v. 1, Deltaic and shallow marine deposits: Amsterdam, Elsevier Publishing Co., p. 114-122.

Good, D. I., 1964, FORTRAN II trend-surface program for the IBM 1620: Kansas Geol. Survey Spec. Distrib. Pub. 14, 54 p.

Harbaugh, J. W., 1963, BALGOL program for trend-surface mapping using an IBM 7090 computer: Kansas Geol. Survey Spec. Distrib. Pub. 3, 17 p.

———— 1964a, 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.

———— 1964b, Trend-surface mapping of hydrodynamic oil traps with the IBM 7090/94 computer, *in* International Symposium; Applications of statistics, operations research, and computers in the mineral industry: Colorado School Mines Quart., v. 59, no. 4, pt. B, p. 557-578.

Harbaugh, J. W., and Preston, F. W., 1965, Fourier series analysis in geology: Symposium on computers and computer applications in mining and exploration, Univ. Arizona, Tucson, March 1965, p. R1-R46.

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

O'Leary, M., Lippert, R. H., and Spitz, O. T., 1966, FORTRAN IV and MAP program for computation and plotting of trend surfaces for degrees 1 through 6: Kansas Geol. Survey Computer Contr. 3, 47 p.

Merriam, D. F., 1964, Use of trend-surface residuals in interpreting geologic structures, *in* Computers in the mineral industries, pt. 2: Stanford Univ. Pub. Geol. Sci., v. 9, pt. 2, p. 686-692.

Merriam, D. F., and Harbaugh, J. W., 1964, Trend-surface analysis of regional and residual components of geologic structure in Kansas: Kansas Geol. Survey Spec. Distrib. Pub. 11, 27 p.

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

Miller, L. H., 1956, Tables of percentage points of Kolmogorov statistics: Am. Stat. Assoc, Jour.. v. 51, p. 111-121.

Preston, F. W., and Harbaugh, J. W., 1965, BALGOL programs and geologic applications for single and double Fourier series using IBM 7090/7094 computers: Kansas Geol. Survey Spec. Distrib. Pub. 24, 72 p.

Spitz, O. T., 1966, Generation of orthogonal polynomials for trend surfacing with a digital computer, *in* Symposium on Computers and Operation Research in Mineral Industries: University Park, The Pennsylvania State Univ., 1966, 6 p.

Whitten, E. H. T., 1963a, A surface-fitting program suitable for testing geological models which involve areally-distributed data: Office Naval Research, Geography Branch, Tech. Rept. 2, ONR Task No. 389-135, Contract Nonr. 1228 (26), 56 p.

———— 1963b, A reply to Chayes and Suzuki: Jour. Petrology, v. 4, p. 313-316.

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.