STATISTICS AND DATA ANALYSIS IN GEOLOGY, 3rd ed.

by John C. Davis

Clarification of zonation procedure described on pp. 238-239

Because the notation used in this section (Eqs. 4.80 through 4.84) is inconsistent with that used elsewhere in the book, it may be confusing. This confusion can be resolved by some redefinition of the equations. Let x_{ij} represent the standardized value of log trace j measured at depth i, where there are n depths and m log traces. Each log trace is standardized by subtracting the mean of the trace from every value and dividing by the standard deviation of the trace. This transforms the log traces so that each trace has a mean of zero and variance of one. The vector $\mathbf{x}_i$ consists of the standardized values for all log traces at depth i. For the purposes of demonstrating the clustering process, we will regard $\mathbf{x}_i$ as a column vector and use inner product notation ($\mathbf{x}'\mathbf{x}$) to represent sums of squares, even though $\mathbf{x}_i$ is actually a row in an $n \times m$ matrix. Using this convention and keeping in mind that the log traces are standardized, the total sum of squares is given by

$$SS_T = \sum_i^n \mathbf{x}_i'\mathbf{x}_i = m \times (n-1) \tag{4.80}$$

Because the log traces are standardized, the total sum of squares is simply the number of log traces times the number of degrees of freedom for the variance of each trace, which is the number of depth intervals minus one. Consequently, we do not have to calculate the sum of the cross-products to find SS_T.

The zonation process consists of iteratively collecting adjacent depths into zones composed of intervals that have similar log characteristics. The measure of similarity (or rather, of dissimilarity) is the ***squared distance***, E^2. (A greater squared distance between two adjacent zones indicates a greater dissimilarity between them. If two zones are identical, the squared distance between them will be zero.) We will refer to zones by the subscript k, and the number of points within a zone as n_k. Because on the first iteration each zone includes only a single depth (that is, $n_k = 1$ for all k), the squared distance between every observation at depth i and its neighbor at depth $i+1$ is given by

$$E_k^2 = E_i^2 = \frac{1}{2}\left(\mathbf{x}_i - \mathbf{x}_{i+1}\right)'\left(\mathbf{x}_i - \mathbf{x}_{i+1}\right) \tag{4.81}$$

After the first iteration, some of the zones will include more than a single depth; these zones are henceforth treated as a single object defined by their vector mean,

$$\overline{\mathbf{X}}_k = \frac{1}{n_k}\sum_i^{n_k} x_i \tag{4.82}$$

again, treating these as column vectors. We also must calculate a measure of the variation within a zone in the form of a within-zone sum of squares, W_k.

$$W_k = \sum_i^{n_k}\left(\mathbf{x}_i - \overline{\mathbf{X}}_k\right)'\left(\mathbf{x}_i - \overline{\mathbf{X}}_k\right) \tag{4.83}$$

Note that the summation extends over the set of n_k depths included in zone k. If we sum this measure over all g zones, we will obtain the total sum of squares within zones, SS_W.

$$SS_W = \sum_k^g W_k = \sum_k^g \sum_i^{n_k}\left(\mathbf{x}_i - \overline{\mathbf{X}}_k\right)'\left(\mathbf{x}_i - \overline{\mathbf{X}}_k\right) \tag{4.84}$$

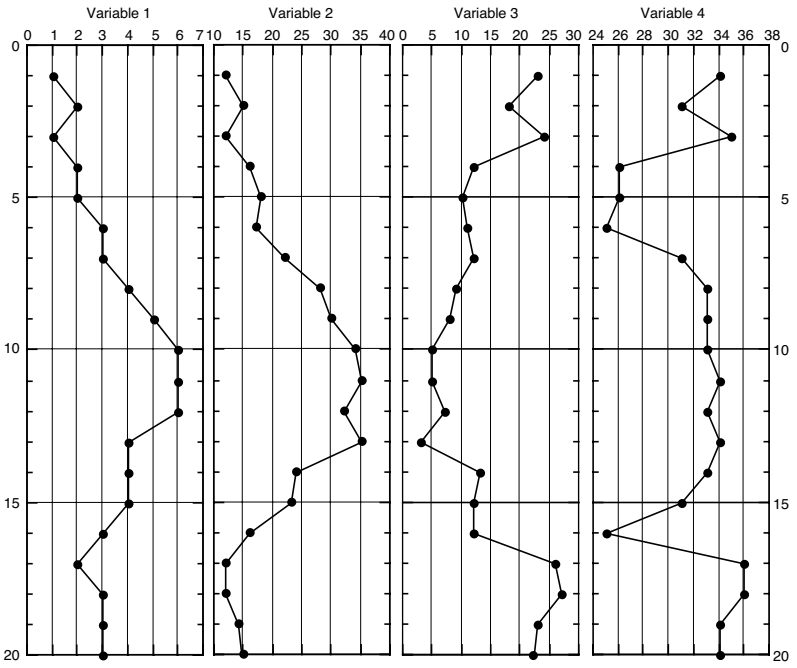


Figure 1. Hypothetical well log with four traces.

As iteration proceeds and depths become grouped into zones, the dissimilarity between adjacent zones is determined as the squared distance between the zone means weighted by the number of depths within each zone. That is, the squared distance between zone k and zone $k+1$ is

$$E_k^2 = \frac{(\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_{k+1})' (\overline{\mathbf{X}}_k - \overline{\mathbf{X}}_{k+1})}{(1/n_k + 1/n_{k+1})}$$

This is the amount by which the merger of zones k and $k+1$ will increase the total sum of squares within zones. As noted previously, on the initial pass each zone consists of a single depth, so $n_k = 1$ for all k and the squared distances E_k^2 are given by Equation (4.81).

A general summary of the steps in the zonation algorithm, after standardizing the log traces, is

(1) Calculate the distances between the vectors of every interval (or zone) and the underlying interval (or zone).

(2) Combine the intervals whose squared distance is smallest to form a zone.

(3) Replace the vectors of intervals in a zone with their mean vector.

(4) Return to (1). Repeat until all intervals are combined into a single zone or some stopping criterion is satisfied.

Operation of the zonation algorithm can be demonstrated using the hypothetical well log shown in **Figure 1**, which consists of four well log variables measured at 20 successive depths; the data are listed in **Table 1**. The means and standard

Table 1. Hypothetical example of a "well log" consisting of four traces (variables) measured at 20 equally spaced depths.

var. 1	var. 2	var. 3	var. 4	depth
1	12	23	34	1
2	15	18	31	2
1	12	24	35	3
2	16	12	26	4
2	18	10	26	5
3	17	11	25	6
3	22	12	31	7
4	28	9	33	8
5	30	8	33	9
6	34	5	33	10
6	35	5	34	11
6	32	7	33	12
4	35	3	34	13
4	24	13	33	14
4	23	12	31	15
3	16	12	25	16
2	12	26	36	17
3	12	27	36	18
3	14	23	34	19
3	15	22	34	20

deviations of the four traces are

	var. 1	**var. 2**	**var. 3**	**var. 4**
mean	3.35	21.1	14.1	31.85
std. dev.	1.5312	8.4161	7.5735	3.5433

which can be used to standardize the readings listed in **Table 1**, giving the transformed values in **Table 2**. The first step is to calculate the total variance in the sequence, using Equation (4.80) as modified above. For depth 1, the multiplication is

$$\mathbf{x}_1'\mathbf{x}_1 = \begin{bmatrix} -1.5346 & -1.0813 & 1.1752 & 0.6068 \end{bmatrix} \begin{bmatrix} -1.5346 \\ -1.0813 \\ 1.1752 \\ 0.6068 \end{bmatrix} = 5.2735$$

The twenty values for the log are given below; the sum of these values is $SS_T = 76.0$, which is also the product $4 \times 19 = 76$.

1.6253	0.1981	7.5347	4.2336
6.0234	1.4112	5.6566	5.7870
3.9471	3.0335	5.4242	5.4943
3.9318	6.8935	0.4254	2.5131
4.1944	5.2735	0.3656	2.0338

Table 2. Hypothetical well log with variables standardized by subtracting column means and dividing by column standard deviations.

var. 1	var. 2	var. 3	var. 4	depth
−1.5347	−1.0813	1.1752	0.6068	1
−0.8816	−0.7248	0.5150	−0.2340	2
−1.5347	−1.0813	1.3072	0.8889	3
−0.8816	−0.6060	−0.2773	−1.6510	4
−0.8816	−0.3683	−0.5414	−1.6510	5
−0.2286	−0.4872	−0.4093	−1.9332	6
−0.2286	0.1069	−0.2773	−0.2399	7
0.4245	0.8199	−0.6734	0.3246	8
1.0775	1.0575	−0.8054	0.3246	9
1.7306	1.5328	−1.2016	0.3246	10
1.7306	1.6516	−1.2016	0.6068	11
1.7306	1.2951	−0.9375	0.3246	12
0.4245	1.6516	−1.4656	0.6068	13
0.4245	0.3446	−0.1452	0.3246	14
0.4245	0.2258	−0.2773	−0.2399	15
−0.2286	−0.6060	−0.2773	−1.9332	16
−0.8816	−1.0813	1.5713	1.1712	17
−0.2286	−1.0813	1.7033	1.1712	18
−0.2286	−0.8436	1.1752	0.6068	19
−0.2286	−0.7248	1.0431	0.6068	20

During this first iteration, the distance between every log measurement and the immediately underlying measurement is calculated. For example, we first find the vector difference between the measurements at depth 1 and depth 2:

$$(\mathbf{x}_1' - \mathbf{x}_2') = \begin{bmatrix} -1.5347 & -1.0813 & 1.1752 & 0.6068 \end{bmatrix} -$$

$$\begin{bmatrix} -0.8816 & -0.7248 & 0.5145 & -0.2399 \end{bmatrix} = \begin{bmatrix} -0.6531 & -0.3565 & 0.6602 & 0.8467 \end{bmatrix}$$

The distance between the log readings at these two depths is found by postmultiplying the vector of differences by its transpose, as indicated in modified Equation (4.81). That is,

$$E_1^2 = \tfrac{1}{2} \begin{bmatrix} -0.6531 & -0.3565 & 0.6602 & 0.8467 \end{bmatrix} \begin{bmatrix} -0.6531 \\ -0.3565 \\ 0.6602 \\ 0.8467 \end{bmatrix}$$

$$= \tfrac{1.7062}{2} = 0.8531$$

After repeating this calculation for every depth interval in the log we obtain the 19 distances listed below.

depth interval	distance
1 and 2	0.8531
2 and 3	1.2278
3 and 4	4.8072
4 and 5	0.0631
5 and 6	0.2689
6 and 7	1.6189
7 and 8	0.7051
8 and 9	0.2502
9 and 10	0.4047
10 and 11	0.0469
11 and 12	0.1382
12 and 13	1.0958
13 and 14	1.7657
14 and 15	0.1751
15 and 16	1.9928
16 and 17	6.8535
17 and 18	0.2220
18 and 19	0.3270
19 and 20	0.0158

The smallest squared distance (greatest similarity) is between depths 19 and 20, the last two values in the sequence. Since these two are most similar, we will combine them to form a new zone defined by the vector mean of the two original intervals. The new, combined zone is

$$\frac{(\mathbf{x}'_{19} + \mathbf{x}'_{20})}{2} =$$

$$\frac{\begin{bmatrix}-0.2286 & -0.8436 & 1.1752 & 0.6068\end{bmatrix} + \begin{bmatrix}-0.2286 & -0.7248 & 1.0431 & 0.6068\end{bmatrix}}{2}$$

$$= \begin{bmatrix}-0.2286 & -0.7842 & 1.1092 & 0.6068\end{bmatrix}$$

Finally, we calculate a measure of variation within the new zone using modified Equation (4.83), subtracting the zone mean, $\overline{\mathbf{X}}_k$, from $\mathbf{x}_{19}$ and from $\mathbf{x}_{20}$, postmultiplying each difference vector by its transpose, and summing the products. The difference vectors are:

$$\begin{aligned}\mathbf{x}'_{19} - \overline{\mathbf{X}}'_k &= \begin{bmatrix}-0.2286 & -0.8436 & 1.1752 & 0.6068\end{bmatrix} \\ &- \begin{bmatrix}-0.2286 & -0.7842 & 1.1092 & 0.6068\end{bmatrix} \\ &= \begin{bmatrix}0.0000 & -0.0594 & 0.0660 & 0.0000\end{bmatrix}\end{aligned}$$

$$\begin{aligned}\mathbf{x}'_{20} - \overline{\mathbf{X}}'_k &= \begin{bmatrix}-0.2286 & -0.7248 & 1.0431 & 0.6068\end{bmatrix} \\ &- \begin{bmatrix}-0.2286 & -0.7842 & 1.1092 & 0.6068\end{bmatrix} \\ &= \begin{bmatrix}0.0000 & 0.0594 & -0.0660 & 0.0000\end{bmatrix}\end{aligned}$$

(Note that the deviations from the zone mean are symmetrical only when a zone consists of just two intervals.)

For combined depths 19 and 20, W_k is

$$W_k = \begin{bmatrix} 0.0000 & -0.0594 & 0.0660 & 0.0000 \end{bmatrix} \begin{bmatrix} 0.0000 \\ -0.0594 \\ 0.0660 \\ 0.0000 \end{bmatrix}$$

$$+ \begin{bmatrix} 0.0000 & -0.0594 & 0.0660 & 0.0000 \end{bmatrix} \begin{bmatrix} 0.0000 \\ -0.0594 \\ 0.0660 \\ 0.0000 \end{bmatrix} =$$

$$0.0079 + 0.0079 =$$

$$0.0158$$

SS_W is found simply by summing all the W_k for all zones that contain more than one original depth. During the initial iteration, there is only one such zone, so $SS_W = W_k = 0.0158$.

In each successive iteration, one depth interval or zone is combined with an adjacent depth interval or zone, so the number of zones decreases by one with each iteration. **Table 3** shows the successsive combination of zones during the 18 iterations needed to combine all of the depth intervals into two zones. As an

Table 3. Order of combination of depth intervals into zones in successive iterations of the zonation algorithm.

Iteration	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1
	3	3	3	3	3	3	3	3	3	3	2	1	1	1	1	1	1	1	1
	4	4	4	4	4	4	4	4	4	4	3	2	2	2	2	2	2	1	1
	5	5	5	4	4	4	4	4	4	4	3	2	2	2	2	2	2	1	1
	6	6	6	5	5	5	5	5	4	4	3	2	2	2	2	2	2	1	1
	7	7	7	6	6	6	6	6	5	5	4	3	3	3	2	2	2	1	1
	8	8	8	7	7	7	7	7	6	6	5	4	4	4	3	3	3	2	1
	9	9	9	8	8	8	8	7	6	6	5	4	4	4	3	3	3	2	1
	10	10	10	9	9	9	9	8	7	7	6	5	5	4	3	3	3	2	1
	11	11	10	9	9	9	9	8	7	7	6	5	5	4	3	3	3	2	1
	12	12	11	10	9	9	9	8	7	7	6	5	5	4	3	3	3	2	1
	13	13	12	11	10	10	10	9	8	8	7	6	5	4	3	3	3	2	1
	14	14	13	12	11	11	11	10	9	9	8	7	6	5	4	4	3	2	1
	15	15	14	13	12	11	11	10	9	9	8	7	6	5	4	4	3	2	1
	16	16	15	14	13	12	12	11	10	10	9	8	7	6	5	4	3	2	1
	17	17	16	15	14	13	13	12	11	11	10	9	8	7	6	5	4	3	2
	18	18	17	16	15	14	13	12	11	11	10	9	8	7	6	5	4	3	2
	19	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2
	20	19	18	17	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2

example, **Figure 2** shows the log traces after 15 iterations, where individual values within zones have been replaced by the mean values of the zones.

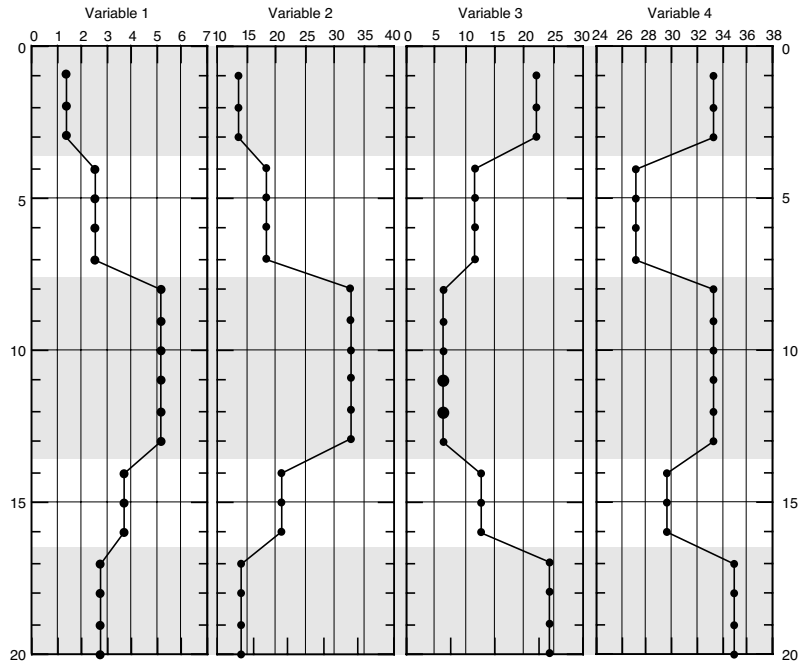


Figure 2. Well log from Figure 1 after replacement of values within each zone with the mean values for the zone. Log is shown after 15 iterations when log has been grouped into five zones.