Mathematical Geology, Vol. 9, No. 1, 1977

Locating Maximum Variance Segments in Sequential Data[1]

T. R. Bement[2] and M. S. Waterman[2]

An automated method is presented for the identification of peaks in sets of sequential data. The method is based upon the location of those segments with maximum variance and has the advantage of guarding against the masking of small-scale effects by large-scale effects. The procedure is illustrated with data taken as part of the National Uranium Resource Evaluation project. KEY WORDS: algorithm, sequential analysis, data processing.

INTRODUCTION

Many types of investigations in geology and other disciplines are related to the problem of locating segments possessing certain properties in a set of sequential data. The problem that motivated the algorithms reported here involves the interpretation of sets of sequential data taken as part of the National Uranium Resource Evaluation (NURE) project. Figure 1 is a set of digitized data of an airborne scan of the gamma-ray signal from ^{214}Bi taken along a transect. This scan, although unusual because of the extremely high peak at A caused by an exposed uranium deposit, was chosen as an example because it illustrates a particular problem encountered in the analysis. A desirable procedure for recognizing certain types of segments should not allow relatively large-scale effects to mask small-scale effects.

The problem of recognizing specific types of segments in sequential data has been considered by several authors. Hawkins and Merriam (1973) present a method of dividing such data sets into homogeneous zones based on minimizing within-zone variation. Kulinkovich, Sokhranov, and Churinova (1966) published an algorithm for identifying boundaries of beds in borehole profiles. Their algorithm is based on a search for the maximum change between two consecutive records in a log.

The purpose of the methods reported here is to present an automated

[1] Manuscript received 19 March 1976. This work was performed under the auspices of the U.S. Energy Research and Development Administration, Grand Junction Office.

[2] Los Alamos Scientific Laboratory, Los Alamos, New Mexico 87545.

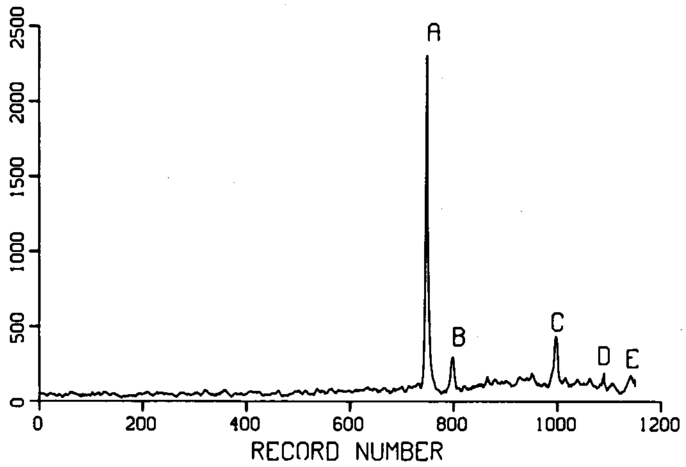

Figure 1. Digitized ^{214}Bi scan taken by airborne gamma-ray detectors along transect.

technique for locating peaks such as those at A, B, and C in Figure 1. The algorithms are based on determining the location of segments whose variance is maximal.

ALGORITHMS

Denote the data by $x_1,x_2,\ldots, x_N$, where N is the number of observations in a scan. The object of the algorithms proposed here is to select the most heterogeneous segments. This problem does not seem to have been considered previously in a mathematical fashion.

Our measure of heterogeneity is the sum of the variances of the selected segments. The variance of the segment $x_i,x_{i+1},\ldots, x_j$ is, as usual, defined to be

$$v(i,j)=(j-i)^{-1}\sum_{l=i}^{j}(x_l-\bar{x}(i,j))^2$$

where

$$\bar{x}(i,j)=(j-i+1)^{-1}\sum_{l=i}^{j}x_l,$$

and

$$v(i,i)=0$$

When we select the k segments n_1 to m_1, n_2 to $m_2,\ldots, n_k$ to m_k, we compute the sum of the variances

$$v(n_1,m_1)+v(n_2,m_2)+\cdots+v(n_k,m_k)$$

as the measure of heterogeneity of those k segments. The largest such sum with $1 \leq n_1 \leq m_1 < n_2 \leq m_2 < \cdots < n_k \leq m_k \leq j$ is denoted by $S_k(j)$ (note that $S_k(j)$ is undefined if $k > j$).

Our algorithm for computing $S_k(N)$ proceeds as follows:

$$\begin{aligned} & S_i(1) = 0, \qquad 1 \leq i \leq N \\ & S_1(j) = \max\{S_1(j-1), \max_{1<l<j} v(l,j)\}, \qquad 2 \leq j \leq N \\ & \quad \ldots \\ & S_i(j) = \max[S_i(j-1), \max_{i \leqslant l < j} \{S_{i-1}(l-1) + v(l,j)\}], \qquad 2 \leq i \leq N, \qquad i < j \leq N \end{aligned} \tag{1}$$

The algorithm can be easily justified. The value of $S_i(j)$ is either the optimal choice of i segments from $x_1, x_2, \ldots, x_{j-1}$ or it must be the situation that the last segment ends with x_j. If the latter is the situation, it must be true that $S_i(j)$ is the maximum of $v(l,j)$ plus the S value for the optimal choice of $i-1$ segments from $x_1, x_2, \ldots, x_{l-1}$ for some l in the interval from 1 to $j-1$. This completes the justification of (1). Clearly, the algorithm (1) can be programmed easily. Essentially, the number of operations to compute $S_k(N)$ is proportional to $kN(N+1)/2$.

Of course, $S(N) = \max S_k(N)$ can be calculated as indicated, but it is useful for considering stopping rules to have a more efficient method of computing $S(N)$. The following algorithm solves this problem.

$$\begin{aligned} & S(0) = 0 \\ & S(j) = \max_{1 \leqslant l < j} \{S(l-1) + v(l,j)\}, \qquad 1 \leq j \leq N \end{aligned} \tag{2}$$

The reasoning for this algorithm is similar to that for algorithm (1). The number of operations is proportional to $N(N+1)/2$. Because $S(N)$ is the maximum variance that can be obtained by the removal of segments, it is useful to compare $S_k(N)$ with $S(N)$. Another procedure would be to compute $S(N)$ and then select all segments which had larger variance than a preassigned percentage of $S(N)$.

The problem of locating the boundaries of the segments which give the sum of variances equal to $S_k(N)$ or $S(N)$ is similar to the corresponding problem in Hawkins and Merriam (1973) but, in our situation, both upper and lower boundaries must be handled.

One useful modification of the algorithms is to impose the restrictions that all segments must be of minimum width w and maximum width W. These modifications are included easily in the algorithm. As presented, the algorithms have $w = 1$ and $W = N$. One useful feature of these modifications is the reduction of computation time. For example, with $w = 1$, the computation time for $S_k(N)$ is proportional to $kW(W+1)/2$. For $N = 1000$, the use of $W = 100$ reduces computation to approximately 10^{-2} times that required for $W = 1000$. It should be pointed out that the algorithm of Kulinkovich, Sokhranov, and Churinova (1966) is equivalent to our algorithm with $W = 2$.

An interesting property of algorithm (1) is that, if all x_i are not members of a segment used to compute $S_k(N)$, then $S_k(N) \leq S_{k+1}(N)$. Therefore, we feel algorithm (1) need not be used to remove segments beyond the point when each x_i is the member of some segment. However, examples exist where continuing beyond this point increases the total variance. Of course, stopping rules of the sort that terminate computation when the value $1 - \mathrm{S}_k(N)/S_{k+1}(N)$ or $1 - S_k(N)/S(N)$ is less than some specified value are reasonable and can be used.

Clearly, our algorithm will handle functions other than variance, and, whereas other such functions are being studied, variance seems satisfactory for our purposes. Also, by replacing max by min in (1) and imposing suitable restrictions such as $w > 1$, a choice of k segments whose sum is minimized can be determined. This could be used to locate homogeneous segments if that were the question of interest. A variant of (2) arose in connection with a problem in molecular biology (Waterman, 1976). Also, the algorithms can be used for questions of interest in numerical analysis (see Hawkins, 1972, for a similar analysis). In addition, a multivariate extension similar to Hawkins and Merriam (1974) should be possible. A theoretical paper is being prepared where these and other points will be dealt with in detail.

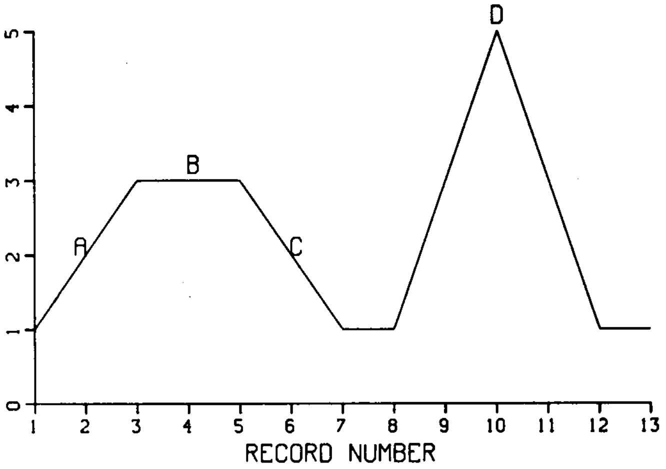

Figure 2. Hypothetical data illustrating increasing and decreasing segments at *A* and *C*, flat spot at *B*, and a peak at *D*.

Table 1. Segments with Maximum Total Variance from Data in Figure 2

Number of segments	Total variance	Boundaries				
1	4.0	10-12				
2	6.0	8-10	11-12			
3	7.0	5-7	8-10	11-12		
4	8.0	1-3	5-7	8-10	11-12	
5	8.0	1-3	5-7	8-10	11-12	13-13
6	8.0	1-3	4-4	5-7	8-10	11-12
		13-13				
7	7.5	1-3	4-4	5-6	7-7	8-10
		11-12	13-13			
8	7.0	1-3	4-4	5-5	6-6	7-7
		8-10	11-12	13-13		
9	6.5	1-1	2-3	4-4	5-5	6-6
		7-7	8-10	11-12	13-13	
10	6.0	1-1	2-2	3-3	4-4	5-5
		6-6	7-7	8-10	11-12	13-13

Maximum total variance = 8.0

A FORTRAN program implementing these algorithms has been written for a CDC 6600 computer at Los Alamos Scientific Laboratory. An elementary example, shown in Figure 2, is used to illustrate the method. In Table 1, results up to $S_{10}(13)$ are given. Of course, $S_{13}(13) = 0$ because each of the 13 selected segments has only one point. By almost any stopping rule, computation would proceed to $j = 5$ segments where $S_4(13) = S_5(13)$. Then the boundaries for $j = 4$ would be chosen. The four segments, when taken together, comprise regions *A* and *C* and the peak at *D* in the data of Figure 2. This illustrates the property of the algorithm that a peak will appear in at least two segments, made up of the ascending and the descending portion. If *A*-*B*-*C* is considered as a peak, note that only *A* and *C* are identified, leaving the flat segment *B*.

APPLICATIONS

As mentioned in the introduction, the data in Figure 1 results from flying a transect at approximately 120 mph and digitizing the ^{214}Bi radiation level at 1-sec intervals. The numbers along the horizontal axis are record numbers and the data set consists of 1150 points or records. It was hoped the algorithms would provide an automated method of locating *A*, *B*, and *C*, and that the large-scale effect of *A* would not mask the smaller effect of *B*.

Table 2. Segments with Maximum Total Variance from Data Shown in Figure 1

Number of segments	Total variance	Boundaries				
1	5.993E+5	740-747				
2	7.868E+5	740-746	747-748			
3	8.968E+5	740-746	747-748	749-760		
4	9.141E+5	740-746	747-748	749-760	995-1010	
5	9.241E+5	740-746	747-748	749-760	982-997	998-1006
6	9.326E+5	740-746	747-748	749-760	784-799	982-997
		998-1006				
7	9.378E+5	740-746	747-748	749-760	784-797	798-808
		982-997	998-1006			
8	9.406E+5	740-746	747-748	749-760	784-797	798-808
		982-997	998-1006	1090-1091		
9	9.426E+5	740-746	747-748	749-760	784-797	798-808
		982-997	998-1006	1090-1091	1119-1144	
10	9.493E+5	740-746	747-748	749-760	784-797	798-808
		982-997	998-1006	1090-1091	1119-1144	1149-1150

Maximum total variance = 9.623E+5

A run of the program was made with this data where the length of the longest segment to be removed was restricted to 50 or less. (That is, $W = 50$.) The results are given in Table 2. The regions *A*, *B*, and *C* of Figure 1 are each identified after the selection of seven segments. The region at *A* is composed of three segments. The total variance "identified" after the selection of seven segments is 97 percent of the maximum total variance in the scan. Because the large peak at *A* has such an influence in this percentage, it is interesting to evaluate the contribution of regions *A* and *B* to total variance in the absence of *C*. By subtracting the effect of *C* from the total variance and from the variance associated with the first seven segments, one finds about 62 percent of the variability has been identified. The last three segments to be selected identify regions *D* and *E*. Again ignoring the region *C*, one finds that when all ten segments have been selected, nearly 72 percent of the maximum total variability has been identified.

ACKNOWLEDGMENTS

The authors wish to express their appreciation to R. K. Zeigler for his many helpful suggestions.

REFERENCES

Hawkins, D. M., 1972, On the choice of segments in piecewise approximation: Jour. Inst. Math. Applications, v. 9, no. 2, p. 250–256.

Hawkins, D. M., and Merriam, D. F., 1973, Optimal zonation of digitized sequential data: Jour. Math. Geology, v. 5, no. 4, p. 389–395.

Hawkins, D. M., and Merriam, D. F., 1974, Zonation of multivariate sequences of digitized geologic data: Jour. Math. Geology, v. 5, no. 3, p. 263–269.

Kulinkovich, A. Ye., Sokhranov, N. N., and Churinova, I. M., 1966, Utilization of digital computers to distinguish boundaries of beds and identify sandstones from electric log data: Intern. Geology Rev., v. 8, no. 4, p. 416–420.

Waterman, M. S., 1976, Secondary structure of single-stranded nucleic acids: draft.