Journal of Geochemical Exploration, 3(1974)129—149

SELECTION OF THRESHOLD VALUES IN GEOCHEMICAL DATA USING PROBABILITY GRAPHS

A.J. SINCLAIR

Department of Geological Sciences, University of British Columbia, Vancouver, B.C. (Canada)

(Accepted for publication November 7, 1973)

ABSTRACT

Sinclair, A.J., 1974. Selection of threshold values in geochemical data using probability graphs. J. Geochem. Explor., 3: 129—149.

A method of choosing threshold values between anomalous and background geochemical data, based on partitioning a cumulative probability plot of the data is described. The procedure is somewhat arbitrary but provides a fundamental grouping of data values. Several practical examples of real data sets that range in complexity from a single population to four populations are discussed in detail to illustrate the procedure.

The method is not restricted to the choice of thresholds between anomalous and background populations but is much more general in nature. It can be applied to any polymodal distribution containing adequate values and populations with appropriate density distribution. As a rule such distributions for geochemical data closely approach a lognormal model. Two examples of the more general application of the method are described.

INTRODUCTION

Tennant and White (1959) were among the first to recognize the usefulness of probability graph paper for concise visual representation of geochemical data. Since the appearance of their publication probability paper has been used somewhat spasmodically, but with increasing regularity for graphical representation and analysis of many types of geochemical data. In particular, Williams (1967) and Lepeltier (1969) have emphasized the ease with which such plots can be used for rapid, graphical analysis of large quantities of data. Bolviken (1971) states that probability graphs are now used routinely by the Norwegian Geological Survey as an aid in interpreting geochemical analytical results. Woodsworth (1972) makes extensive use of probability plots as the basis for a thorough statistical analysis of about 2000 reconnaissance stream sediment analyses from an exploration program in central British Columbia. Numerous other examples could be cited. None of these papers, however, treats in detail the problem of useful and efficient selection of threshold values.

Threshold is a term used throughout the mineral exploration industry to signify a specific value that effectively separates high and low data values of fundamentally different character that reflect different causes. Commonly, the term is applied to a value that distinguishes an upper or anomalous data set from a lower or background set. For many types of data, particularly those of a geochemical nature, anomalous values are related to mineralized rock. Consequently, the choice of a threshold value has considerable importance in directing exploration to specific anomalous sample sites where the chances of discovery of an economic mineral deposit are greatly enhanced.

Thresholds in geochemical data are chosen in a variety of ways. A method recommended in several publications involves the estimation of the mean and standard deviation of a data set with an arbitrary choice of a threshold at a value corresponding to the mean plus two standard deviations (see Hawkes and Webb, 1962; Lepeltier, 1969). In some cases this procedure might be adequate but it ignores the fact that no a priori reason exists for exactly the upper 2½% of every data set being anomalous. Furthermore, the method does not take into account adequately, the fact that anomalous and background populations have fairly extensive ranges of overlap in some cases, and as they are two populations the mean and standard deviation derived from the whole data set really have no statistical validity and are just numbers. These failings are recognized by many field practitioners who rely on subjective visual examination of histograms of data sets to choose threshold values.

A third approach is to define thresholds at points of maximum curvature in cumulative probability plots (e.g. Woodsworth, 1972). The procedure entails approximating segments of a probability curve by straight lines and picking threshold values at ordinate levels that correspond to intersections of these "linear" segments. At best, this method is approximate, at worst it can result in a high proportion of anomalous values going unrecognized.

Obviously, a procedure is desirable for choosing threshold values that maximizes the likelihood of recognition of anomalous values and minimizes the number of background values included with anomalous data. Cumulative probability plots provide an effective graphical means of meeting these ends.

PROBABILITY PAPER

Arithmetic probability paper is a special kind of commercially available graph paper generally designed with an arithmetic ordinate scale and an unusual abscissa scale of probability (or cumulative frequency percent) arranged such that a normal (gaussian) cumulative distribution plots as a straight line. Lognormal probability paper differs only in that the ordinate scale is logarithmic. Arithmetic values of a single lognormal distribution grouped in exactly the same manner as required for the construction of a cumulative histogram, plot as a straight line on log probability paper. A bimodal distribution consisting of two lognormal populations plots as a curve. Examples of a single lognormal distribution and bimodal lognormal distributions are shown in Fig.1. In these examples, and throughout the remainder

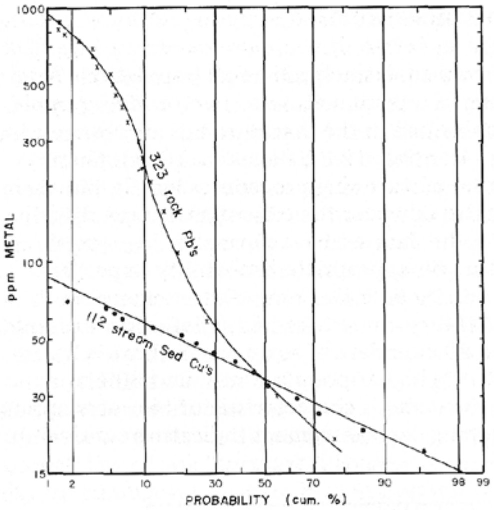

Fig.1. Examples of unimodal and bimodal real distributions plotted on logarithmic probability paper.

of this paper, values are cumulated for plotting by starting at the upper or high value end (cf. Lepeltier, 1969). The probability scale is taken as the abscissa because most commercially available probability paper in North America is arranged in this manner.

There are numerous advantages to probability plots that are worth noting here:

(1) The form of density distribution of a data set can be examined.

(2) Parameters of normal and lognormal populations can be estimated rapidly and with adequate accuracy for most sets of geochemical data.

(3) Several data sets can be represented on a single graph with much greater clarity than multiple histograms.

(4) Plots of several data sets can be compared visually for rapid recognition of similarities or differences.

Additional advantages resulting from the ability to partition polymodal distributions into their individual populations will become apparent in examples presented later. Of course, there are limitations to these plots as well, that must be recognized: (1) data might not have normal or lognormal distributions; (2) construction of a probability graph normally requires a minimum of about 100 values, although techniques are available for dealing with fewer data (see Koch and Link, 1970); (3) scatter of data on a probability plot can be too great to permit a confident analysis of the data.

Despite these limitations a high proportion of geochemical data sets can be analysed usefully and confidently on probability graph paper.

PARTITIONING OF POLYMODAL DISTRIBUTIONS

Partitioning refers to methods used to extract individual populations from a polymodal distribution consisting of a combination of two or more populations. The methods are not well described in the literature but are referred to, or implied by various writers (e.g., Harding, 1949; Bolviken, 1971). Cassie (1954) and Williams (1967) describe partitioning procedures briefly but their publications are not widely available. Consider the case of a bimodal distribution: providing that populations in the data set have normal or lognormal density distributions and are plotted on appropriate probability paper, an estimate of their proportions is given by an inflection point or change in direction of curvature on the probability curve (Harding, 1949). For example, in Fig.2, an inflection point at the 20 cumulative percentile, indicated by an arrow, shows the presence of 20% of a higher population A, and 80% of a lower population B. The form of the curve is characteristic of two overlapping populations, a relatively gently sloping central segment indicating considerable overlap of the two.

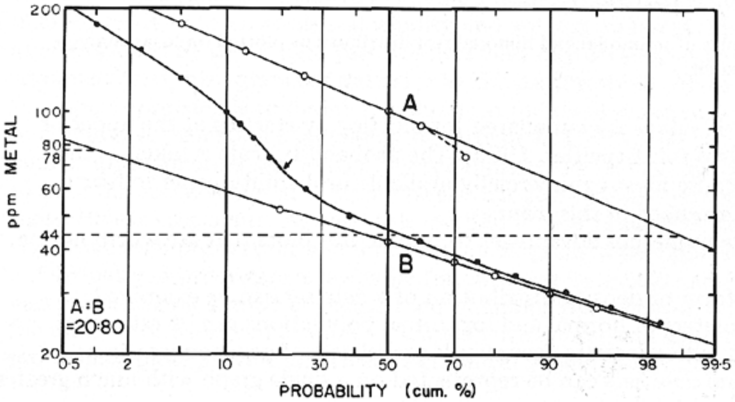

Fig.2. Two idealize hypothetical populations A and B are combined in the proportions A/B = 20/80 to produce the intermediate curved distribution drawn through calculated points shown as solid dots. An inflection point is shown by the arrowhead. Arbitrary thresholds at the 1% level of B population and the 99% level of A population correspond to 78 and 44 ppm, respectively.

The uppermost plotted point on the curve at the 180-ppm ordinate level represents 1% of the total data. However, it also represents (1/20 × 100) = 5 cumulative percent of population A because at this extremity of the data set there is no effective contribution from population B. Consequently, a point on A population is defined at 5 cumulative percent on the 180-ppm level. In the same manner, the point plotted on the curve at the 150 ordinate level represents (2.6/20 × 100) = 13 cumulative percent of population A and a second point on population A is obtained. This procedure is repeated until sufficient points are obtained to define population A by a straight line or until the replotted points begin to depart from a linear pattern indicating that population B is present in significant amounts. When sufficient points are obtained, a line is drawn through them as an estimate of population A.

Population B can be obtained in precisely the same way, providing the probability scale is read as complementary values, e.g., 90 cumulative percent is read as (100 − 90) = 10 cumulative percent. Calculated points for both A and B populations are shown as open circles in Fig.2.

Validity of the two-population model can be checked by combining them in the proportions 20% A and 80% B at various ordinate levels. In this hypothetical example, check points are not shown because it has been constructed ideally. Throughout the remainder of the paper, however, check calculations are indicated by open triangles. The checking procedure involves the calculation of ideal combinations of the partitioned populations at various ordinate levels using the relationship $P_M = f_A P_A + f_B P_B$ where P_M, the probability of the "mixture", is to be calculated (see Bolviken, 1971); P_A and P_B are cumulative probabilities of populations A and B read from the graph at a specified ordinate level; f_A is the proportion of population A, and $f_B = 1 - f_A$ is the proportion of population B. In practice, several trials might be necessary to obtain a good fit of the ideal mixture with the real data because of the difficulty in defining the inflection point accurately. In most cases, the partitioning procedure is as straight forward as outlined. In other cases, a slight modification is necessary when dealing with real data as will become apparent in some of the examples that follow.

Partitioning of polymodal curves containing three or more populations is somewhat more complex but is done in an analogous way, proceeding in stages. Generally, partitioning begins with the populations represented by the extremities of the probability curve, followed by partitioning of more centrally located populations.

Note that in this idealized example, parameters of the individual partitioned populations can be estimated. The geometric mean of each can be read at the 50 percentile and the range including 68% of the values can be determined at the 84 and 16 cumulative percentiles. This range encompassing 2 standard deviations is asymmetric about the geometric mean. The method of representation adopted here is to quote the geometric mean, followed in brackets by the range that includes 68% of the values. These parameters for the partitioned populations A and B are 100 (144, 71) and 42 (55, 33), respectively.

Estimates of the arithmetic mean and variance can be obtained from this information as described by Krumbein and Graybill (1965), but normally are not required.

CHOICE OF THRESHOLDS

The hypothetical example in Fig.2 illustrates a common general situation of high and low populations with an effective range of overlap. If no significant overlap of values existed, the central moderately steep segment of the curve would be nearly vertical and a single threshold could be chosen rapidly at its mid-point. In the general case, however, choice of thresholds is more complex.

Consider 2 thresholds chosen arbitrarily at the 99 and 1 cumulative percentiles of the partitioned populations A and B, respectively of Fig.2 (recall that A and B are present in the ratio A/B = 20/80). These percentiles divide the data into 3 groups at the 44- and 78-ppm ordinate levels. 16% of the total data is above the upper threshold of 78 ppm. In a hypothetical sample of 100 values, this upper group would consist approximately of 15 values from A population and 1 value from B population. The lower group below 44 ppm contains 46% of the total data. It consists of 1% of population A (at most, 1 value in this case) and 57% of population B (about 46 values). The intermediate group between the two thresholds contains about 38% of the total data consisting of 42% of the B population and 33% of the A population. In our hypothetical sample this corresponds to about 6 or 7 A values and 33 or 34 B values (Table I).

TABLE I

	Total data		A population		B population	
	%	No.*	%	No.*	%	No.*
Group I	16	16	76	15.2	1	0.8
Group II	38	38	23	4.6	42	33.6
Group III	46	46	1	0.2	57	45.6
	100	100	100.0	20.0	100	80.0

*Sample = 100 of which 20 are A and 80 are B population.

The procedure, although arbitrary, has thus divided the data rather effectively into three groups, two of which contain significant proportions of the upper A population and a third that almost exclusively represents the lower B population. Let us assume for the moment that A and B represent anomalous and background populations, respectively. The upper group above the upper threshold can be considered top priority for follow up examination because practically all values are anomalous. Lower priority can be attached to values in the intermediate group because although it contains virtually all remaining anomalous values, an increased amount of exploration manpower per anomalous sample is required to check them and sort them out from background values in the same range.

There is nothing sacrosanct about the percentiles used to define thresholds. In this case, values were chosen that corresponded with 99 and 1 cumulative percentiles of the A and B populations, respectively. Thresholds could equally well have been defined by the 98 and 2 cumulative percentiles of the appropriate partitioned populations. Whatever choice is made, it is possible to determine estimates of the proportions of each population occurring in the groups thus delimited. In the writer's experience, the two sets of figures mentioned above have proved most useful but different values could be chosen depending on the nature of the data and the required probability that all anomalous values be retained in the upper two groups.

Note that in this hypothetical but typical case the choice of a threshold at the mean plus two standard deviations would have placed most of the anomalous values with background. The same effect would be obtained with a common variation of this procedure, the assumption that the upper 2½% of values are anomalous. Were the probability curve approximated by three linear segments, their intersections would have provided thresholds at approximately 103 and 55 ppm. The common procedure of adopting the upper value as threshold would result in rejection of more than 50% of the anomalous values. Even the choice of the lower value would result in rejection of about 5% of anomalous values.

Zn IN SOILS, TCHENTLO LAKE AREA (CENTRAL BRITISH COLUMBIA)

Fig.3 is a probability graph of 173 zinc analyses of B horizon soils taken on a grid pattern in an area of known Mo—Cu mineralization near Tchentlo Lake in central British Columbia. Underlying rock is a texturally and mineralogically uniform, well-jointed diorite. Joints are mineralized, principally with quartz and pyrite, but in some places molybdenite is abundant and small amounts of chalcopyrite occur. A thin layer of overburden covers the area except for sporadic outcrop knolls.

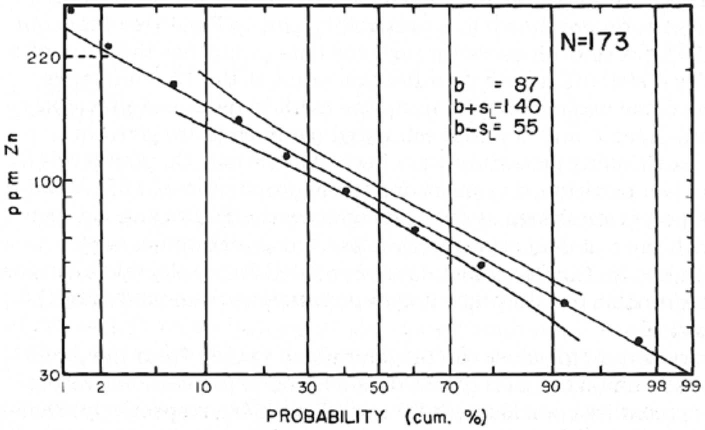

Fig.3. Probability plot of 173 values of Zn in B zone soils, Tchentlo Lake, B.C. Listed parameters of the distribution were obtained from the straight line drawn through original data points (solid dots). 95% confidence limits are shown after Lepeltier (1969).

The probability plot is linear if one neglects slight divergences at the extremities, that commonly result from sampling error. Consequently, an estimate of the distribution can be obtained by a straight line through the plotted points. 95% confidence limits of the population were determined graphically (cf. Lepeltier, 1969). Woodsworth (1972) suggests that a useful procedure for recognizing significant curvature in a probability graph is to assume the presence of a single population and construct its 95% confidence belt. Significant curvature to the plot is assumed at points that plot outside the zone of 95% confidence. None of the plotted points for Tchentlo Lake data lie outside the band defined by the 95% confidence limit suggesting that only a single population is present.

In this case, the range of values and the form of the probability graph suggest that the data represent a single background population. A wise procedure, however, is to assume that the few highest values are anomalous until proven otherwise. This is a convenient safety precaution in cases where anomalous values are present in too low proportion to define a second population. To standardize a procedure for dealing with such data, it is convenient to pick an arbitrary threshold at an ordinate level corresponding to the mean plus 2 standard deviations as recommended by Hawkes and Webb (1962). This procedure assumes that approximately the upper 2½% of values are anomalous until shown otherwise, and should be applied only when a single population is indicated from an examination of the probability graph. In this example, the upper 5 zinc values were found to plot on a plan of the grid, sporadically, but away from known mineralized areas.

Cu IN STREAM SEDIMENTS, MT. NANSEN AREA (YUKON TERRITORY)

Copper analyses for 158 stream sediment samples from the Mt. Nansen area, Yukon Territory, are shown as a probability plot in Fig.4 (see Bianconi and Saagar, 1971). A smooth curve through the data points has the form of a bimodal density distribution with an inflection point at the 15 cumulative percentile. The curve was partitioned using the method described previously to obtain populations A and B whose estimated parameters are given in Table II. The partitioning procedure was checked at various Cu ppm levels by combining the two partitioned populations in the proportion of 15% A and 85% B. Check points are shown as open triangles on the Figure and are seen to coincide with the real data curve. In this case, some high values are associated with known Cu—Mo mineralization related to porphyritic intrusions and it seems reasonable to interpret the two populations as anomalous (A) and background (B).

Two arbitrary threshold values can be determined readily from the graph at the 1.0 and 99 cumulative percentiles of the B and A populations, respectively. These percentiles coincide with 70 and 37 ppm Cu, respectively. Hence, the data are divided into 3 groups, an upper group of predominantly anomalous values, a lower group of predominantly background values, and an intermediate

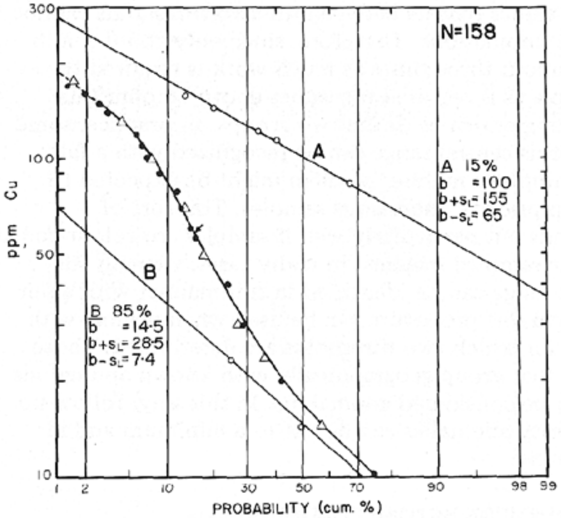

Fig.4. Bimodal probability plot of 158 Cu's in stream sediments, Mt. Nansen, Yukon. Open circles are partitioning points used to establish populations A and B. Open triangles are check points obtained by combining A and B in the ratio 15/85.

TABLE II

Estimated parameters of partitioned populations, Cu in stream sediments, Mt. Nansen area (Yukon Territory)

Population	Proportion (%)	No. of samples	Values in ppm Cu		
			b	$b + s_L$	$b - s_L$
A: anomalous	15	24	101	155	63
B: background	85	134	14.7	28.5	7.4
A + B	100	158			

group containing both anomalous and background values. Of the 158 values, about 23 are anomalous, and 135 are background. 80% or about 18 of the anomalous values are above the 70-ppm threshold; and 5 are below it, for all practical purposes, in the intermediate range. Of the 135 background values, 91.5% or 124 values, are below the lower threshold, the remaining 11 background values are above the lower threshold in the intermediate range.

Consequently, anomalous values occur in only two ppm intervals to which priorities can be assigned for follow up exploration. Virtually all values above 70 ppm are anomalous and have top priority. Second priority is assigned to the 16 values in the intermediate range, about 5 of which are anomalous.

Theoretically, individual values that lie between the two thresholds cannot be assigned to either A or B populations. Therefore, since only about 1 in 3 is anomalous in this range, about three times as much work is required to check each anomalous sample as is required for values above 70 ppm Cu; hence, the reason for assigning priorities to the two groups. In practice, some of the anomalous values in this central range can be recognized with a fair degree of certainty. For example, a number of them might be expected to occur down stream from top priority anomalous samples. This sort of geographic relationship stands out particularly well if samples are colour-coded as to group, on a plan of the sampled streams. In many cases, virtually all samples in the intermediate range can be identified in this manner with a fair degree of certainty. A comparable procedure can be used when dealing with soil or whole rock analyses for which two thresholds are determined. Those intermediate range samples that group geographically with known anomalous samples commonly can also be considered anomalous. In this way, follow-up examination of second priority anomalies can be cut to a minimum and in many cases avoided completely.

Ni IN SOILS, HOPE AREA (SOUTHERN BRITISH COLUMBIA)

Fig.5 is a log probability graph of 166 Ni analyses of soils obtained from a grid superimposed on a known Cu—Ni mineralized zone. The mineral showing is associated with ultramafic rocks enclosed in regionally metamorphosed fine-grained clastic sedimentary rocks, near Hope in southern British Columbia. A smooth curve drawn through the data points has the form of at least three populations based on inflection points at 5.5 and 25 cumulative percentiles. The A and C populations were partitioned using the method described in a previous section. Population B was then estimated using the relationship:

$$P_M = f_A P_A + f_B P_B + f_C P_C$$

In this equation: $f_A = 0.055$, $f_B = 0.195$, $f_C = 0.75$ and P_M, P_A, P_C can be read from the graph for any ordinate level. Hence, P_B is the only unknown and can be estimated for various ordinate levels, plotted, and an estimate of population B determined by passing a straight line through the calculated points. The three partitioned populations A, B and C were then combined ideally in the proportion: 5.5/19.5/75 for a number of ordinate values, to check the partitioning procedure. These check values are shown in Fig.5 as open triangles that almost coincide with the smooth curve through the original data.

Population A is obviously not well defined as indicated by the scatter of points about its linear estimator. The reason is that only a small proportion of the total data represents population A, thus its estimation by partitioning is based on very few data points — four in this case. Populations B and C appear well defined, principally because their ideal combination in the ratio 19.5/75.0 agrees with the real data curve. Estimated parameters of the three

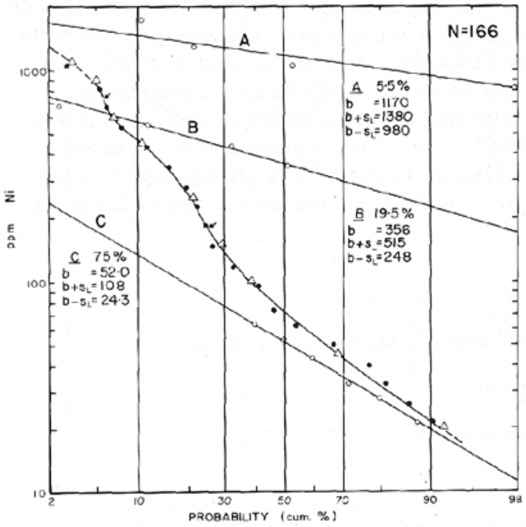

Fig.5. Probability plot of 166 Ni's in soils, Hope, B.C., with 2 inflection points (indicated by arrowheads) suggesting it results from the combination of three lognormal populations in the ratio 5.5/19.5/75. A, B and C are the three partitioned populations estimated by lines through the calculated points (open circles). Parameters of each population are listed. Open triangles are check points that agree well with the original data (black dots).

populations are given in Table III. On the basis of the partitioned populations, a single threshold at 780 ppm Ni can be chosen to distinguish effectively between populations A and B. Populations B and C overlap somewhat and two thresholds must be chosen. These thresholds are arbitrarily taken at the 2 cumulative percentile of population C (i.e. 236 ppm Ni) and the 98 cumulative percentile of population B (i.e. 170 ppm Ni).

TABLE III

Estimated parameters of partitioned populations, Ni in soils, Hope area (southern British Columbia)

Population	Proportion (%)	No. of samples	Values in ppm Ni		
			b	$b + s_L$	$b - s_L$
A: anomalous	5.5	9	1170	1380	980
B: background (ultramafic)	19.5	32	356	515	248
C: background (metaseds)	75	125	52	108	24.5
A + B + C	100	166			

These three threshold values divide the data into 4 groups, 3 of which each consist principally of a single population and a fourth containing values from two populations (Table IV). The thresholds can now be used as contour values on a plan of the grid, or can be used to code data on a plan using colour or symbols, to aid in interpreting the significance of each population. In this case, population A is related to Ni mineralization and is therefore interpreted as an anomalous population. Population B corresponds to areas underlain by ultramafic rocks, and population C occurs in areas underlain by metasedimentary rocks.

TABLE IV

Estimated thresholds, Ni in soils, Hope area (southern British Columbia)

Threshold	Principal content of group
	almost exclusively population A
780	
	almost exclusively population B
236	
	combination of populations B and C
170	
	almost exclusively population C

The choice of thresholds is arbitrary. For example, one could equally well have chosen the two thresholds for the B and C populations at the 1 and 99 cumulative percentile of the C and B populations respectively, or the 2.5 and 97.5 cumulative percentile and so on. . . A choice should be made with the idea of defining a short range of overlap of the two populations, and, at the same time, producing adjacent ranges that to all intents and purposes contain values of a single population, with negligible or minor amounts of other populations.

Cu IN SOILS, SMITHERS AREA (BRITISH COLUMBIA)

A probability plot of 795 soil copper analyses is shown in Fig.6. The sinuous character of the plot is probably real because of the large number of values in the data set. This type of data is characteristic of the sort obtained from reconnaissance surveys where large quantities of information are obtained in a relatively short time. The area sampled is underlain predominantly by acid to intermediate intrusive bodies that cut a thick monotonous sequence of volcanic rocks.

Inflection points are evident at approximately the 1, 2 and 32 cumulative percentiles indicating the presence of at least four populations. These populations can be estimated by partitioning the curve in stages. In this case, it is most convenient to begin with the population C for which most data points

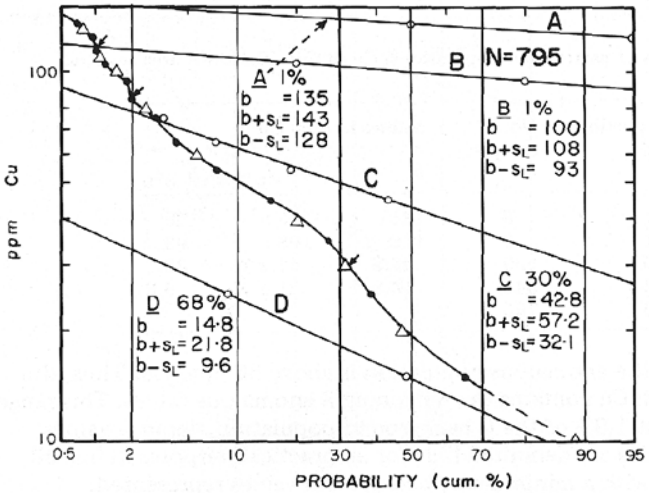

Fig.6. Probability plot of 795 Cu's in B-horizon soils, Smithers area, B.C. Symbols are as defined for Fig.5.

are available. Once C has been defined, population D can be estimated using C and the original data curve. These two populations can be specified reasonably well. The upper two populations A and B can be approximated roughly but cannot be delineated with much precision because of the small percentage of total data that each represents and hence the small number of points available for partitioning. Crude estimates of populations A and B are shown based on the limited data available.

A number of check points, shown as open triangles on the curve were calculated for the partitioned populations A, B, C and D, combined in the ratio 1/1/30/68. These points agree almost perfectly with the smooth curve describing the data, suggesting that the partitioning represents a plausible model for the data. Estimated parameters of partitioned populations are listed in Table V. Comparison of the data with a geological map of the sampled area suggested that populations C and D represent background Cu in soils over volcanic and plutonic rocks, respectively. By the same means, it was concluded that populations A and B are anomalous populations in areas underlain by volcanic and plutonic rocks, respectively.

In choosing thresholds for distinction between anomalous and background values there is no need to consider either population A or D. The critical part of the graph is the range of overlap of populations B and C.

We know that about 2% of the data, or about 16 values are anomalous. Of these, 11 are above 100 ppm Cu as is 1 value of C population. Hence, one of 12 values above 100 ppm Cu is not anomalous and 100 can be chosen as an arbitrary upper threshold.

TABLE V

Estimated parameters of partitioned populations, Cu in soils, Smithers area (central British Columbia)

Population	Proportion (%)	No. of samples	Values in ppm Cu		
			b	$b + s_L$	$b - s_L$
A	1	8	135	145	128
B	1	8	100	108	93
C	30	239	42.8	57.2	32.1
D	68	540	14.8	21.8	9.6
A + B + C + D	100	795			

Virtually all of the anomalous population is above 85 ppm Cu. Thus, the range 85—100 ppm Cu contains the remaining 5 anomalous values. This range also contains about 1.0% of the C background population, about 2 values. Thus, two thresholds are delimited that for all practical purposes define all anomalous values with a minimum of background values represented.

This example illustrates several important points in procedure:

(1) It is wise to carry through with a complete partitioning procedure in examining complex distributions in order to check the realism of the interpretation.

(2) Even when individual populations cannot be defined particularly accurately, thresholds can commonly be determined with adequate accuracy.

(3) Inflection points in a probability curve based on abundant data are probably real and should form a basis for interpretation.

(4) An alternative approach would have been to group the data into two subclasses based on presence of underlying volcanic or plutonic rock. This procedure was not used here only because adequate thresholds could be obtained without spending additional manpower in carrying out a more detailed analysis.

(5) The bottom population, D, is reasonably well known despite the fact its partitioning was based on only two points.

The foregoing examples show that the major advantage of probability plots is to provide a useful grouping of data. Commonly, this grouping is not simply for the purpose of obtaining thresholds between anomalous and background populations — but more generally to derive thresholds between populations that aid in a general interpretation of the significance of the data.

pH MEASUREMENTS OF STREAMS

pH measurements are commonly an integral part of stream sediment surveys. A probability plot of pH values from one such survey in southern British Columbia is shown in Fig.7. The plot is on arithmetic probability paper — a logarithmic transform being incorporated in the original data

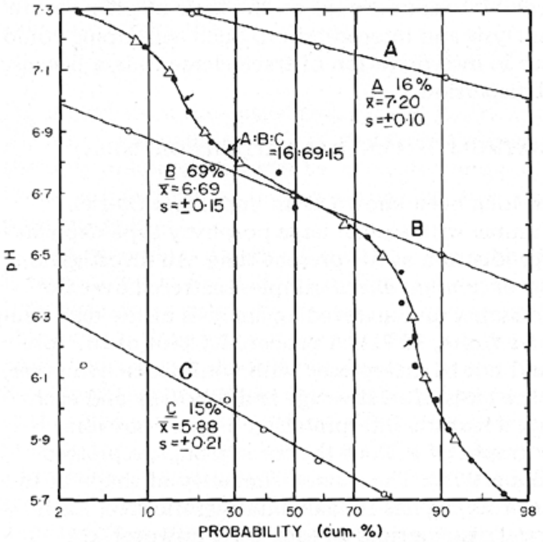

Fig.7. Probability plot of pH values obtained from a stream sediment survey in southern British Columbia. Symbols are as defined for Fig.5.

because of the very nature of pH values. A smooth curve through the data has the form of a trimodal distribution with inflection points at the 16 and 85 cumulative percentiles. The curve has been partitioned using the method described previously to obtain populations A, B and C. Check points based on ideal mixtures of the three populations in the proportion 16/69/15 agree remarkably well with the real data curve.

Thresholds arbitrarily chosen at the 99 cumulative percentiles of A and B populations, and the 1 cumulative percentiles of the B and C populations, provide the information in Table VI.

TABLE VI

Estimated thresholds, pH values (southern British Columbia)

pH		% of total data
	principally population A	15
7.00		
	populations A + B	4.5
6.93		
	principally population B	64.5
6.37		
6.36		
6.35		
	principally population C	16

Thus, the data can be divided into four groups on the basis of pH measurements and prior to further analysis and interpretation. Such a grouping could have fundamental significance in interpretation of trace element data because of the effect of pH on metal dispersion.

WHOLE ROCK Cu, GUICHON BATHOLITH (CENTRAL BRITISH COLUMBIA)

The Guichon batholith has long been known as an important Cu-rich pluton in central British Columbia with several large porphyry-type deposits either producing or nearing production at the present time. An investigation of the whole rock Cu content of *unmineralized* samples scattered over the batholith was undertaken by Brabec and involved an analysis of the data using probability graphs (Brabec and White, 1971). A probability plot of the total data, some 330 analyses, could not be interpreted with confidence. However, when data were grouped on the basis of relative age and lithology and each such group plotted separately, a realistic interpretation became possible.

Fig.8 contains probability graphs of each of the three groups, replotted from data of Brabec and White (1971). The general similarity of shape of the three curves suggests that the grouping has fundamental significance. Each curve has the form of a bimodal distribution. In each case, however, the bottom part of the bimodal curve is partly missing due to the bar interval chosen for construction of the probability plots (15 ppm Cu). Assuming that all distributions are lognormal it is possible to partition each curve using a modification of the procedure described earlier. The upper population can be

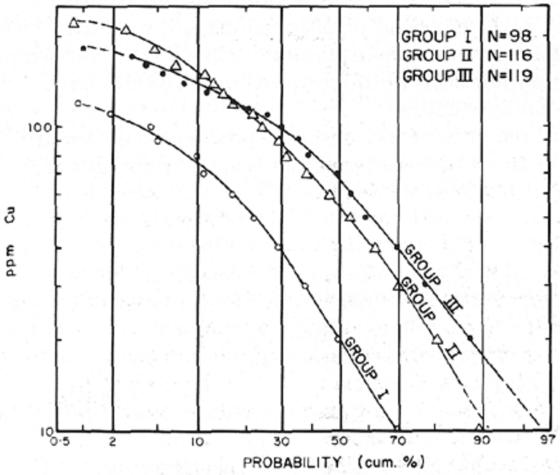

Fig.8. Probability plots of whole rock Cu's for 3 rock groups of the Guichon batholith, central British Columbia. Group I = youngest age, Group II = intermediate age and Group III = oldest age (after Brabec and White, 1971).

determined in the normal manner. Points on the lower population are then calculated using the expression:

$$P_M = f_A P_A + f_B P_B$$

P_M is read from the data curve, f_A and f_B are known from the position of the inflection point and P_A is read from the partitioned population A. P_B is the only unknown and can be calculated and plotted for various ordinate levels. A line can then be passed through these calculated points to estimate population B.

One example is described in detail. The probability plot for group II rocks is reproduced in Fig.9. Some difficulties were encountered in specifying an inflection point precisely, because the two populations overlap to a considerable extent. However, a series of trial values were used until the upper population plotted as a straight line, leading to an inflection being assigned at the 80 cumulative percentile. One additional problem with the data is a flattening at the upper end of the curve. In fact, this flattening is present to some extent in plots for each of the 3 groups and is a characteristic pattern obtained when a symmetric population has been top-truncated. Brabec and White (1971) arbitrarily rejected a small proportion of high values from their analysis to impose this artificial top truncation on their data. Since the truncated values account for only about 2% of the data, no effort was made to correct for their absence. The upper extremities of all curves, however, were ignored during the partitioning.

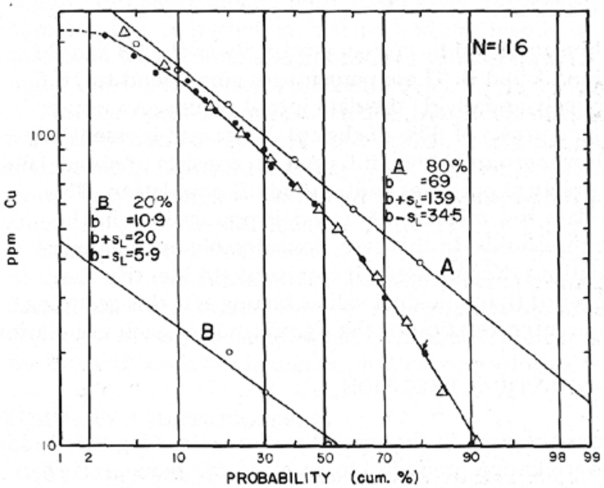

Fig.9. Probability plot of 116 whole rock Cu's in Group II rocks (intermediate age) of the Guichon batholith, central British Columbia, showing partitioned populations and their parameters. Symbols are those defined for Fig.5.

Once the upper population A is defined, population B can be estimated using the relationship:

$$P_M = f_A P_A + f_B P_B$$

as described earlier. Check points of ideal mixtures of partitioned populations A and B, shown as open triangles in Fig.9, coincide with the real data curve except at the upper truncated end. Parameters of the partitioned population for each of the 3 groups are given in Table VII.

TABLE VII

Estimated parameters, whole rock Cu, Guichon batholith (central British Columbia)

Lithologic group	Population	Proportion	No. of samples	Values in ppm Cu		
				b	$b + s_L$	$b - s_L$
I	A	60	56	98	142	68
	B	40	39	26.7	46.4	15.2
	A + B	100	95			
II	A	80	93	69	139	34.5
	B	20	23	10.9	20	5.9
	A + B	100	116			
III	A	40	28	54	85	34.5
	B	60	91	10.3	20.2	5.1
	A + B	100	119			

For group A data thresholds can be chosen arbitrarily as the 98 and 2 cumulative of populations A and B. These percentages correspond to 16.5 and 39 ppm Cu, respectively and divide the data into 3 groups. An upper group above 39 ppm Cu, consists of 63% of the total data and is essentially only A population. A lower group below 16.5 ppm Cu consists of about 16% of the data and for all practical purposes contain only B population. The remaining 21% of the data is a mixture of A and B populations in the range between the two thresholds. In this case, considerable overlap exists between the two populations. Nevertheless, it is possible to identify the population to which most of the individual values belong and this grouping could aid considerably in interpretation of the significance of each population.

THE IMPORTANCE OF ANALYTICAL PRECISION

Thus far, an implicit assumption in the procedure for estimating thresholds is that analytical values are known precisely. In practice, of course, recorded values include a combined sampling and analytical error. Consequently, some values above the threshold actually belong below it and vice versa. Normally this confusion affects only a small proportion of the data, but becomes more and more pronounced as the precision becomes poorer and poorer.

In some cases the confusion is minimal relative to the problem on hand and can be ignored. More generally, however, the sampling and analytical error should be taken into account in defining thresholds. A convenient procedure to achieve this end is to consider the threshold a range of values centred about the single threshold obtained by assuming that values are perfectly known. The threshold range is a confidence belt based on the precision of the data. Average precision is normally adequate for defining such threshold ranges. Precision, however, does vary with absolute amount of the variable being estimated (e.g., Bolviken and Sinding-Larsen, 1973) and this can be taken into account where adequate data are available. Such threshold ranges define narrow bands on contour maps.

This procedure *increases* the number of potentially anomalous samples and therefore involves additional time and money in checking such added samples. These efforts can be minimized by examining the geographic positions of the additional samples relative to known anomalous samples.

DISCUSSION

The method for choosing thresholds described here is a standardized technique applicable to the vast quantity of geochemical data. It can be used for any polymodal distribution if sufficient data of adequate quality are present so that partitioning is feasible. A grouping of the data values is obtained that can be invaluable in interpretation. For this reason, the method is more fundamental and potentially more useful than other methods in common use. In particular, the method outlined here stresses the concept that both background and anomalous values represent populations that in many cases overlap (see Bolviken, 1971).

The procedure is not restricted to the choice of thresholds between anomalous and background populations. It is much more general in nature, permitting grouping of many types of data with appropriate density distributions. In addition, probability graph analysis of data is simple, rapid and amenable to use in the field (see Lepeltier, 1969).

Examples used to illustrate the selection of thresholds give ample evidence of the general usefulness of probability plots in dealing with geochemical data. This is true even if three or four populations are represented in the data, although, in general, simpler interpretations result if data are first grouped on the basis of some fundamental physical or geological criterion.

SUMMARY AND CONCLUSIONS

(1) Geochemical analyses commonly approximate lognormal density distribution sufficiently closely that the distributions can be represented usefully on lognormal probability paper.

(2) Providing a data set contains adequate values, normally a minimum of about 100, a polymodal cumulative probability plot can be partitioned to produce estimates of the individual populations that make up the total distribution.

(3) The partitioned populations can be used to define arbitrary but meaningful thresholds that divide the data into groups that have fundamental significance.

(4) In the special case of no effective overlap between anomalous and background populations, a single threshold can be defined. In the common simple case of two overlapping anomalous and background populations, two thresholds are obtained that divide the data into three groups. An upper group of predominantly anomalous values, a central group of anomalous *and* background values, and a third group of background values.

(5) Polymodal distributions of geochemical data consisting of more than two populations can commonly be treated in the same way as bimodal distributions to obtain useful threshold values. In some cases, however, the procedure can be simplified by grouping data on the basis of some fundamental characteristic (e.g., pH, underlying rock type) to produce simpler probability plots that permit greater confidence in partitioning and interpreting.

(6) The method described for choosing thresholds is not confined to the distinction between anomalous and background values but has general application to any type of data, providing the individual populations approximate lognormal (or normal) density distribution. Fortunately, this criterion is met in the bulk geochemical data.

ACKNOWLEDGEMENTS

This paper is an outgrowth of a more extensive study of the use of probability paper in dealing with various kinds of data obtained from mineral exploration programs. The study is supported by a grant from the Department of Energy, Mines and Resources of Canada. Technical assistance through much of the study was given by Mr. A.C.L. Fox. Examples are drawn entirely from real problems encountered in industry and in university research projects. Appreciation is expressed to the numerous individuals and companies involved for permission to publish them. Dr. W.K. Fletcher offered constructive criticism of an earlier draft of the paper.

REFERENCES

Bianconi, F. and Saagar, R., 1971. Reconnaissance mineral exploration in the Yukon Territory, Canada. Schweiz. Mineral. Pet. Mitt., 51:139—154.

Bolviken, B., 1971. A statistical approach to the problem of interpretation in geochemical prospecting. Can. Inst. Min. Metall., Spec. Vol., 11:564—567.

Bolviken, B. and Sinding-Larson, R., 1973. Total error and other criteria in the interpretation of stream sediment data. In: M.L. Jones (Editor), Geochemical Exploration, 1972. Institute for Mining and Metallurgy, London, pp.285—295.

Brabec, D. and White, W.H., 1971. Distribution of copper and zinc in rocks of the Guichon Creek batholith, British Columbia. Can. Inst. Min. Metall., Spec. Vol., 11:291—297.

Cassie, R.M., 1954. Some uses of probability paper in the analysis of size frequency distributions. Aust. J. Mar. Freshwater Res., 5:513—523.

Harding, J.P., 1949. The use of probability paper for the graphical analysis of polymodal frequency distributions. J. Mar. Biol. Assoc., U.K., 28:141—153.

Hawkes, H.E. and Webb, J.S., 1962. Geochemistry in Mineral Exploration. Harper and Row, New York, N.Y., 415 pp.

Koch, G.S., Jr. and Link, R.F., 1970. Statistical Analysis of Geological Data, Vol.I. John Wiley and Sons, New York, N.Y., 375 pp.

Krumbein, W.C. and Graybill, F.A., 1965. An Introduction to Statistical Models in Geology. McGraw-Hill, New York, N.Y., 475 pp.

Lepeltier, C., 1969. A simplified statistical treatment of geochemical data by graphical representation. Econ. Geol., 64:538—550.

Tennant, C.B. and White, M.L., 1959. Study of the distribution of some geochemical data. Econ. Geol., 54:1281—1290.

Williams, X.K., 1967. Statistics in the interpretation of geochemical data. N.Z. J. Geol. Geophys., 10:771—797.

Woodsworth, G.J., 1972. A geochemical drainage survey and its implications for metallogenesis, central Coast Mountains, British Columbia. Econ. Geol., 68:1104—1120.