CLAUDE LEPELTIER

A SIMPLIFIED STATISTICAL TREATMENT OF GEOCHEMICAL DATA BY GRAPHICAL REPRESENTATION

Economic Geology
Vol. 64, 1969, pp. 538–550

A Simplified Statistical Treatment of Geochemical Data by Graphical Representation[1]

CLAUDE LEPELTIER

Abstract

In the course of a mineral exploration sponsored by the United Nations Development Programme in two selected zones of Guatemala, a stream sediment reconnaissance was carried out, and graphical methods of interpretation were attempted in the search for a simplified statistical treatment of about 25,000 geochemical results. The data were grouped by drainage and lithological units, and the frequency distributions of the abundance of Cu, Pb, Zn and Mo were studied in the form of cumulative frequency curves. The four elements appear to be approximately lognormally distributed. Background, coefficients of deviation and threshold levels were graphically estimated. Examples are given of simple and complex populations. Mineral associations were studied by correlation diagrams.

Contents

Introduction

THE United Nations Mineral Exploration Programme in Guatemala relied heavily on geochemical prospecting. During one year (1967) 60 percent of the total Project area was covered systematically by a geochemical reconnaissance carried out in the drainage systems. Nine thousand stream sediment samples were collected over about 12,000 km² (rounded figures). All the samples were analyzed for copper and zinc, and the total number thinned out to approximately 4,000 before being run for lead and molybdenum. Finally about 25,000 geochemical results were available for compilation and interpretation. As they accumulated, it became apparent that high-contrast anomalies which are obvious targets for follow-up operations would not be encountered but rather more subtle features not so easy to pinpoint and interpret.

The interpretation phase of the survey was characterized by two essential features: the great amount of data to be analyzed and the lack of precision of these data.

Sampling and analytical methods must sacrifice precision for speed due to the nature of geochemical prospecting, and the first consequence of this fact is that an isolated result has little meaning in geochemistry. It must be part of a population as numerous and homogeneous as possible. Indeed in all kinds of phenomena, individual inaccuracies shade off progressively when observation is extended to larger and larger populations.

The first phase of geochemical interpretation is to condense large masses of numerical data and extract from them the essential information. The most objective and reliable way to do it (and sometimes the only one) is statistically. Large sets of numbers, cumbersome and difficult to interpret, may be reduced to a useful form by the use of descriptive statistics. This is best done by the graphical representation of the frequency distribution of a given set of data; then the average value, an expression of the degree of variation around the average, and the limit above which the anomalies start are immediately and precisely determined as well as the existence of one or several populations in the surveyed area.

This treatment of the data also simplifies the comparison of the geochemical behavior of an element in various geological surroundings or of several elements in the same lithological unit.

I am grateful to Mr. Henry H. Meyer, Project Manager of the Guatemala and El Salvador Mineral

[1] This article is published with the authorization of the United Nations. The opinions expressed are not necessarily endorsed by this Organization.

Surveys, and to Mr. Stephen S. Steinhauser for technical criticism and much helpful discussion.

Difficulty of Statistical Approach in Stream Sediment Surveys

A reliable statistical interpretation requires that a great quantity of data be treated and that these data be homogeneous.

In drainage reconnaissance surveys, the first condition is easily filled but not the second. As a matter of fact, the importance of sampling technique is sometimes overlooked in this type of prospecting. But even if given the appropriate attention, too many types of rivers and too many lithological units are generally sampled to result in a homogeneous collection of samples. The best way to limit the inconvenience of the heterogeneity of the samples (particularly pH, organic content and grain size) is to split the survey area into drainages and lithological units, when possible, and to make the statistical interpretation for each of them separately. However, even if this is done, the same degree of precision cannot be achieved as in the case of a soil survey where good homogeneity is possible.

Adjustment to a Lognormal Distribution

Definitions

When dealing with a large mass of geochemical data, the first step is to find what sort of distribution pattern best fits the various sets of observations. And, thus far, the lognormal distribution pattern appears to be the one most applicable to the results of most geochemical surveys (Ahrens, 1957).

In geochemical prospecting, we study the content of trace elements in various natural materials, and to say that the values are lognormally distributed means that the logarithms of these values are distributed following a normal law (or Gauss' law) well known as the bell-shaped curve (Monjallon, 1963).

Many natural or economic phenomena can be expressed by a value varying between zero and infinity, represented by a skewed distribution curve. If, instead of the actual value of the variable itself, we plot its logarithm in abscissae, the frequency curve takes a symmetrical, bell-shaped form, typical of the normal distribution. This happens when a phenomenon is subject to a proportional effect, that is to say when independent initial causes of variations of the studied value take effect in a multiplicative way. It is the case, for instance, for the distribution of trace elements in rocks, for the area of the different countries of the world, for the income of individuals in a country, for the grain size in samples of sedimentary rocks, and others (Coulomb, 1959; Cousins, 1956).

In all these examples, the character studied follows the lognormal law, which is probably more common than the normal one.

It is interesting to note here that the lognormal law fits very well in the case of low-grade deposits like gold but for high-grade deposits, iron for instance, the experimental distributions are generally negatively skewed because of the limitation towards the high values. G. Matheron gives a thermodynamic interpretation of the proportional effect in the case of ore deposits and relates it to the Mass Action Law (Matheron, 1962). To the extent in which geochemical anomalies are extrapolations of ore deposits this theory should apply to geochemical prospecting.

Construction of the Cumulative Frequency Curve

A lognormal distribution curve is defined by two parameters: one dependent on the mean value, and the other dependent on the character of value-distribution. This latter parameter is a measure of the range of distribution of values, that is whether the distribution covers a wide or narrow range of values. The two parameters can be determined graphically as will be explained on following pages. For practical purposes, we work on cumulative frequency curves, and their construction shall be explained by means of a concrete example.

The various steps of this construction are the following:

(a) Selection of a precise set of data ("population") as large and homogeneous as possible.

(b) Grouping of the values into an adequate number of classes.

(c) Calculating the frequency of occurrence in each class and plotting it against the class limits; this gives a diagram called the "histogram."

(d) Smoothing the histogram to get the frequency curve.

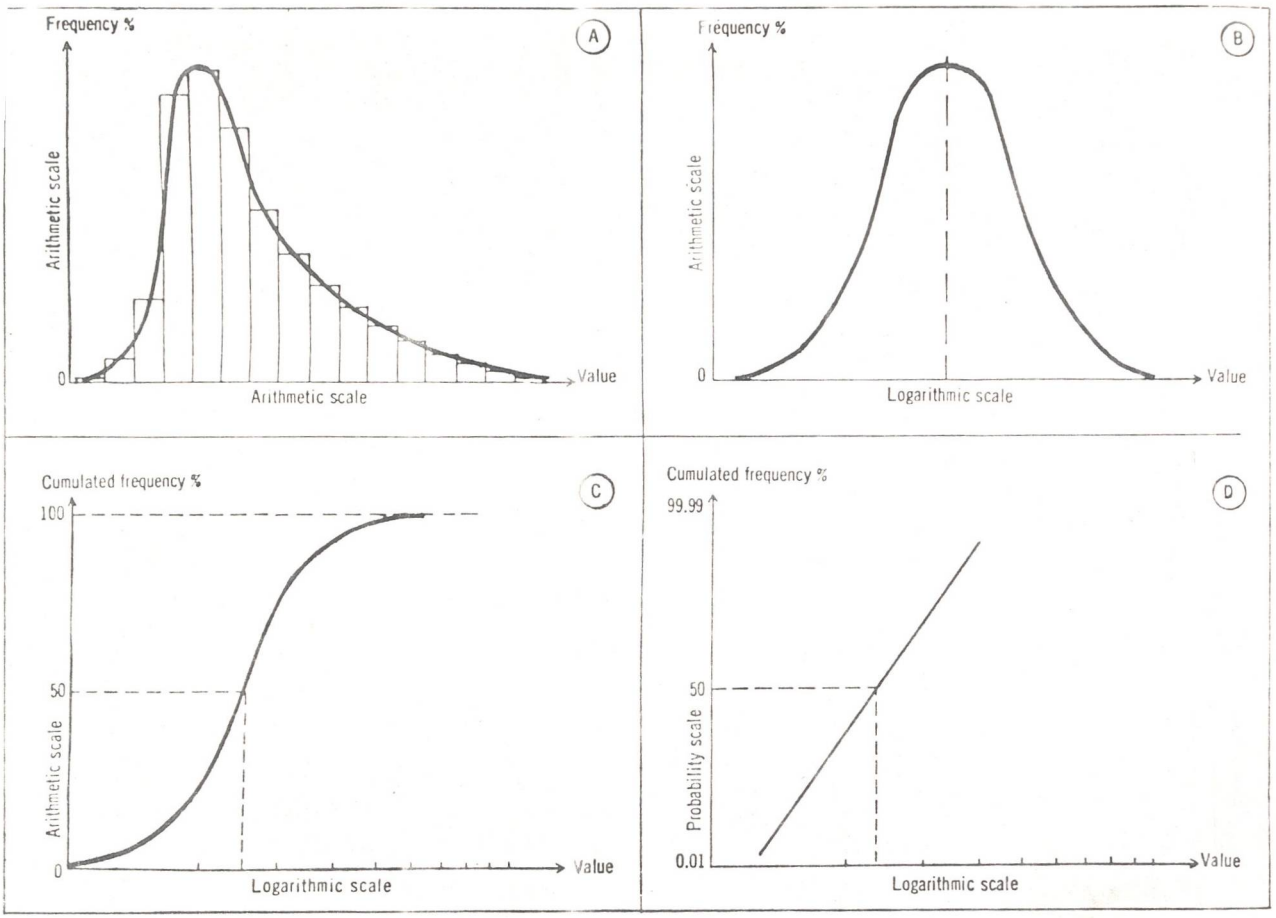
(e) Plotting the cumulated frequencies as ordinates gives the cumulative frequency curve, which is the integral of the frequency curve.

(f) By replacing the arithmetic ordinate scale with a probability scale the cumulative frequency curve is represented by one or more straight lines. Examples of lognormal frequency curves are shown in Figure 1.

Some brief comments on the different steps follow:

(a) The larger the population to be analyzed, the more precise and reliable the results. If necessary, as few as 50 values may be treated statistically but

Figure 1. Logno mal distribution curves



the confidence limits must be calculated to see if the analysis is meaningful.

(b) A correct grouping of the values is mandatory if some precision is to be achieved in the statistical interpretation; too few classes will result in shading out important features of the curve; too many in losing significant details amidst a cloud of erratic ones. The results are distributed in classes, the modulus of which should be proportional to the precision of the analyses: the more precise the analyses, the smaller the modulus. The logarithmic interval must be adapted to the variation amplitude of the values and to the precision of the analytical methods (Miesh, 1967).

In statistics, working with 15 to 25 intervals (or classes) is recommended. As a rule, the width of a class, expressed logarithmically, must be kept equal to or smaller than half of standard deviation (Shaw, 1964).

For geochemical purposes, it is convenient to work with 10 to 20 points on the cumulative frequency line, that is to say with 9 to 19 intervals or classes. There are three variables to consider: the number of points (n) necessary to construct a correct line; the range of distribution of the values (R), expressed as the ratio of the highest to the lowest value of the population; and the width of the classes expressed logarithmically (log. int.) which has to be selected in function of the two first parameters. These three variables are linked by the relation:

$$\text{log. int.} = \frac{\log R}{n}$$

In most of the cases R varies from 6 to 300 (experimental average values), then, with (n) varying from 10 to 20, log R from 0.78 to 2.48, the extreme values for the logarithmic interval will be:

$$\text{log. int.} = \frac{0.78}{20} = 0.039$$

$$\text{log. int.} = \frac{2.48}{10} = 0.25$$

The 0.10 was selected as the best suited logarithmic interval for the classes because it suits most distri-

bution, giving reasonable number of classes and a good definition of the curve. In case of very reduced dispersion of the values around the mean, it may be necessary to use 0.05, and if the dispersion is specially large, 0.2 will be chosen. When the logarithmic interval is selected, it is easy to calculate a table giving the class limits in ppm. The only precaution is to avoid starting with a round value so that no analytical results will fall on the limit of two classes. The most useful and commonly employed in geochemical work is the 0.1 log. int. classs table, a part of which is given below:

class limit (log) . . 0.07, 0.17, 0.27, 0.37, 0.47, 0.57
class limit (ppm) . 1.17, 1.48, 1.86, 2.34, 2.95, 3.72

It can be extended in both directions as far as necessary.

(c-d) After selecting the class table, the values are grouped and the frequency calculated for each class (in percentage); then the frequencies are plotted against the class limits (the latter being logarithmically calculated, ordinary arithmetic-arithmetic paper must be used), giving a histogram which is smoothed to a frequency curve. But histograms are often misleading, being strongly affected by slight changes in class intervals, and frequency curves are difficult to draw and handle: for instance, it is necessary to determine the inflexion points of the curve in order to evaluate the standard deviation. Practically, the histogram-frequency curve step is skipped and the cumulative frequency directly constructed. However, note here an advantage of the histogram: it clearly illustrates the effect of the sensitivity of the analytical method and more precisely the bias brought to the low values by the use of colorimetric scales of standards. As a matter of fact, experience shows that there is an inevitable concentration of the readings, whoever the analyst, on the values actually represented in the colorimetric scale. For instance, in the case of copper, the lower part of the standard colorimetric scale reads 0,2,4,7 . . . ppm. Usually this results in an excess of 2, 4 and 7 values, and a conspicuous lack of 1, 3, 5 ppm values. This is of importance for a correct construction of the frequency curve, and the raw values must often be corrected by extrapolating the general shape of the curve.

(e-f) By plotting the cumulated frequencies as ordinates instead of the frequencies, one obtains the integral curve of the preceding. It has the form of a straight line when using the appropriate graphpaper (probability-log), and it is the one used in geochemical presentation and interpretation of the results. Then two questions have to be answered: where to start accumulating the frequencies, and where to plot the cumulated frequencies?

As for the first point, the normal procedure followed by many authors is to start cumulating the

Figure 2 - Cumulative Frequency Distribution for Zn and Cu

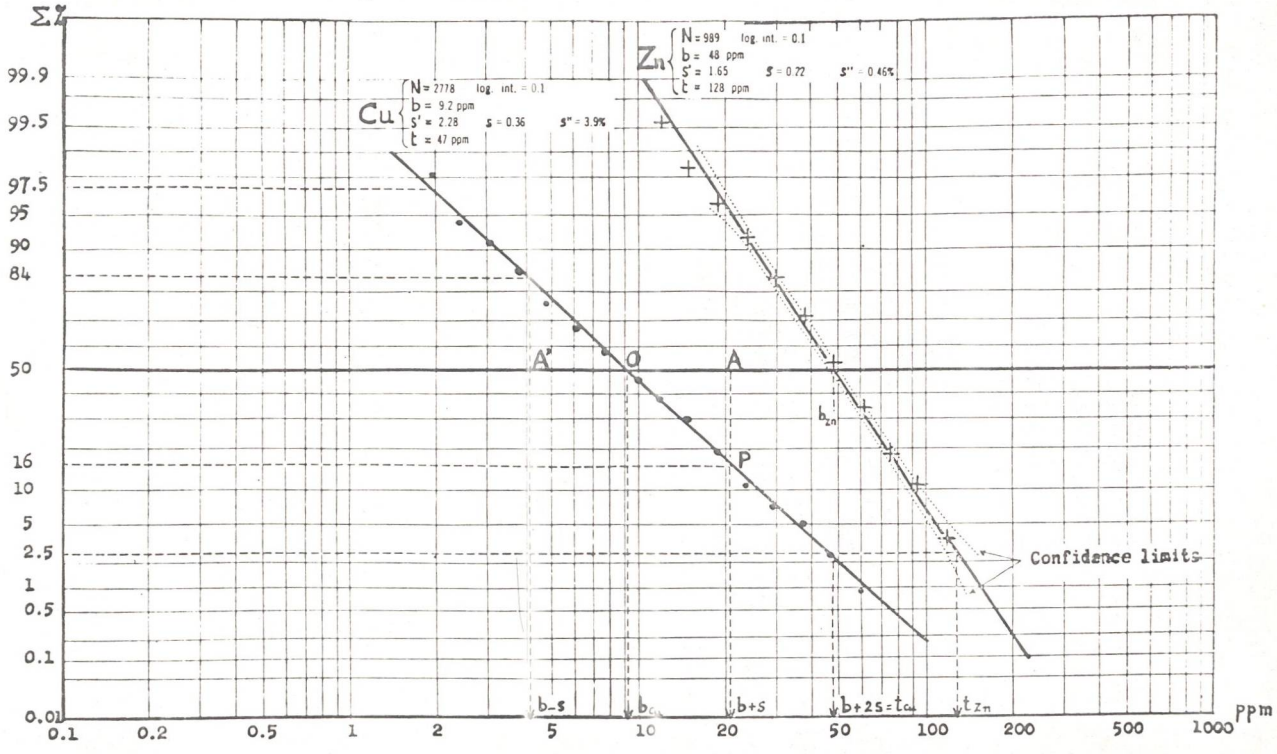
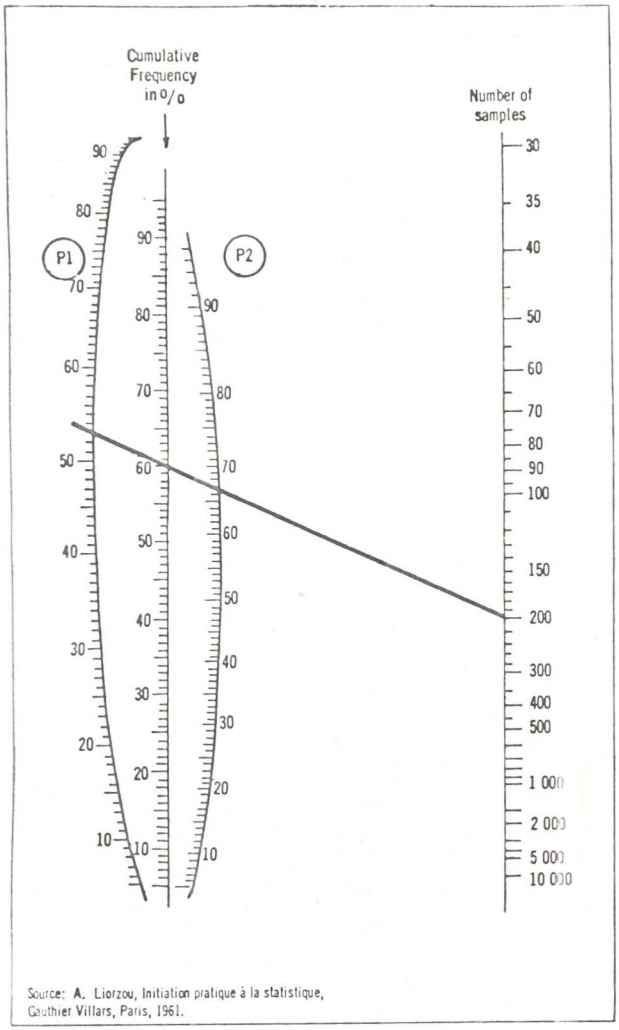


Figure 3. Confidence limits (P_1, P_2) at 0.05 probability level



Source: A. Liorzou, Initiation pratique à la statistique, Gauthier Villars, Paris, 1961.

frequencies from the lowest values toward the highest (Fig. 1) (Hubaux, 1961; Tennant and White, 1959). However, one has to consider a property of the probability scale used as ordinates: the values zero and 100% are rejected at the infinite; it does not matter for zero because zero% never occurs, but in each case the last cumulated frequency is 100%, and this value is impossible to plot, lost for the curve. Then considering the lack of precision in the low values and the importance of the high ones for the determination of the threshold level, I consider it much better to *cumulate the frequencies from the highest to the lowest values;* thus, the 100% will correspond to the lowest class and be eliminated.

As for the second point, the curve being an integral one, the ordinates must be plotted at class limits and not at class center; then, since one cumulates the frequencies from the highest values to the lowest, cumulated frequencies are to be plotted against the lower class limits. Using the class center will entail an error of excess on the central tendency parameters (background and threshold) but not on the dispersion parameter (coefficient of deviation). This error, or difference, varies with the type of classes used and is easily calculated (6% for the 0.05 logarithmic class interval, 12% for the 0.1 log. int. and 26% for the 0.2 log. int.). If the class limit is used, curves constructed from different log. int. classes can be directly compared without correction.

Let us take a concrete example: the distribution of Zn in the quaternary alluvial deposits of Block I (Fig. 2). There are 989 results ranging from 10 to 230 ppm.

$$\text{population: } N = 989 \quad \text{range: } R = \frac{230}{10} = 23$$

The best class interval is selected as explained above:

$$\text{log. int.} = \frac{\log R}{n} = \frac{1.36}{14} = 0.097$$

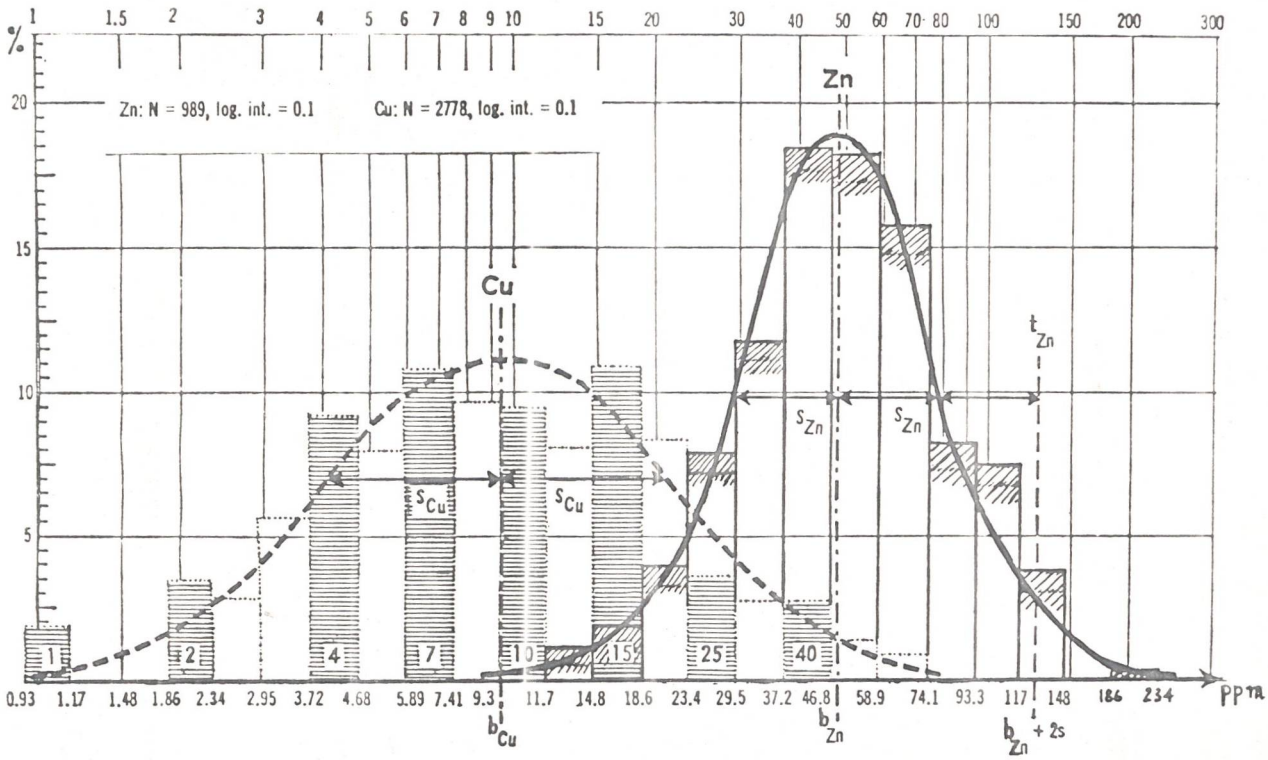
A 0.1 log. interval will give 14 intervals, which is acceptable. Usually, the histogram-frequency curve step is skipped and the cumulative frequency diagram directly constructed.

In Figure 2, the points fit fairly well along a straight line, suggesting a lognormal distribution of zinc in the alluvial deposits. Actually, the points never fit the line exactly, but this does not matter provided they stay in a channel delimited by the confidence limits usually taken at the 5% probability level. This confidence interval has been drawn on Figure 2 by using a graph (Fig. 3), which avoids fastidious calculation and gives a fairly good precision for the cumulative frequency values between 5% and 95%. The width of the confidence channel is inversely proportional to the importance of the population considered: the bigger the population, the narrower the confidence interval. To check that a distribution fits a lognormal pattern, one should use the Pearson's test (Rodionov, 1965; Vistelius, 1960), but this longer operation is generally not warranted in this type of interpretation and, for practical purposes, the graphical control described above is satisfactory.

Comparison with Histograms

For comparison purposes the cumulative frequency curve for Cu in the Motagua drainage (Fig. 2) was also constructed, then, in Figure 4, the corresponding histograms and frequency curves for Cu and Zn. Figures 2 and 4 present the same data in two different ways. Before enumerating and com-

Figure 4. Histogram and frequency curve for Zn and Cu



menting on the advantages of the former presentation over the latter, an interesting feature of the histogram should be mentioned: in the case of colorimetric determinations made in the lower range of sensitivity of the analytical method, the histogram shows clearly the bias introduced in the readings by the human factor and by the accuracy and sensitivity limits of the method. This effect is illustrated for copper in Figure 4, where the classes including a colorimetric standard are shaded and the value of the standard itself is given as a larger figure (1, 2, 4 . . . ppm); the cumulation of the frequency reduces this effect, particularly if it is started from the high values, but it may be necessary to bring some corrections to the low value frequencies in order to construct a precise distribution curve.

Comparing Figures 2 and 4, one sees immediately that it is easier to compare two straight lines than two overlapping bell-shaped curves; many more populations can be presented on the same diagram by using cumulative frequency curves than by using histograms. Cumulative frequency curves are of easier construction and more precise than ordinary frequency curves; it is simpler to draw a line that fits a set of points than to draw a bell-shaped curve with inflexion points.

Information Given by Cumulative Frequency Curves

The main purpose in constructing the cumulative frequency curve for a given population is to check if it fits a lognormal distribution, and if it does, to estimate graphically its basic parameters: background (b), coefficients of deviation (s, s', s'') and threshold level (t).

(b) gives an idea of the average concentration level of the elements in a given surrounding.

(s) expresses the scatter of the values around (b): it corresponds to the spread of the values and their range, from the lowest to the highest.

(t) is a complex notion which might be termed "conditional": statistically it depends on the probability level chosen; geologically, and for practical purposes, it is supposed to be the upper limit of the fluctuations of (b): it depends on (b) and (s). The values equal to or higher than (t) are considered anomalous.

Adjustment to the lognormal law is generally the case when soil samples are considered: in the drainage reconnaissance survey in Guatemala, we found that trace element contents in stream sediments appear also to be lognormally distributed.

Background

A straight line denotes a single population lognormally distributed. In this simple case, the background value (b) is given by the intersection of the line with the 50% ordinate. In the examples given in Figure 2, we have:

background value for copper .. b (Cu) = 9.2 ppm
background value for zinc .... b (Zn) = 48 ppm

Of course, these values must be rounded off; it will be illusory to imply a precision far out of reach of the analytical methods. In the illustrated example, 10 and 50 ppm are taken as reasonably good approximations of the background levels.

In the case of a perfect frequency distribution curve, the background thus calculated corresponds to the mode (most frequent) and median (50% of the values above, 50% below it) values, and is *the geometric mean* of the results. This geometric mean is a more significant value that the arithmetic mean. It is also a more stable statistic, less subject to change with the addition of new data and less affected by high values.

Deviation

Before explaining how to determine graphically the deviation coefficient, an essential property of the normal distribution (i.e., fitting the "bell-shaped" curve) must be recalled here:

(b) being the median value and (s) the standard deviation then:

68.26% of the population falls between $b - s$ and $b + s$
95.44% of the population falls between $b - 2s$ and $b + 2s$
99.74% of the population falls between $b - 3s$ and $b + 3s$

This holds true in the case of the lognormal distribution since the logarithms of the values are normally distributed. Then, rounding off the above-mentioned percentages and taking (b) as the background, we can say that 68% of the population falls between $b - s$ and $b + s$ or that 32% is outside these limits. The distribution curve being symetrical around an axis of abscissa (b) (Fig. 4), 16% of the values will fall above $b + s$ and 16% below $b - s$. In Figure 2, the values $b + s$ and $b - s$ will be obtained by projecting the intersection of the distribution line with the ordinates 16 and 84% on the abscissa axis. Working with logarithms, one has to consider the ratios and not the absolute values thus established. Taking the same example of Cu (Fig. 2), one determines the points P (at the 16% ordinate) and A. OA is the geometrical expression of the deviation: it is inversely proportional to the slope of the line. We call it the *geometric deviation* (s'); it has no dimension: *it is a factor* obtained by dividing the value read in A by the value read in O:

$$s' = \frac{21}{9.2} = 2.28$$

Then multiplying or dividing the background value by the geometric deviation will give the upper and lower limits of a range including 68% of the population (from $b - s$ to $b + s$, or $A'A$ on the figure). Multiplying or dividing by the square of the geometric deviation gives a range including about 95% of the values ($b - 2s$ to $b + 2s$).

Because all the reasoning is made on logarithms, it is also necessary to express the deviation by a logarithm: the *coefficient of deviation* (s) is the logarithm (base 10) of the geometric deviation (s').

$$s' = 2.28$$

$$s = \log s' = 0.36$$

It will be seen later that it might be interesting to consider a third deviation index: the *relative deviation* (s'') sometimes called *coefficient of variation*. It is expressed as a percentage:

$$s'' = 100\,\frac{s}{b}$$

$$s'' = 100\,\frac{0.36}{9.2} = 3.9\%$$

Threshold

After the background and the coefficient of deviation, the third important parameter is the threshold level (t), which is a function of the two former. It has been seen that in the case of symmetrical distribution (either normal or lognormal) 95% of the individual values fall between $b + 2s$ and $b - 2s$, that is to say that only 2.5% of the population exceeds the upper limit $b + 2s$. This upper limit is conventionally taken as the threshold level (t) above which the values are considered as anomalies:

$$\log t = (\log b) + 2s$$

or to avoid using logarithms:

$$t = b \times s'^2$$

$$t = 9.2 \times 5.2 = 47.8 \text{ ppm}$$

Practically, (t) as well as (b), is read directly on the graph as the abscissa of the intersection of the distribution line with the 2.5% ordinate. In this example one reads 47 ppm, and the slight difference is due to the rounding off of the exact

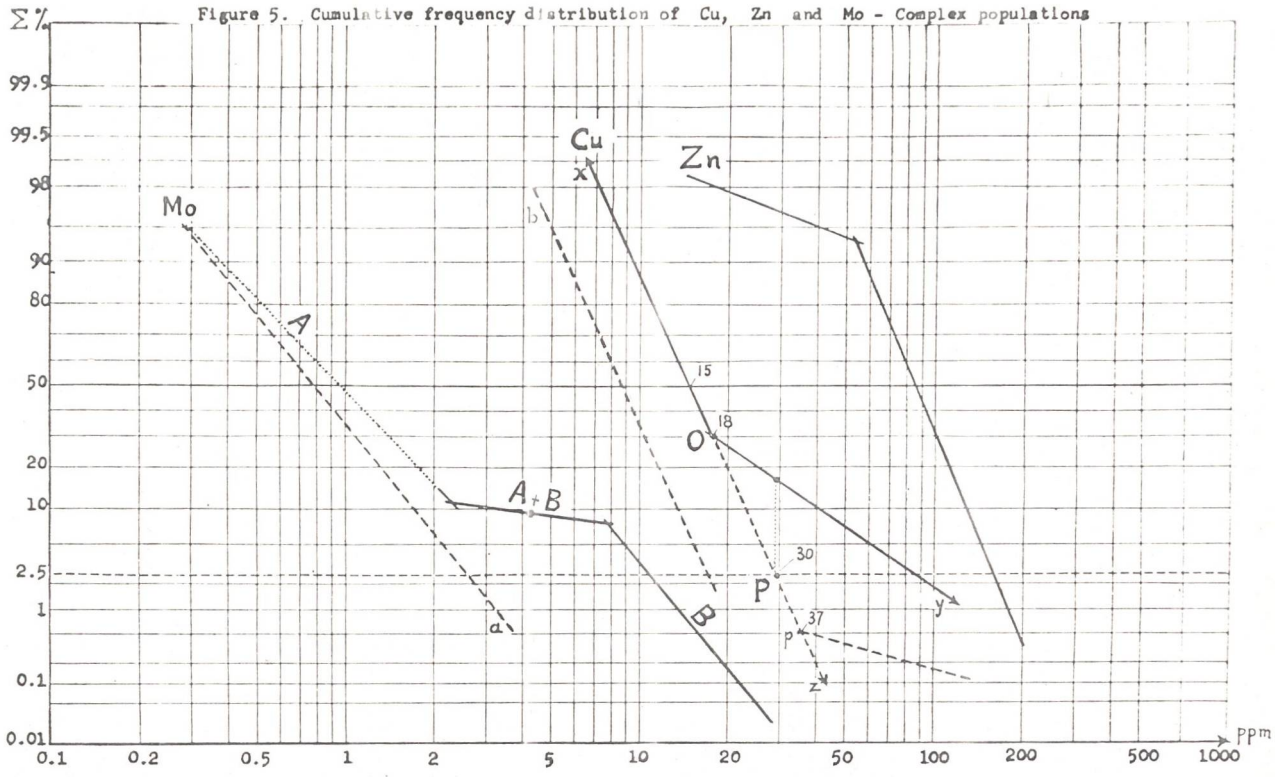


Figure 5. Cumulative frequency distribution of Cu, Zn and Mo - Complex populations

ordinate 2.28% to 2.5%. This shows the importance of the deviation in the estimation of the threshold; two populations may have the same background but, nevertheless, different thresholds if their coefficients of deviation are different. In Figure 2, the threshold is five times the background for Cu and only 2.7 times for Zn.

In all the foregoing, I have considered the simplest case: a single lognormal population, the diagrammatic expression of which is a straight line. However, when constructing cumulative frequency curves, a broken line is frequently obtained suggesting that the set of data considered consists of a complex population or of different ones. Whenever possible in practice, the interpretation is made on sets of data selected so as not to include more than two different distributions; for instance, a lithological unit may include two types of mineralization showing up in soil or sediment samples; one representative of the normal or background content of the material sampled, and the other, a superimposed mineralization related to ore.

Examples

The three main cases of non-homogeneous distribution that are the most likely to occur are, in decreasing frequency order:

a. an excess of high values in the considered population;

b. a mixture of two populations in a given set of data; and

c. an excess of low values in the considered population.

These three cases are represented graphically in Figures 5. They correspond to real distributions encounetred in the Guatemalan drainage survey and appear as solid lines with slope breaks on the diagram. Some indications are given below showing how to interpret such lines.

Copper Distribution (in a lithological unit). The cumulative frequency line (Fig. 5) shows a break to a flatter slope at the 30% level. This is the case when there is an excess of high values in the population; the histogram will give a frequency curve skewed to the right, in the direction of the high values (positive skewness). If the population was lognormally distributed, the main branch Ox should extend as a straight line in Oz whereas, in this case, Oz is lifted to Oy which means that instead of having 2.5% of the values 30 ppm or greater, there are 17% of them. The abscissa of the breaking point, O, (in this case 18 ppm) indicates the limit above which there is a departure from the norm (i.e., from the lognormal distribution), an excess of high values. In this case, background and coefficients of deviation are calculated with the main branch Ox. The abscissa of the breaking point may be conveniently

taken as threshold value if the break occurs above the normal threshold level of 2.5%. If, however, the break occurs below 2.5% level (at point p for instance) the threshold should be taken as usual (abscissa of point P). Positively broken distribution lines are the more interesting because they indicate an excess over the background mineralization.

Molybdenum Distribution (in a lithological unit). The cumulative distribution line shows two breaks: first a positive, then a negative one. Such a graph is the expression of a dual distribution, suggesting the existence of two distinct populations in the set of data considered. It gives a double-peaked histogram. We shall consider here only the most frequent case of a main "background" population mixed with a smaller one of higher average value, the two of them being lognormally distributed. On the diagram (Fig. 5), branch A corresponds to the main or normal population, branch B to the anomalous population and the central branch $A + B$ to a mixture of the two. By splitting the data at a value taken around the middle of $A + B$ (at 4 ppm for instance), it is possible to separate the total population into two elementary ones appearing as a and b on the diagram. The general background will be taken with branch A and the threshold as the abscissa of the middle of branch $A + B$, though the threshold of population a may also be considered, but we have not enough examples of such complex distributions to make definite recommendations, and we lacked computing facilities to calculate theoretical distributions. The coefficients of deviation must be calculated separately for distributions a and b.

Zinc Distribution (in a drainage unit). The negatively broken line on Figure 5 is the expression of an excess of low values in an essentially lognormal distribution; in this case, the histogram is skewed to the left, toward these low values (negative skewness). Provided their proportion is not too high (20% or less or instance), they do not interfere in the interpretation, which is done on the main branch of the distribution line in the usual way. This excess of low values may be due to the inclusion in the population of a low-background lithological unit or, more often, to poor sampling (for instance, collecting an important set of sediment samples that are too coarse).

When the results do not fit a lognormal distribution, an explanation may generally be found among these three factors: (1) lack of homogeneity in sampling, (2) complex geology (imprecision in the lithological boundaries), and (3) analytical errors.

It should also be kept in mind that some elements in some surroundings may not be lognormally distributed.

Advantages of Cumulative Frequency Curves

Plotting the distribution of an element in a selected unit as cumulative frequency curve on probability graph paper is the easiest and most precise way to present a great amount of data (for instance, presenting Figure 5 as histograms and frequency curves will result in an overloaded and illegible diagram). All the characteristic parameters of the distribution can be estimated without cumbersome calculations. Comparison between various populations are easy and complex distributions are clearly identified. Furthermore, the adjustment to a lognormal distribution can be checked graphically.

Comparing the geochemical features of the various units of a survey area is important in assessing their mineral potential. This is conveniently done by plotting the corresponding distributions on the same diagram—for instance Cu distribution in three or four different drainages in the case of a stream sediment reconnaissance. Distribution heterogeneities will be spotted and the corresponding units selected for further investigations. On a broader scale, the geochemical behavior of trace elements in a given geological environment from different countries or metallogenic provinces can be readily compared. This is an approach to a better understanding of the distribution laws of trace elements in naturally occurring materials.

The Coefficients of Deviation

A lognormal distribution is completely determined by two parameters: the geometric mean (b) and the coefficient of deviation (s). It has been seen that the *absolute deviation* can be expressed as a geometric factor s' or, more commonly, as a logarithmic coefficient s. The term "deviation" is preferred to "dispersion" which might be more expressive, because there is no genetic implication in the concept of statistical dispersion whereas there is one in the notion of geochemical dispersion; however, many people use the term "dispersion" in statistical interpretation of geochemical data.

The coefficient of deviation is a dispersion index specific for the distribution of a given element in a given environment and expresses the degree of homogeneity of this distribution. When rocks are considered, a similarity in the coefficient of deviation, together with similar average values, may indicate similar geochemical processes in their formation.

It is possible that a given value of s corresponds to each type of mineralization in a lithological unit. Confirming this assumption would require very extensive geological-statistical studies encompassing all metallogenic cases.

There is also a relationship between the background (b) and the coefficient of deviations (s)

which is the expression of the geochemical law which states that the dispersion of an element is inversely proportional to its abundance. This is expressed very clearly by the relative dispersion s'' (or relative deviation), a percentage related to b and s as follows:

$$s'' = 100 \frac{s}{b}$$

The higher the background, the lower the relative deviation. This is best shown on a log/log correlation diagram by plotting s'' as abscissa and b as ordinate. Figure 6, for instance, shows the variation of s'' in function of b in the different lithological units of Blocks I and II, for Cu, Zn, Pb and Mo. The diagram has been constructed by taking, for each element, the extreme values for b and s'', thus determining parallelograms including all the individual values. One sees immediately that there is an inverse linear relationship between b and s'' (which is evident from the definition of s'') and that the average absolute deviation s (graphically estimated in Fig. 6) also decreases when the abundance of the element increases.

The weighted mean values of b, s and s'' for each element have been calculated separately for Blocks I and II:

Block I	b	s	s''	Block II	b	s	s''
Zn	55.	0.23	0.42	Zn	70.	0.17	0.24
Cu	8.	0.34	4.2	Cu	8.	0.30	3.8
Pb	6.8	0.32	4.7	Pb	5.8	0.30	5.2
Mo	0.38	0.37	97.5	Mo	0.35	0.40	125.

The fact that the absolute deviation for Pb is equal to or slightly lower than that for copper is due to two factors: (1) the sensitivity limit of the analytical method for lead, which entailed a number of assumptions and extrapolations in the interpretation—determination of b and s, and (2) the existence of some Pb mineralized zones in the survey area where b was high and s low.

Figure 6. Correlation diagram b/s" for blocks I and II

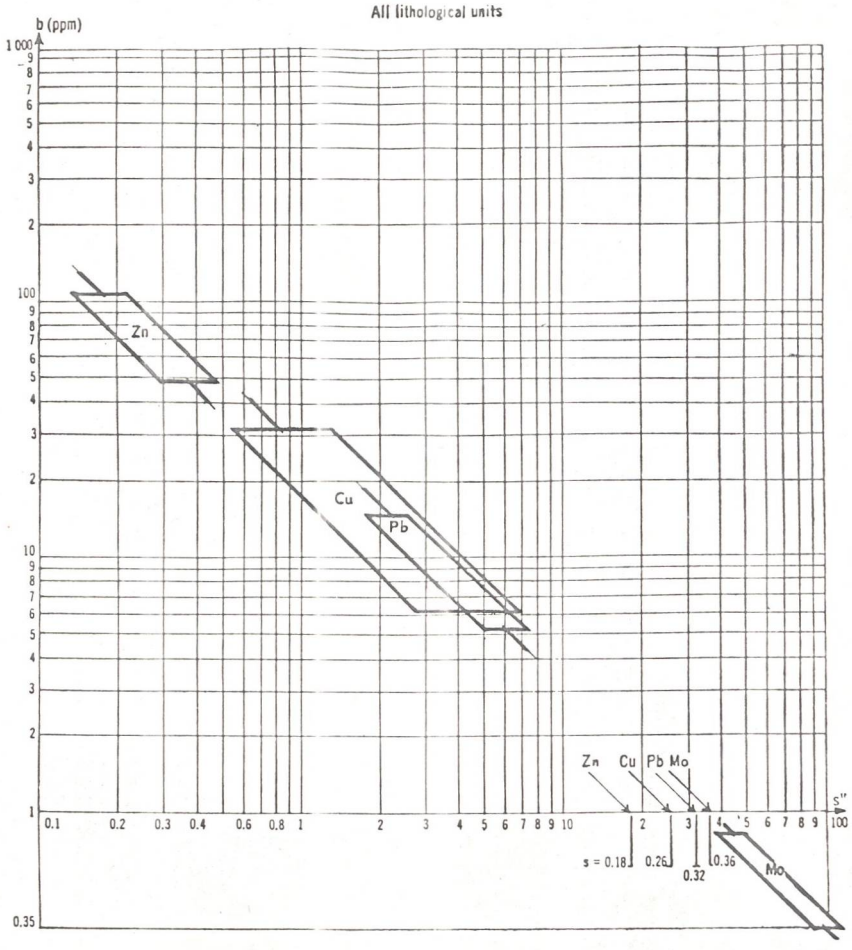
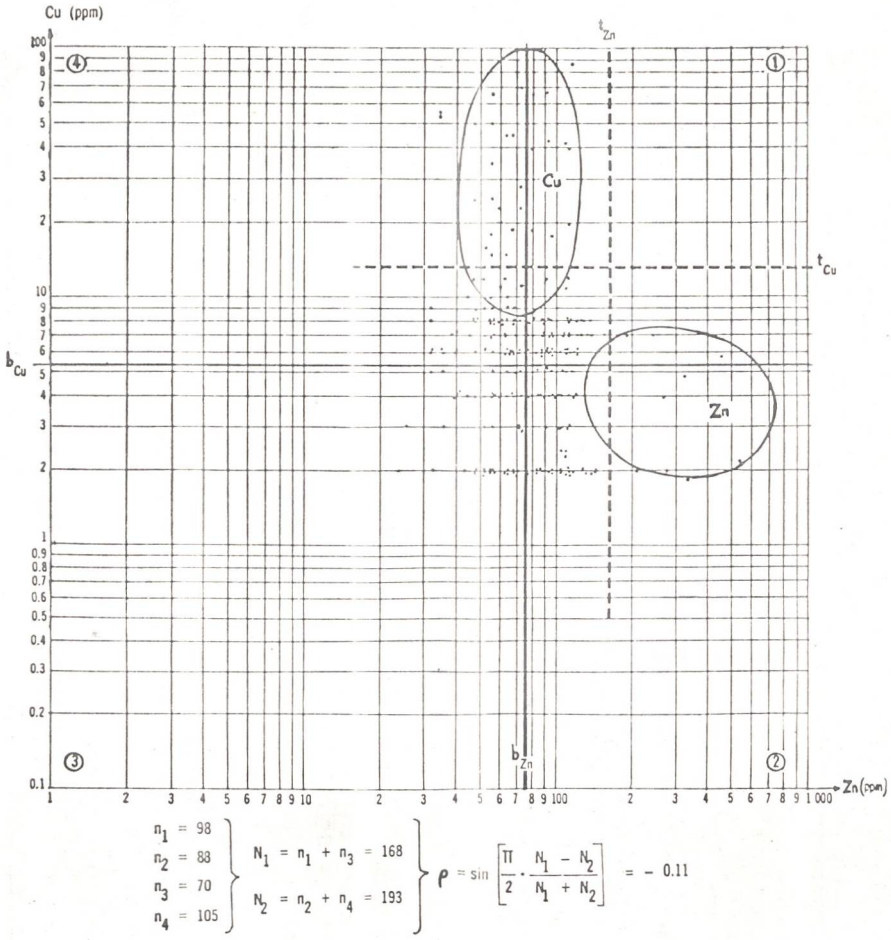


Figure 7. Correlation diagram Cu/Zn



In Figure 6, it is also interesting to note the variations of the dispersion of the same element in different lithological units which is particularly noticeable for copper; the width of each parallelogram indicates the range of variation of s for each element.

The coefficient of deviation is a very important character of the distribution of an element in a given surrounding; it is probably related to the type of geochemical dispersion, mechanical or chemical, and consequently might give an indication of the type of anomaly encountered: syngenetic or epigenetic. It appears that a higher coefficient of deviation indicates a preponderantly mechanical dispersion, but this has not been proved. Much remains to be done in this field.

Correlation Diagrams

In the case of a polymetallic mineralization, with two or more elements lognormally distributed, there is generally a positive correlation between them; for instance between lead and zinc, a sample high in Pb is commonly also high in Zn. This geologic concept of a relationship between two types of mineralization (only qualitative and rather vague) may be substituted by a precise factor, the coefficient of correlation ρ, which gives a rigorous measure of their degree of dependency. In the case of geochemical prospecting, ρ measures the degree of dependency of two lognormal variables namely the tenors of two elements in a sample population (Matheron, 1962).

The coefficient ρ always falls between -1 and $+1$. $\rho = 0$ means a complete independence between the two elements, $\rho = \pm 1$ indicates a functional relationship, direct or inverse, between them (it is a linear relationship between the logarithms of the tenors).

Simplified Calculation of ρ.—There is a graphical way to estimate ρ, slightly less precise but much faster than the complete statistical calculation: constructing a *correlation cloud* in full log. coordinates (Fig. 7, 8). Each sample of the population under study is plotted following its two coordinates: its tenor in element A and its tenor in element B and the total population appears as a cloud of points.

Practically, this presentation of the data is very convenient because it gives a geometric image of the distribution laws. The axes passing by the gravity center (b_A, b_B), that is to say by the point whose coordinates are the background values for the two considered elements, are then drawn. In Figure 7, the axes will pass through the point ($b_{Cu} = 5.3$ ppm, $b_{Zn} = 75$ ppm). The points falling in each quadrant are summed up and counted as follows:

N_1 = number of points in first and third quadrants
N_2 = number of points in second and fourth quadrants.

Then ρ is given by the formula:

$$\rho = \sin\left[\frac{\pi}{2} \cdot \frac{N_1 - N_2}{N_1 + N_2}\right]$$

Practically, ρ is never equal to ± 1 (which would be the case if all the points were on a straight line) and the points form an elliptical cloud. Two cases may happen:

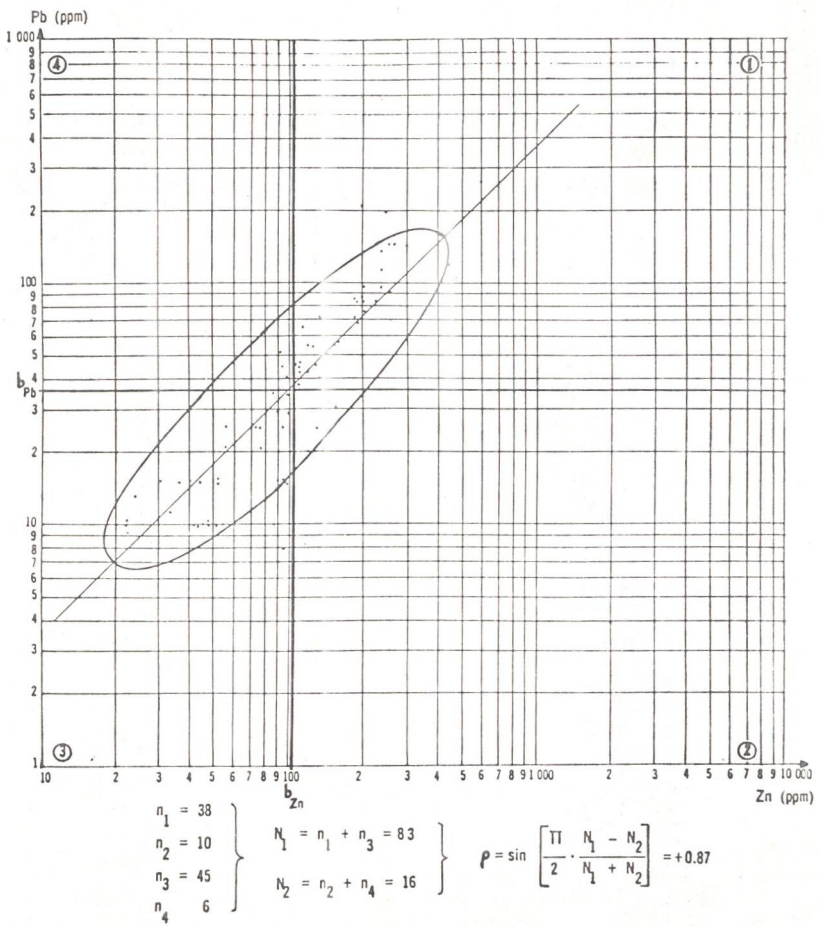
(1) either ρ is equal or near to zero: the elliptical cloud has its axes parallel to the coordinate axes and the two variables are independent,

(2) or ρ is clearly different from zero and the cloud is an ellipse whose axes are inclined relative to the coordinates. The slope of the main axis has the same sign as ρ (if $\rho > 0$ the two elements vary in the same direction; if $\rho < 0$ the two elements vary inversely).

The correlation cloud is in fact a two dimensional histogram; it is the best and simplest way to establish whether a population is homogeneous or heterogeneous: in the first case, the points tend to group in a single elliptical cloud; in the second, they split into 2 or several attraction centers and form several elliptical clouds more or less overlapping. G. Matheron points out that the relation expressed by ρ is an expression of the Mass Action Law if $\rho = \pm 1$ (or of the order of ± 0.95) (Matheron, 1962); then it is likely that a geologically based chemical equilibrium exists between the two elements considered.

In geochemical prospecting, correlation coefficients

Figure 8. Correlation diagram Pb/Zn



may be used to assess mineral associations of elements in natural samples. The correlation diagram shows whether two elements are spatially associated and if one may be used as a pathfinder for the other.

Let us consider two examples: the relationship of Cu/Zn in the drainage of the Suchiate River (Fig. 7) and the relationship of Pb/Zn in the Rio Grande drainage (Fig. 8).

The first example, in Figure 7, is intended only to illustrate the lack of relationship between two types of mineralization. The cloud of points has no definite shape, but it can be divided into three zones: one around the intersection point of the axes, including the majority of the points which are spread more or less equally among the four quadrants; an elliptical one, marked Cu, in the range high-Cu/background-Zn values; and a third one, including only a few high-Zn/background-Cu points. This shows that, in the Suchiate drainage, there is no relationship whatsoever between the Cu and Zn mineralization, that the Cu anomaly is more important than that for Zn and that the two anomalies are well separated spatially. All this is expressed by the coefficient of correlation:

$$\rho = -0.11$$

Its low absolute value indicates a nearly complete independence of the two mineralizations, with a tendency to inverse relationship (negative value).

On the contrary, Figure 8 shows an example of direct relationship between two types of mineralization. In the Rio Grande drainage, Pb and Zn are associated: the correlation cloud is an elongated ellipse whose main axis has a 45° slope and the correlation coefficient $\rho = +0.87$. In this drainage, lead and zinc anomalies will have the same pattern and will be spatially related. In similar geological conditions, one element may be used as a pathfinder for the other.

Conclusion

In the Guatemalan geochemical reconnaissance, the statistical analysis of the data, although elementary, was useful in outlining subdued anomalous patterns in a complex geochemical surrounding, but much more information can certainly be extracted from the analytical results by a more thorough, computer-oriented, treatment.

The graphical methods described above have the great advantage of being quick, cheap and easy to use in the field without any special mathematical knowledge. It is a convenient and synthetic way to present a great amount of geochemical data, and I think it might be useful to any geologist involved in geochemical prospecting.

United Nations Mineral Survey,
Guatemala City, Guatemala,
January 20; March 28, 1969

REFERENCES

Ahrens, L. H., 1957, The lognormal distribution of the elements—a fundamental law of geochemistry: Geochim. et Cosmochim. Acta, v. 11, no. 4.

Coulomb, R., 1959, Contribution à la Géochimie de l'uranium dans les granites intrusifs: Rapport C.E.A. 1173, Centre d'Etudes Nucléaires de Saclay, France.

Cousins, C. A., 1956, The value distribution of economic minerals with special reference to the Witwatersrand Gold Reefs: Geol. Soc. South Africa Trans. v. LIX.

Hubaux, A., 1961, Représentation graphique des distributions d'oligo-éléments: Ann. Soc. Géol. Belgique, T. LXXXIV—Mars 1961.

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

Matheron, G., 1962, Traité de géostatistique appliquée, tome 1: Mémoire no. 14 du Bureau de Recherches Géologiques et Minières, Paris.

Miesh, A. T., 1967, Methods of computation for estimating geochemical abundance—U. S. Geological Survey Professional Paper 574-B.

Monjallon, A., 1963, Introduction à la méthode statistique: Vuibert, Paris.

Rodionov, D. A., 1965, Distribution functions of the elements and mineral contents of igneous rocks: Consultant Bureau, New York.

Shaw, D. M., 1964, Interprétation géochimique des éléments en trace dans les roches cristallines: Masson et Cie, Paris.

Vistelius, A. B., 1960, The skew frequency distributions and fundamental law of the geochemical processes: Journal of Geol. Jan. 1960.