PPlot, a webapp to partition geochemical data and isolate mixed subpopulations using probability plot modeling

PPlot, um aplicativo web para analisar dados geoquímicos e isolar subpopulações através da modelagem de gráficos de probabilidade

Francisco Ferreira de Campos[1]*
Otavio Augusto Boni Licht[2]
Nivaldo Benedito Ferreira Campos[3]

[1] Serviço Geológico do Brasil – SGB/CPRM
Rua Costa, 55
São Paulo SP Brazil
CEP 01304-010

[2] Universidade Federal do Paraná
Programa de Pós Graduação em Geologia
Av. Cel. Francisco H. dos Santos, 210
Curitiba PR Brazil
CEP 81530-900

[3] Universidade Federal do ABC
Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas
Alameda da Universidade s/nº
São Bernardo do Campo SP Brazil
CEP 09606-045

*Corresponding author
francisco.campos@sgb.gov.br

DOI:10.21715/GB2358-2812.202337002

RESUMO

Os métodos estatísticos são, em sua maioria, adequados para lidar com conjuntos de dados formados por uma única população normal ou log-normal, mas dados geoquímicos e geofísicos geralmente não atendem esse requisito. Isso se dá pela heterogeneidade na ocorrência dos objetos geológicos, de forma que o conjunto de dados completo pode ser formado pela mistura de diversas subpopulações. Especificamente, essa mistura de diversas subpopulações pode se referir às diferenças entre áreas mineralizadas e estéreis, ou diferentes fácies geoquímicas de uma unidade geológica, ou ainda entre áreas contaminadas e não contaminadas. Isso implica numa restrição no uso de estimadores estatísticos, tanto os clássicos ou os robustos, a menos que essas subpopulações presentes no conjunto possam ser identificadas e isoladas. O gráfico de probabilidade pode ser usado para analisar um conjunto de dados e inferir uma possível combinação de subpopulações, normais ou log-normais, cuja mistura pode gerá-lo. O aplicativo online PPlot, apresentado neste artigo, permite a construção do gráfico de probabilide de um conjunto de dados e a modelagem das subpopulações presentes nele, tanto de forma automática quanto manual. Após a modelagem do conjunto de dados pelo aplicativo, o usuário obterá resultados numéricos e gráficos dos intervalos de valores que delimitam cada subpopulação, bem como a média e desvio-padrão de cada uma delas. Para validar tanto o procedimento estatístico quanto o código de progração desenvolvido foram usados conjuntos de dados reais e fictícios, e um exemplo de uso do app é apresentado. O aplicativo foi desenvolvido utilizando HTML5 e Javascript e pode ser executado em qualquer navegador moderno e está disponível para uso livre em https://pplotweb.firebaseapp.com/.

Palavras-Chave: mapeamento geoquímico, populações multimodais, mistura de subpopulações, isolando subpopulações, gráfico de probabilidade

ABSTRACT

Statistical methods are mostly designed to handle datasets comprising statistically single normal or log-normal populations, but geochemical and geophysical surveys usually deviate from this expectation. A reason for this is the heterogeneity in the occurrence of geological objects, so the complete dataset may correspond to multiple mixed subpopulations. Specifically, multiple mixed subpopulations can refer to differences between mineralized and barren areas, different geochemical facies of a geological unit, or contaminated and healthy areas. This implies a restriction on using classical or even robust statistical estimates, unless the underlying subpopulations can be extracted from the dataset. The probability plot can be used to assess a dataset and to infer a possible combination of subpopulations, either normal or log-normal, whose combination may generate it. The web-based app PPlot, presented in this paper, allows the plotting of the probability plot of a dataset and modeling the underlying subpopulations present in it, either automatically or manually. After modeling the dataset by the application, the user will obtain numerical results and plots of the range of values that delimit each subpopulation, as well as the mean and standard deviation for each of them. Computer-generated and real datasets were used to validate the procedure and coding, and an example of usage is presented. The app was developed using HTML5 and JavaScript and it runs in any modern browser, and is freely available in https://pplotweb.firebaseapp.com/.

Keywords: geochemical mapping, multimodal populations, mixed subpopulations, isolating subpopulations, probability plot

1 INTRODUCTION

Measures and determinations of numerous quantitative variables are made during geological mapping and geophysical or geochemical surveys, aiming to characterize the differences and contrasting zones from the geological background of the studied area. These differences can be outlined in a wide range of situations and scales of observation, from detailed mapping of crystals for studying their geochemical zoning under electronic scanning microscopy to large structures as seen in geochemical or geophysical regional surveys.

Today, with the analytical facilities and the low cost of determining a wide range of chemical elements and compounds, as well as geophysical measurements, exploration databases have become increasingly larger and more complex. Accompanying this complexity, the availability of data processing computer applications and packages has disseminated sophisticated statistical techniques *e.g.,* discriminant, cluster, factor, and principal component analysis. Many researchers adopt these sophisticated techniques surpassing a previous and fundamental step of recognizing their variables by applying simple but powerful techniques, housed in the field of exploratory data analysis (EDA). Further, it is important to consider that in the real world, advanced techniques for processing geochemical or geophysical data are not in the domain or comprehension of most exploration geologists, being restricted to the academic environment or geologists with advanced statistical expertise. During the handling of geochemical data, it is a very common procedure that stratification of the database is made using as reference the domains or units defined in the geological map. This *a priori* selection of allegedly homogeneous sets can lead to misconceptions that may have undesirable impacts on the results. For this reason, it is necessary to delimit statistically homogeneous domains in the geochemical and geophysical datasets, which may or may not coincide with the geological map.

This paper aims to introduce the free and interactive web application PPlot, available at https://pplotweb.firebaseapp.com/. It is a very simple to use but highly effective data analysis tool in the data handling for delimitation of statistically homogeneous domains. This is achieved by partioning naturally mixed subpopulations in geochemical and geophysical databases using the method described by Sinclair (1974a). Thus, it is possible to divide a dataset into subsets with the necessary statistical support, avoiding the adoption of arbitrary assumptions.

It is clear that the technique of isolating mixed subpopulations, like most of those that compose the classical and robust statistics toolbox, does not consider the spatial component of the sampling stations or the neighborhood between points. The construction of satisfactory and reliable geochemical maps will be achieved with the application of the results obtained with the separation of the subpopulations and their respective statistical estimates.

2 BACKGROUND OF THE TECHNIQUE

Pearson (1894) discussed the meaning of symmetric and asymmetric distribution curves in natural data. He stated that curves whose shape was very close to the symmetry and normal curves constituted the majority of cases. In some cases, however, he observed a well-marked deviation from the normal curve. This asymmetry could happen when the units grouped in the dataset under analysis were not homogeneous. Thus, he studied not only the process of transforming an abnormal frequency curve into a normal curve but also the problem of "*Given an asymmetrical frequency-curve to break it up, if possible, into two component probability-curves, or into two normal curves"* (p. 76). The mathematical solution was so complex that he stated, "*the majority* [of the relations] *lead to an exponential equation, the solution of which seems more beyond the wit of man than that of a numerical equation even of the ninth order"* (p. 75-76).

Seeking practical solutions that were applicable to concrete problems, Hazen (1913) first used probability plots to analyze water flow in rivers; Rissik (1942), Doust and Josephs (1941) used them in engineering applications and in the industry. These authors, however, used probability plots only to analyze unimodal curves, making no mention of their application to the analysis and break-up in the case of bi- and multimodal distributions (HARDING, 1949).

To our best knowledge, Harding (1949) was the pioneer in applying the "*Hazen's*

probability graph paper" for handling complex data sets. He used biological data, justifying it as follows: "*He* [the biologist] *has long been aware that the mean and standard deviation of populations he is confronted with are often of little biological significance; because these populations are compounded of individuals belonging to the two sexes, to different species, or to different age-groups, and are therefore necessarily bimodal or polymodal in character.*" (p. 142). The technique applied by Harding (1949) gave the direction to the subsequent authors who perfected the graphical handling of uni-, bi- and multimodal distributions, mostly of geochemical datasets.

Tennant and White (1959) examined the behavior of geochemical data using a logarithmic probability graph paper. In most cases the results suggested that more than one distribution was present. On logarithmic probability graph paper, a single log-normal distribution gives a single straight line, however on these geochemical datasets, the expected "straight" lines sometimes showed breaks. The breaks suggest that more than one log-normal distribution occurs in the data. The authors assumed that the presence of two straight lines with different slopes indicates two distributions mixed in the geochemical dataset, each representing different geochemical environments. Following these studies, Lepeltier (1969) perfected the method and, considering that in geochemical data there is a lack of precision of the lowest values and the importance of the highest ones for the determination of anomalies, for practical reasons, he suggested: "*to cumulate the frequencies from the highest to the lowest values"* (p. 542). Having isolated the straight lines representing each subpopulation, he could graphically estimate some basic parameters.

Subsequently, Parslow (1974) and Coppens (1977) also published articles on this partitioning method, considering more closely multimodal curves, log-normal datasets, and how the range of each subpopulation should be determined. In successive articles, Sinclair (1972, 1974a, 1974b, 1976, 1986, and 1991), proposed a substantial improvement in the technique presented by previous authors for plotting and interpreting probability plots. The main innovation is to have transformed it into a generic technique, making it possible to be applied to datasets that follow uni-, bi, or multimodal distribution curves. This improvement extended the concept of representativeness of geochemical data to much more than the background and threshold of a bimodal curve. It showed that multimodal curves, which are very common in geochemical data of regional surveys, reflect the diversity of environments that are covered by these large-scale works. From the modeling of a complex multimodal curve, when the several straight lines are drawn, each will represent a geochemical-statistical domain. Since a straight line graphically represents each domain, the domain is a normal and unimodal subpopulation. The subpopulations allow for establishing background and anomaly levels for each domain and graphically representing them on geochemical maps. The technique is still used in recent research, as can be seen in Seyedrahimi-Niaraq and Hekmatnejad (2020), Cabassi *et al.* (2021), Moradpouri and Hayati (2021), Giustini *et al.* (2022) and Apollaro *et al.* (2022).

Two computer programs were developed to model probability plots following Sinclair's proposals: ProbPlot by Stanley (1987) and P-Rez by Bentzen and Sinclair (1993). Unfortunately, both run under MS-DOS and do not have adequate graphical outputs. A more recent approach was included in the software SoilExp (BOUDOIRE *et al.,* 2020), however, it is limited to a specific kind of data from soil surveys. Thus, the lack of an application aimed at modeling probability plots, executable in a modern environment with interactivity and good graphic output, led the authors of this article to design and develop the PPlot app. For this, the authors based mainly on the technique proposed by Sinclair (1972, 1974a, 1974b, 1976, 1986 and 1991), with some modifications, and took as reference the above mentioned computer programs. Additionally, Reimann *et al.* (2005) point out that searching for the inflection points is "subjective and experience plays a major role", which may have hindered a wide use of the method. Integrated with the app, we present an automated method to determine the inflection points and obtain the best-fitting modeled curve based in an interactive error minimization algorithm.

3 CONCEPTUAL REVIEW

The classical techniques for statistical data processing have been established for a very particular type of data distribution that is represented by a unimodal symmetric curve, also called the Gaussian curve, or bell-shaped distribution (Figures 1 and 2). Thus, the calculation of statistical estimates, *e.g.,* mean, variance, and standard deviation, is based on this distribution model.

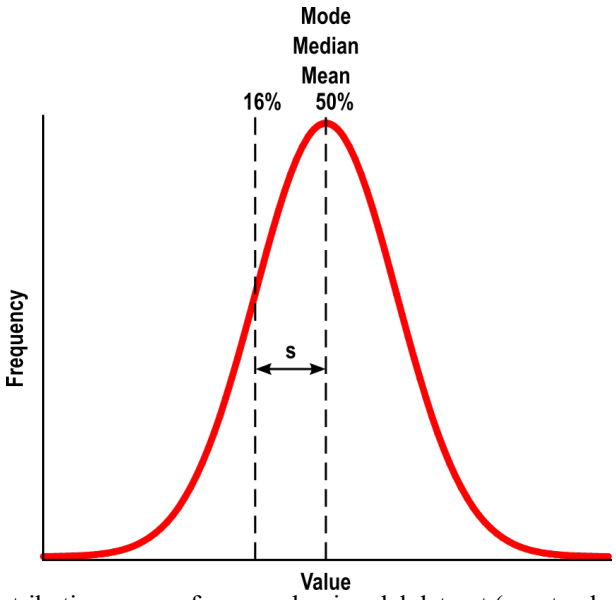


Figure 1. Distribution curve of a normal unimodal dataset (s – standard deviation).

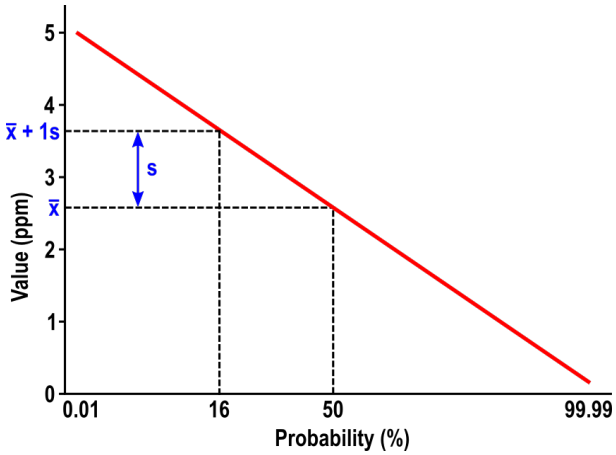


Figure 2. Probability plot of a normal unimodal dataset (s – standard deviation; $\bar{x}$ – mean).

From a misunderstanding of a Hawkes and Webb (1962) statement, "*For a single population of values that are distributed symmetrically (either normally or lognormally), the threshold for that material may be conventionally taken as the mean plus twice the standard deviation*" (p. 30), an equivocated criterion to establishing geochemical anomalies was for long time adopted. It is clear that what the authors stated is valid only for a single population. Still, it was misunderstood and applied to any geochemical dataset regardless of whether it is uni-, bi-, or multimodal. Then, based on this misconception, geochemical anomalies were considered at three hierarchical levels: 3rd order anomaly (mean + 1 standard deviation), 2nd order anomaly (mean + 2 standard deviations) and 1st order anomaly (mean + 3 standard deviations). Since the estimates $\bar{x}$ and s are easily calculated, this mistaken criterion for establishing three levels of anomalies spread out (REIMANN *et al.,* 2005).

An additional complication that has been exhaustively demonstrated in the geochemical literature since the pioneering works of Ahrens

(1953, 1957), Vistelius (1960), and Matheron (1962) is that geochemical data, especially of minor and trace elements, do not follow a normal distribution. Therefore their histograms and distribution curves are asymmetric. The positive skewness makes it clear that geochemical datasets contain a large proportion of low values and a scarcity of high ones. In this case, the transformation of the original values to their logarithms usually promotes an acceptable normalization of the distribution curve (Figures 3 and 4). Thus, having the distribution normalized, and if it is unimodal, it is possible to calculate the statistical estimates of the logarithms of values such as mean, variance, and standard deviation.

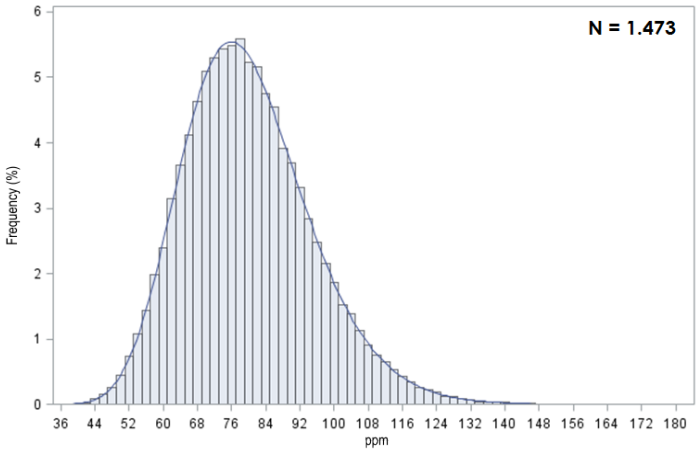


Figure 3. Histogram and distribution curve of a unimodal dataset with positive skew.

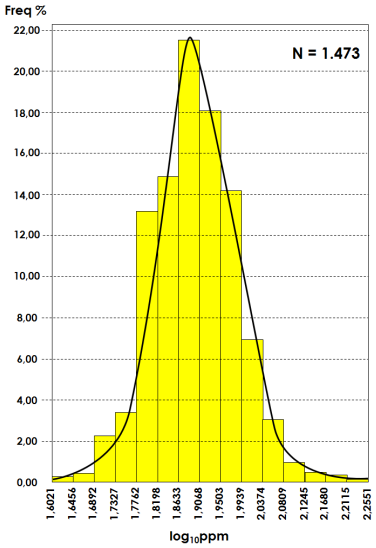


Figure 4. Histogram and distribution curve of a normalized dataset by using the logarithm transformation on the original values.

Moreover, unimodal distributions constitute very particular situations for any database describing any phenomena of nature, whether biotic or abiotic and geochemical in our case. This is because geochemical datasets represent areas that are made up of varied geological environments with different geochemical signals. Thus, it is impossible to expect the geochemical response of one environment to be the same as another.

To better exemplify the need to isolate subpopulations in a dataset, let us consider a hypothetical case. In a geochemical survey carried out over an area with two very different geological environments, one of the elements in the dataset is able to clearly distinguish the two units. The histograms undoubtedly show this: *unit B*, with a low-content geochemical signal, and *unit A*, with a more intense geochemical response (Figure 5). However, a histogram constructed with the complete dataset received from the lab will have the structure presented in Figure 6. It looks like a histogram with a subtle positive skew, highlighting a major mode representing 42.74% of the data, and a much subtler secondary mode with 12.39% of the

data. Due to the mixture in this dataset, it is impossible to discriminate each subpopulation. If any statistical estimate such as mean or standard deviation were calculated with the A+B dataset, it is clear that they would not have any significance. When this hypothetical dataset is plotted on a probability plot, the resulting curve shows the trace of a sigmoid, composed of two smoother sloping segments, slightly deviating from a straight line (Figure 6).

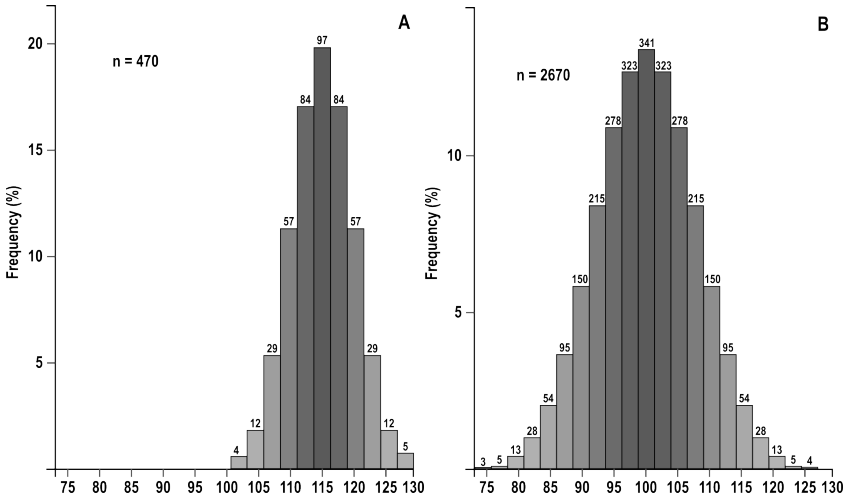


Figure 5. Histogram of hypothetical data from a distinguishing element characterizing the geological units A (A) and B (B).

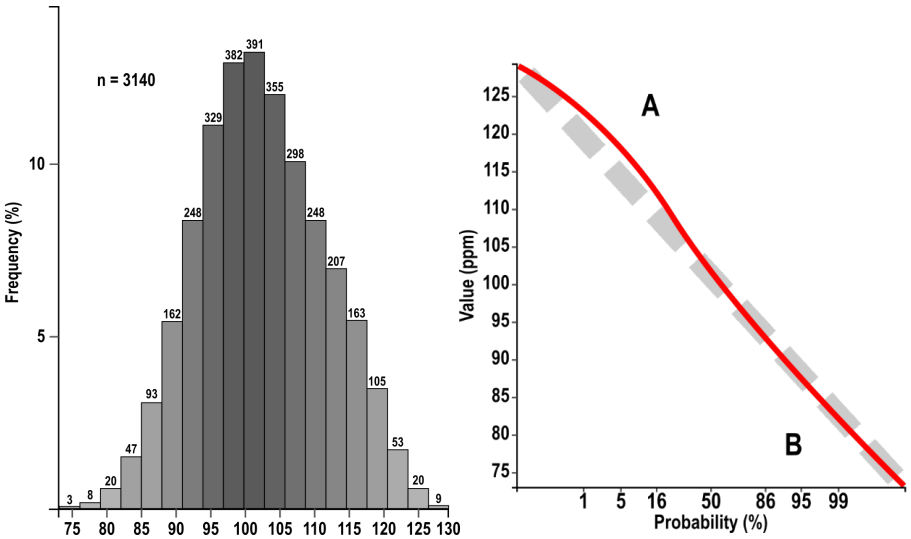


Figure 6. Histogram of hypothetical data from a distinguishing element as received from the lab and its corresponding sigmoidal-shaped probability plot.

A more complex dataset containing the contrasted geochemical signal from three geological units will be expressed in a three-mode histogram, representing the mixture of the subpopulations A, B, and C (Figure 7). The corresponding probability plot will have a complex line with three gentle slope segments, connected by two nearly vertical ones (Figure 7). Each gently sloping segment represents one mode of the histogram, *i.e.*, a subpopulation.

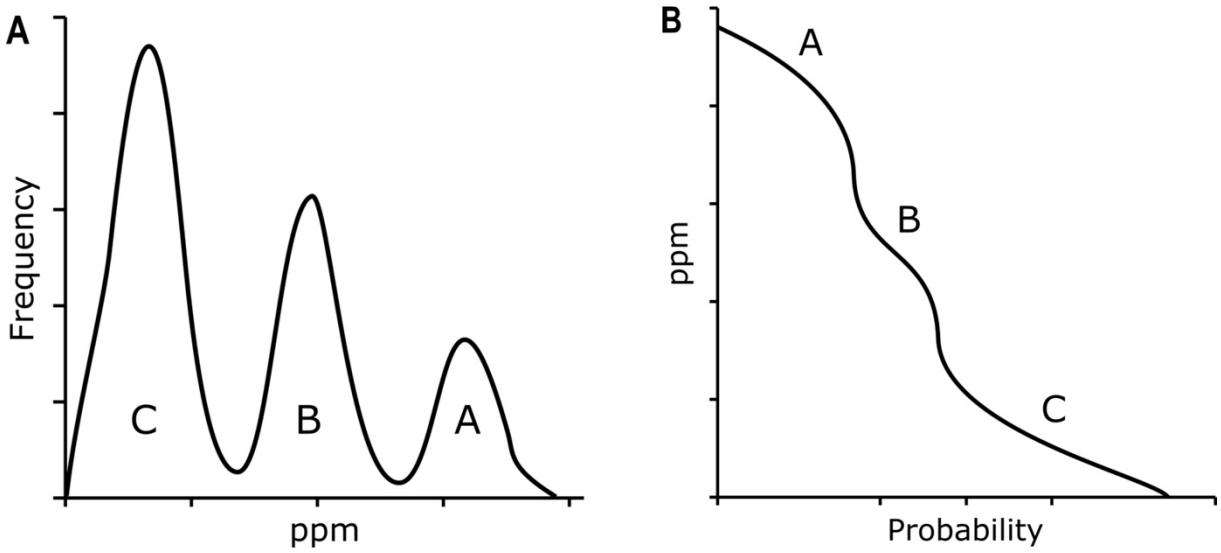


Figure 7. Distribution curve (A) and probability plot (B) of a dataset with three modes (multimodal).

In even more complex situations where the geochemical database represents great geological and geochemical diversity, the probability plot may be multimodal, showing several subpopulations. Thus, it seems clear that isolating mixed subpopulations in a geochemical dataset is a fundamental step in the correct interpretation of the data, as each subpopulation has a specific meaning and can represent different situations or environments, such as lithologies and/or facies, orebodies and host rocks, or even low-grade ore zone and high grade orebody. After partitioning the dataset, it is essential to analyze the spatial distribution of the modeled subpopulations (REIMANN *et al.,* 2005) plotting the classified samples on a map, and associate it to other available data from the studied area (elevation, hydrographic network, lithologies, geological structures, geophysics, etc.).

4 MATERIALS AND METHODS

Probability plots are very efficient tools for delimiting statistically homogeneous domains represented by subpopulations, mixed in a geochemical dataset. They are easy-to-construct diagrams sensitive to deviations from normality and facilitate the identification and characterization of subpopulations, besides providing reliable information for the interpretation of geochemical data. The technique of manually constructing and modeling a probability plot has been exhaustively demonstrated by Sinclair (1972, 1974a, 1974b, 1976, 1986 and 1991).

The web app PPlot was developed using JavaScript and HTML5 to create a userfriendly and robust data analysis tool for partitioning subpopulations using a probability plot that delivers good graphical outputs and can be used in any modern web browser. Besides the core functions of JavaScript and HTML5, several JavaScript libraries were used, as follows:

- jQuery: manipulation of HTML elements and dependency of the jExcel library (source: https://github.com/jquery/jquery);
- jExcel: preparation of editable tables with an interface similar to the software "Microsoft Office Excel"® (source: https://github.com/paulhodel/jexcel);
- Statistics: utilities for statistical data analysis (source: https://github.com/thisancog/statistics.js)
- D3: creation of data-driven elements (such as plots) (source: https://github.com/d3/d3)
- svgsaver: export of the generated drawing for download in SVG (vector) and PNG (raster) file formats (source: https://github.com/Hypercubed/svgsaver).

The following description of the internal mathematical calculations are presented for validation of the method, but the end user is only exposed to the interface presented in section 5. To demonstrate the process, we will consider a sample a with n elements ($a_1, a_2, \ldots, a_i, \ldots, a_n$) composed of m subpopulations ($S_1, S_2, \ldots, S_j, \ldots, S_m$) and $(m - 1)$ inflection points. Contrarily than using a ready-to-use printed probability paper to plot the data as in the classical method, we need to calculate and

devise our own virtual probability "paper" (Figure 8). To do this, the input data is sorted in descending order (LEPELTIER, 1969) and its approximate percentile (Eq. 1) and the inverse of the cumulative normal distribution function (Eq. 2) are calculated.

$$p_i = \frac{i-\frac{3}{8}}{n-\frac{1}{4}} \qquad \text{Eq. 1}$$

$$z = \phi^{-1}(p) = probit(p) \qquad \text{Eq. 2}$$

where: p_i = approximated probability;
i = position of the data value in the ordered list;
n = number of observations;
z = observed values;
φ^{-1} and *probit* = inverse of the cumulative normal distribution function.

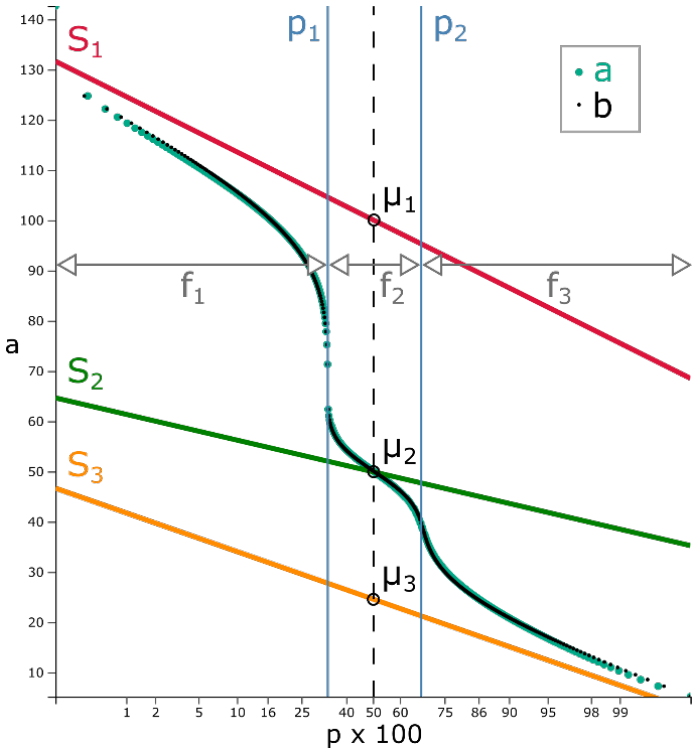


Figure 8. Probability plot with the elements and notation used in this section (*a* – sample; *b* – modeled sample; *p* – inflexion point; *S* – subpopulation; *f* – proportion of the subpopulation *S*; μ – the mean of subpopulation *S*).

The probit represents "*how many standard deviations from the mean a given cumulative probability is spread*" (statitics.js Documentation) since the data is standardized and can be plotted in a linear scale in the abscissa against the data values in the ordinate to generate the probability plot. Since the probit values have no direct meaning to the geoscience researcher, the values are replaced with their respective quantiles (Figure 9).

To begin the partitioning process, we must first determine how many inflection points are present in the probability plot and in which position they are. When selecting the inflection points graphically, their value is read on the abscissa and converted to percentile using the cumulative normal distribution function (Eq. 3) for $N(0,1)$. At the same time, the proportion of each subpopulation in the mixture is calculated by Eq. 4.

$$p_j = \phi(z_j) \qquad \text{Eq. 3}$$

$$f_j = \begin{cases} p_1\,,\ j = 1 \\ p_j - p_{j-1},\ 1 < j < m \\ 1 - \ p_{m-1},\ j = m \end{cases} \qquad \text{Eq. 4}$$

where: z_j = the observed value read on the abscissa for the j inflection point;
p_j = the probability for the j inflection point;
f_j = the proportion of the S_j subpopulation in the sample.

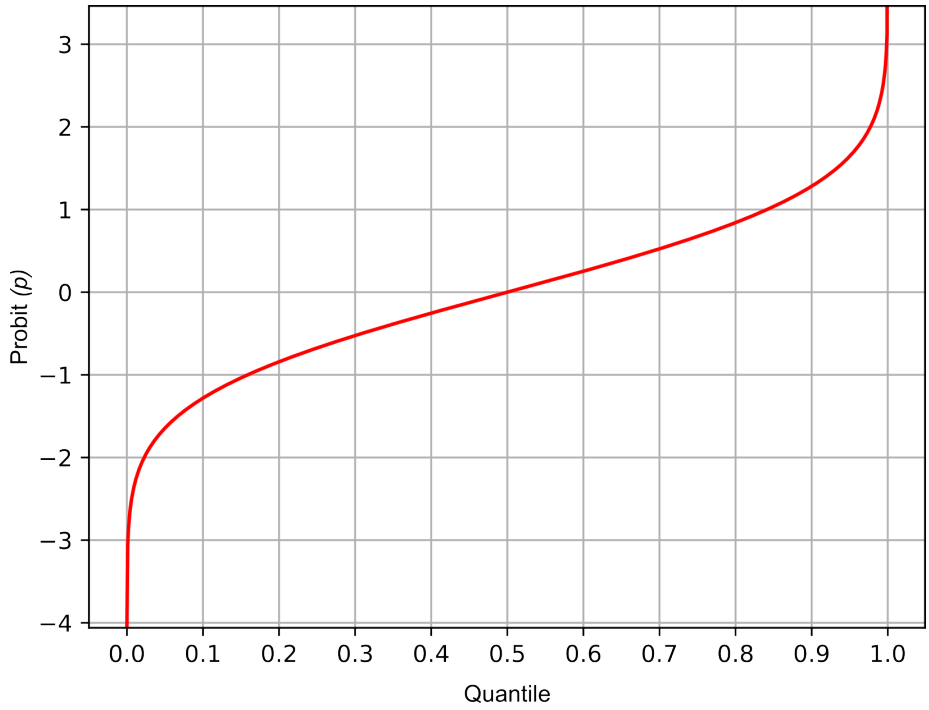


Figure 9. Relationship between probit values and quantiles. Probit(50) = 0, so the y-axis in the probability plot (Figure 8) is displaced from the origin of the x-axis (source: https://en.wikipedia.org/wiki/Probit#/media/File:Probit_plot.png)

Then, the mean (Eq. 5) and standard deviation (Eq. 6) of each subpopulation is calculated using as approximation the data values that are respective of the inflection such as

$$\mu_j = \bar{x}\left\{a_{p_{j-1}}, \ldots, a_{p_j}\right\} \qquad \text{Eq. 5}$$

$$\sigma_j = s\left\{a_{p_{j-1}}, \ldots, a_{p_j}\right\} \qquad \text{Eq. 6}$$

where: μ_j = the mean of subpopulation S_j;
σ_j = the standard deviation of subpopulation S_j.

In the probability plot, a normal distribution can be plotted as a linear function where the mean is the constant term and the standard deviation is the slope (Eq. 7 and Figure 8).

$$S_j = f(z) = \mu_j - \sigma_j\, z \qquad \text{Eq. 7}$$

After subpopulations are defined, a sample b resulting from this mixture must be calculated considering the f_j proportions, made up of n elements (b_1, b_2, ... b_i, ..., b_n). Therefore, for each element a_i of the original sample we calculate the corresponding b_i element of sample b, as the weighted mean of the expected normal distribution value for each subpopulation. This is achieved using Eq. 8.

$$probit(p_{b_i}) = probit\left(1 - \sum_{j=1}^{m}\left(\phi\left(\frac{a_i - \mu_j}{\sigma_j}\right) \times f_j\right)\right) \quad \text{Eq. 8}$$

The fit of the points generated by the calculated sample *b* in regard to the original sample *a* can be evaluated graphically or by the error coefficient *E* defined by Eq. 9. The lower the value of *E* the more similar sample *b* is to the original sample *a*.

$$E = \frac{\sum_{i=1}^{n}\left|probit(p_{b_i}) - probit(p_{a_i})\right|}{n} \quad \text{Eq. 9}$$

The limits of each subpopulation are defined by the mean plus or minus two standard deviations, which in a normal distribution account for 95.45% of the data, or the limits of the neighboring subpopulation, choosing the most restrictive of the two, as showed in Eq. 10 and Eq. 11.

$$S_j(min) = max\{(\mu_j - 2\sigma_j), S_{j-1}(max)\} \quad \text{Eq. 10}$$
$$S_j(max) = min\{(\mu_j + 2\sigma_j), S_{j+1}(min)\} \quad \text{Eq. 11}$$

5 USAGE, OPERATION, AND FUNCTIONALITY

5.1 DATA INPUT

Data should be inputted as a table where the results (*e.g.* geochemical samples or geophysical signal) are ordered in rows. More than one variable can be inputted simultaneously, and each variable should be placed in a column, as is usual in these datasets. A label, such as the chemical element name, can be placed on the first line. Only numbers should be present in the table, but if values under the detection limit are present *(e.g.* < 0.1, commonly reported by analytical labs), these can be automatically transformed to half of their value. Additionally, a missing data code can be inputted to automatically exclude such results from the analysis. Since geochemical databases usually contain sensitive and confidential information, it is important to point out that all processing is done locally on the user browser, and no inputted data is transferred over the internet or stored in the server.

In sequence, the app displays the selected variable's probability plot, histogram, and box plot. The histogram is initially constructed with the number of bins defined by *k* (Eq. 16), but this can be interactively altered to any value arbitrarily selected by the user. The probability plot is initially presented with all the original data points, but the user may choose to use the log-transformed values, checking the corresponding check box. Furthermore, the user may also choose to use the data intervals from the histogram instead of the original data points, which may be useful when the dataset is very large and/or noisy.

$$k = \frac{a(max) - a(min)}{\frac{1}{4} \times s} \quad \text{Eq. 16}$$

5.2 GRAPHICAL ANALYSIS

The analysis may be carried out manually by the user or automatically by the internal routines of the app. To begin the manual analysis, the user must first choose how many subpopulations will be partitioned in the dataset. This is done visually by the user, who identifies the number of inflection points in the probability curve. Clicking over the plot creates a vertical line which marks the inflection point, and the first estimate of the corresponding subpopulations are calculated and plotted as colored lines while the points of the modelled sample are plotted as black points. All generated lines can be graphically moved, and the model is dynamically recalculated on change. The inflection line can be dragged sideways, while the subpopulation lines can be dragged up- and downwards (to alter the mean value, μ) or, while holding the Ctrl/Command key, tilted (to alter the standard deviation, σ) (Figure 10). The values of *p* (probability) for each inflection and the values of μ and σ for each subpopulation are displayed in an editable table. Any value changed in the table will also be automatically updated in the plots.

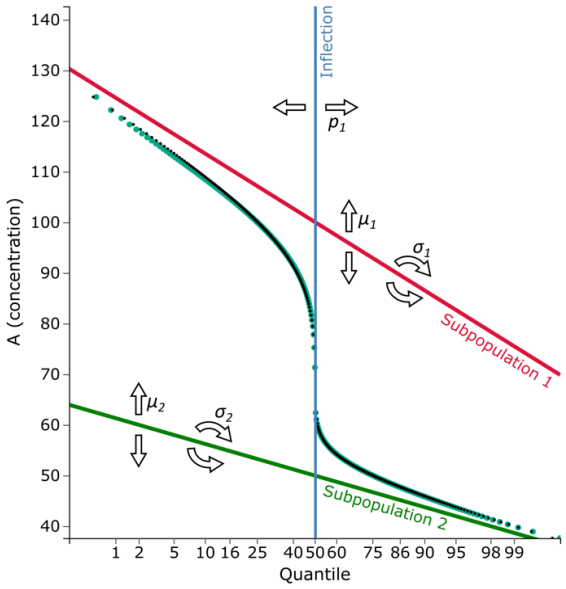
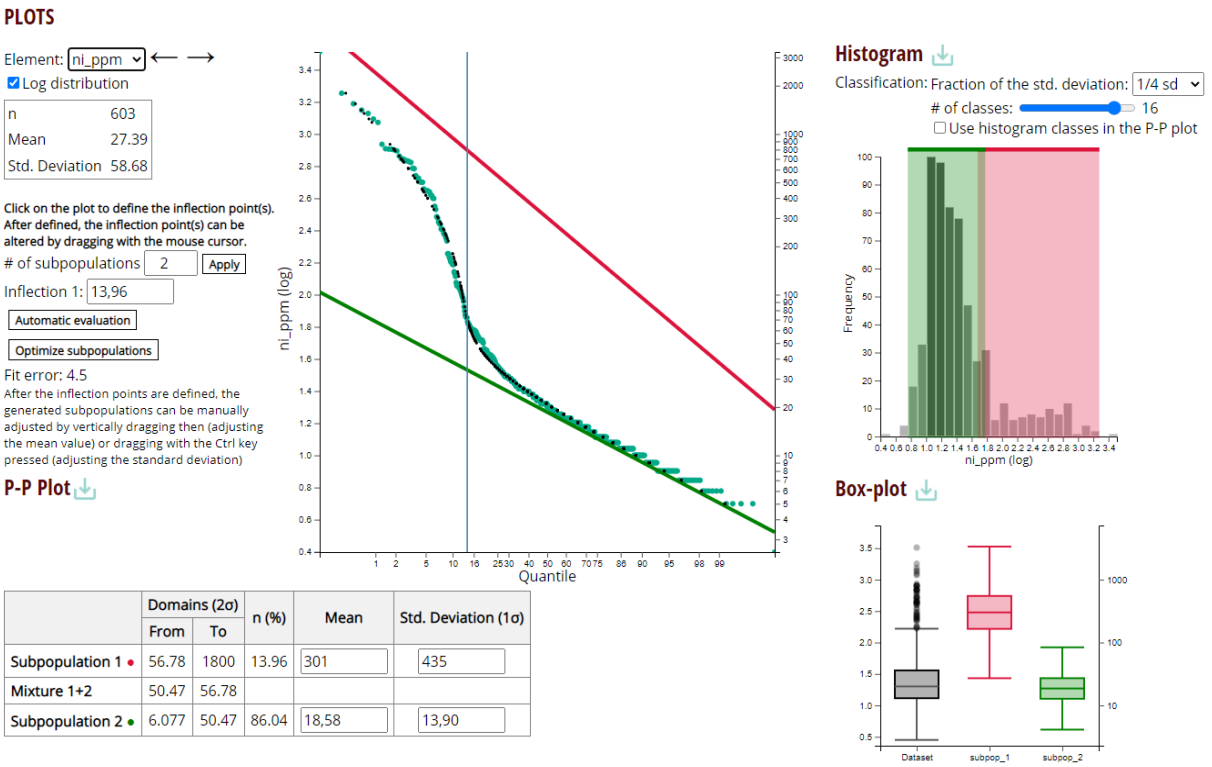


Figure 10. Probability plot elements with arrows indicating all possible user interactions with the plot (p – inflection point; μ – mean; σ – standard deviation).

The objective is to vary these parameters until the modeled population B is as closest as possible to the original population A. The fit of the curve can be assessed visually or by the calculated error value. The fine tuning of the inflection point placement and of the mean and standard deviation values (optimization) can be done mathematically by the app using a Penalty Function Method that uses a sensitivity matrix (FRISWELL; MOTIERSHEAD, 1995), which interactively alters all parameters simultaneously in order to identify the best possible fit of the modelled population in regard to the original dataset, that is, the one with the lowest error.

Besides the manual analysis, the user can initiate an analysis by clicking the "automatic analysis" button. It is based on the principle that the lowest error value will be obtained when the correct number and position of inflection points are placed in the plot, as well as the mean and standard deviation for each subpopulation are fine tuned. The routine begins with placing a single inflection in $p = 1$, calculating the corresponding error, and repeating the operation up to $p = 99$ in 1 percentile intervals. The percentile with the minimum obtained error is selected and the first inflection point is fixed. Subsequently, the routine continues by placing a second inflection in $p = 1$ and following the same procedure as before, selecting the percentile corresponding to the lowest error value. After each selection of a new inflection point, the model is optimized. The routine stops when adding a new inflection reduces the error value by less than 10%. The determined inflection points are evaluated, all those closer to *probit(p)* = ± 0.15 of each other are merged, and a last optimization is run. The result obtained by the automatic analysis should represent the best fit of the curve using the least possible number of inflection points. Still, the user must evaluate the geological and statistical meaning of the generated subpopulations. The user can interact with the result in the same way as described in the manual analysis.

After the user is satisfied with the obtained model, a table displays the range for each subpopulation according to the criteria presented in section 4. If there is an overlap, the table displays this range using the term "mixture". Additionally, the box plot of the subpopulations is plotted, and the ranges are overlaid in the histogram (Figure 11). The resulting plots can be saved as an image vector file format (SVG). Furthermore, the user can save the PPlot project as a local file (JSON format) and later reopen it exactly as it was saved with all the data, modeling, and plots. Supplementary data 1 contains a JSON file with the data and modeled probability plots presented in the next section, that can be opened in PPlot.



	Domains (2σ)		n (%)	Mean	Std. Deviation (1σ)
	From	To			
Subpopulation 1	56.78	1800	13.96	301	435
Mixture 1+2	50.47	56.78			
Subpopulation 2	6.077	50.47	86.04	18,58	13,90

Figure 11. Resulting table and respective histogram and box-plots.

6 TESTING THE APPLICATION

6.1 TESTING THE CODE

First of all, to test that the written code computes exactly the mathematical development presented in section 4, we will use a computer-generated sample (n = 300). The dataset inputted into PPlot is composed by the mixture of three normal distributions $N(\bar{x},s)$: N_1(150,5) (n = 50); N_2(100,15) (n = 100), and N_3(50,10) (n = 50) (see supplementary data 2). The numerical values for the inflection points, means, and standard deviations were numerically informed, generating the probability plot (Figure 12) and corresponding table of ranges (Table 1). As can be assessed graphically and by the obtained error values (0.77), the fit of the modeled sample with the original sample is almost perfect, as expected.

6.2 TESTING THE PARTITIONING IN A GENERATED DATASET

A computer-generated sample *a* (n = 250) composed by the mixture of two normal populations (*v, n = 100* and *w, n = 150*) with some associated randomness, so that $v(\bar{x},s)$ = (100.81, 5.38) and $w(\bar{x},s)$ = (50.10, 9.77), was inputted into PPlot (see supplementary data 2). The "automatic evaluation" option was used, which identified one inflection point (0.400) to obtain the smallest error possible (E = 2.12). As can be seen by the fit of the curve in Figure 13 and the obtained values for each subpopulation in Table 2, the partitioning was very efficient in estimating the parameters for the two subpopulations. Classifying the dataset by the obtained ranges, the method correctly identified 98% of *v* and 94% of *w*.

6.3 TESTING THE PARTITIONING IN A REAL GEOCHEMICAL DATASET

A real multi-elemental dataset of a soil grid (n = 134) sampled over an alkaline intrusion (BRUMATTI *et al.*, 2015) was inputted into PPlot (see supplementary data 2). We analyzed the probability plot for Ce and Ni, choosing a logarithmic distribution, using the automatic inflection point, and graphically adjusting the parameters to obtain the minimum error value (Figure 14). In both cases, two subpopulations were identified and adequately separated (Table

3). The obtained ranges were plotted on a map to analyze their spatial distribution and compare it to the geological map (Figure 15). In both cases, the subpopulations formed well-delineated groupings. For Ce, the higher values of $S_1(Ce)$ delimit the carbonatite unit. In contrast, for Ni, the higher $S_1(Ni)$ values spatially correlate well with the olivine clinopyroxenite and nepheline syenite units, so the statistically defined subpopulations are geologically coherent.

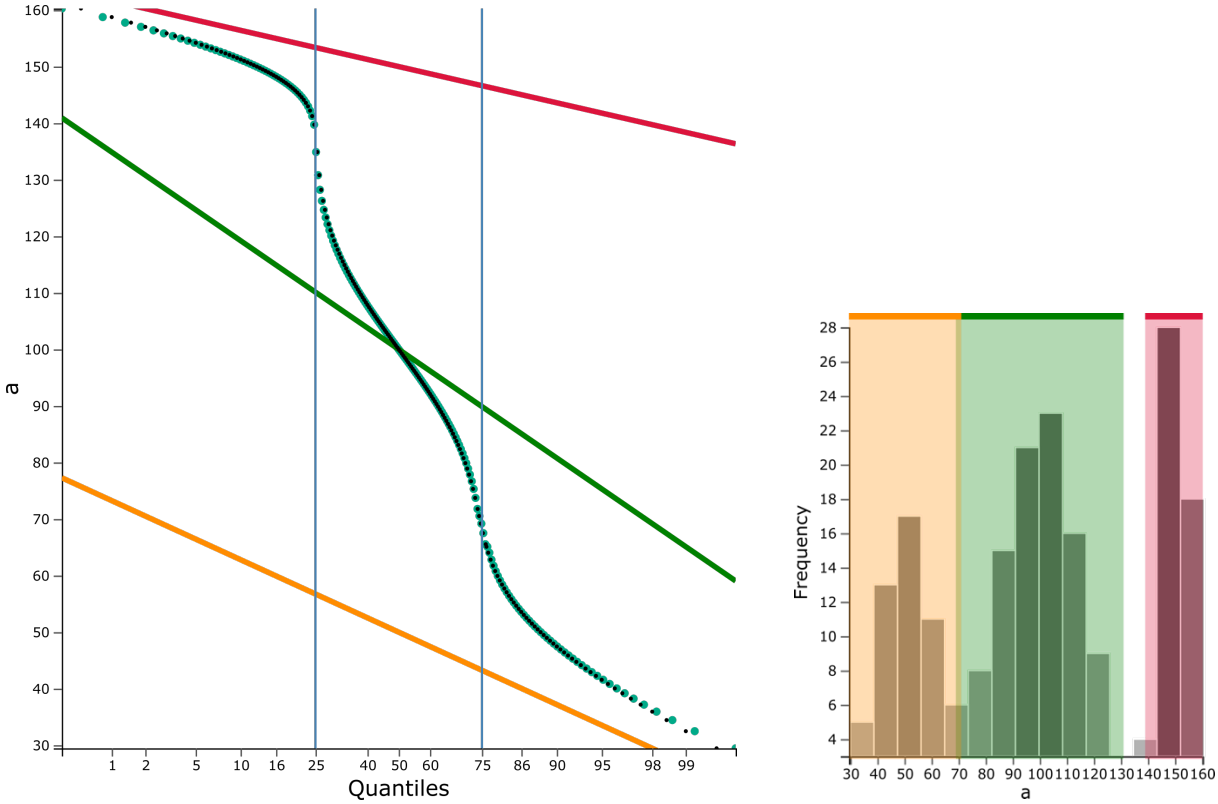


Figure 12. Probability plot and histogram for computer-generated sample (n = 300) composed by the mixture of three normal distributions N(150,5) (n = 50), N(100,15) (n = 100), and N(50,10) (n = 50), in green, and respective distribution generated by the numerical input of the inflection points, means and standard deviations, in black.

Table 1. Inputted parameters for generating the modeled distribution shown in Figure 12, and obtained values for "Error" and "Range".

	Parameters	Range
S_1	$\mu_1 = 150$; $\sigma_1 = 5$; $f_1 = 25\%$;	140 - 160
S_2	$\mu_2 = 100$; $\sigma_2 = 15$; $f_2 = 50\%$;	70 - 130
S_3	$\mu_3 = 50$; $\sigma_3 = 10$; $f_3 = 25\%$;	30 - 70
p_1	0.25	-
p_2	0.75	-
Error	0.77	-

Table 2. Obtained values for the modeled distribution shown in Figure 13.

	Parameters	Range
S_1	$\mu_1 = 101$; $\sigma_1 = 5.375$; $f_1 = 40.00\%$;	90.06 - 112
S_2	$\mu_2 = 50.10$; $\sigma_2 = 9.772$; $f_2 = 60.00\%$;	30.56 - 69.65
p_1	0.400	-
Error	2.12	-

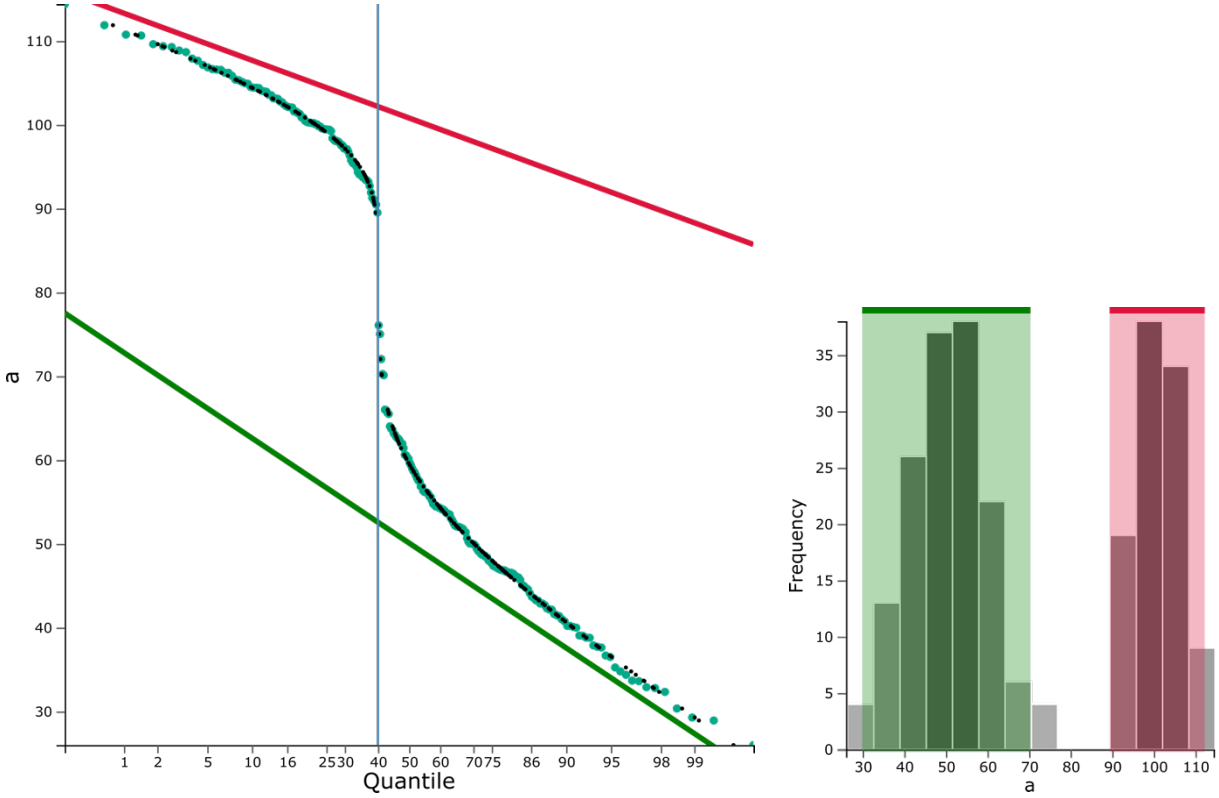


Figure 13. Probability plot and histogram for computer-generated sample (n = 250) composed by the mixture of two normal populations with some associated randomness, in green, and respective modeled distribution generated by the automatic definition of the inflection point, mean, and standard deviation values, in black.

Table 3. Obtained values for the modeled distribution shown in Figure 14.

		Parameters	Range
Ce	S_1	$\mu_1 = 2292$; $\sigma_1 = 3125$; $f_1 = 13.75\%$;	504 - 12802
		$S_1 + S_2$ *(overlap)*	411 - 504
	S_2	$\mu_2 = 149$; $\sigma_2 = 125$; $f_2 = 86.25\%$;	44 - 411
	p_1	0.1375	-
	Error	4.95	-
Ni	S_1	$\mu_1 = 449$; $\sigma_1 = 480$; $f_1 = 44.21\%$;	105 - 1921
	S_2	$\mu_2 = 39$; $\sigma_2 = 17$; $f_2 = 55.79\%$;	19 - 81
	p_1	0.4421	-
	Error	5.6	-

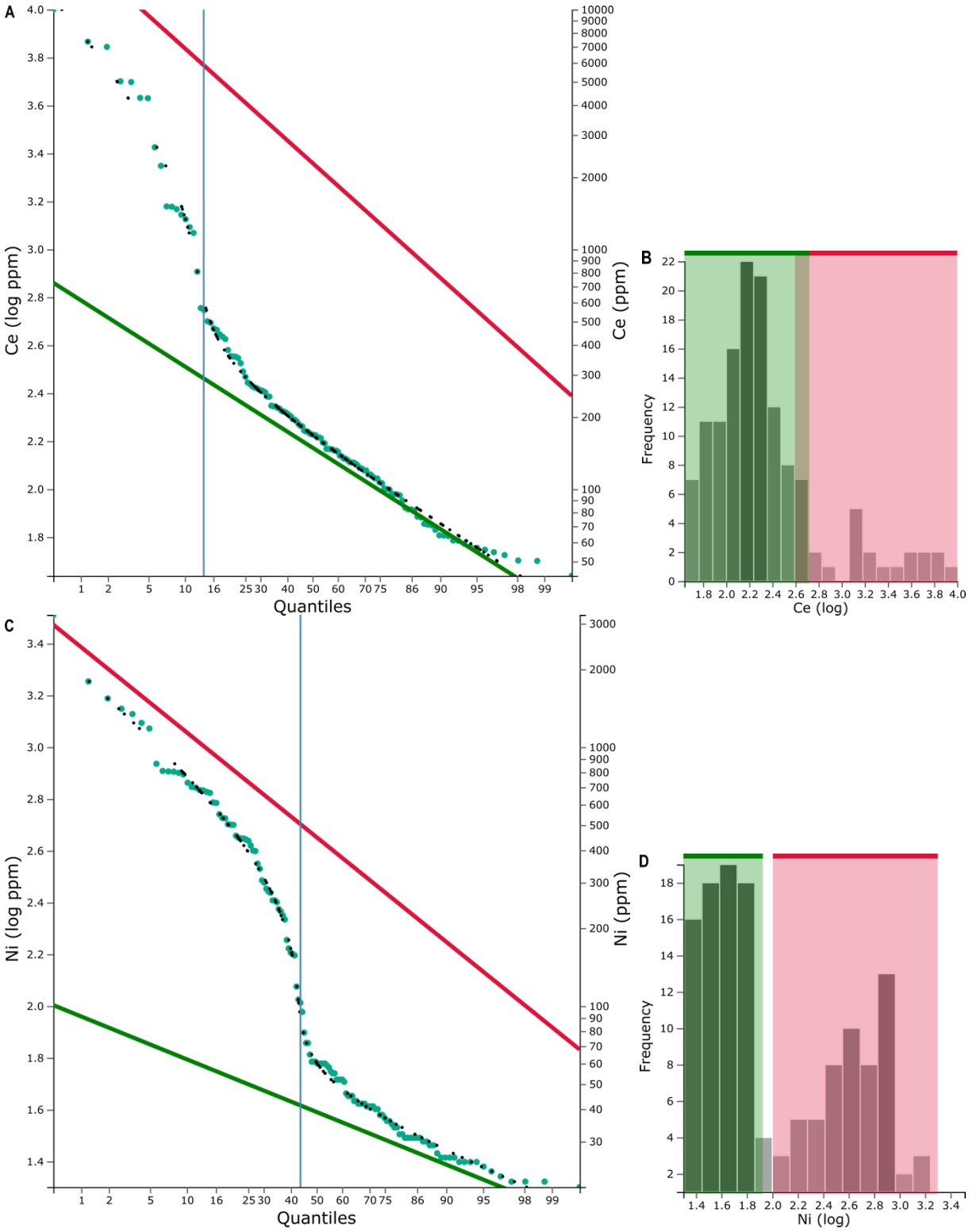


Figure 14. Probability plot and histogram for Ce (A and B) and Ni (C and D) of a multi-elemental dataset of a soil grid (n = 134) sampled over an alkaline intrusion (BRUMATTI *et al.*, 2015). The original data is in green, and the respective modeled distribution is in black.

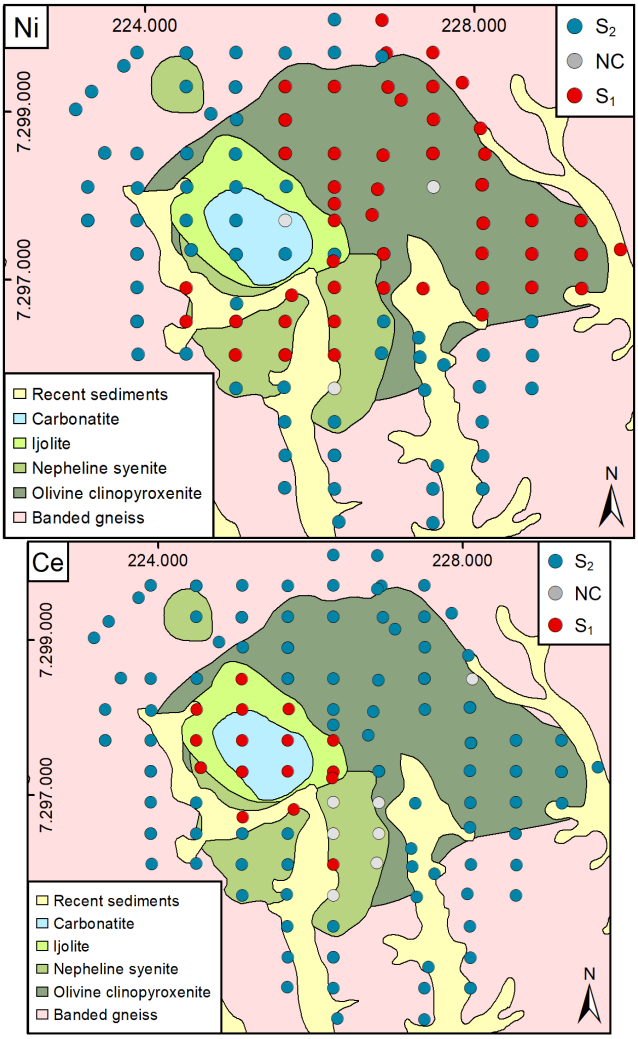


Figure 15. Sampling sites classified by the modeled subpopulations and underlying geology (UTM zone 22S) (BRUMATTI *et al.*, 2015). S_1 – subpopulation 1; S_2 – subpopulation 2; NC – not classified.

7 FINAL REMARKS

Geochemical and geophysical datasets are often composed of a mixture of several statistical subpopulations. The presented application is an easy-to-use yet powerful tool for partitioning normal subpopulations from a dataset. Besides that, it also generates a dashboard of the main aspects an exploration geologist usually needs to assess when doing univariate analysis: the histogram, the box-plot, and the probability plot. This partitioning technique, while effective, had its usage hindered nowadays mainly due to software limitations, which were not updated to current user systems. Partitioning of the subpopulations may reveal geochemical anomalies, map geochemical domains or classify the data into subsets which will then undergo classical statistical analysis.

Like all data analysis techniques, this method has many advantages, but has its limitations. For example, the cumulative probability plot of the geochemical data set is usually very smooth and therefore, the positions of the inflection points are not always evident. Sometimes identification is subjective, and experience plays an important role. To surpass this limitation, a feature designed for automatic identification of the inflection points, based on multiple iterations, has been included on the app. Another limitation of the technique is the inability to separate values positioned in the overlapping zones of bi- or multimodal subpopulations. However, while generating geochemical maps, the overlapping zone samples can be identified as belonging to the mixed zone between subpopulations. This is a very straightforward technique that can be easily applied and understood by exploration geologists, creating practical results and adding value to the analysis and interpretation of

results. It is important to emphasize that a good input database is essential to obtain good results when partitioning subpopulations, as well as performing any statistical tests in PPlot or any other software. Regarding geochemical samples, representativeness of the sampling environment is critical to generate good quality data that will subsequently generate a reliable definition of subpopulations that can be interpreted as geochemical signatures.

PPlot runs under any modern web browser, including mobile devices. It can be used by any geology student, researcher, or practitioner without any other specific or additional software installation. It has a graphical interface that, besides its primary intended use, also allows a didactic approach to the method, so that professors can use it to demonstrate the relationships between the three plots and how subpopulations are represented on it. Furthermore, although it has been designed with a mainly geochemical perspective, the technique can be used on any dataset composed of several normal subpopulations.

8 ACKNOWLEDGEMENTS

Acknowledgements are due to Prof. Dr. Alastair James Sinclair who, in lectures given in Itaipava (1982) and Rio de Janeiro (1984) (Brazil), presented his innovative and elegant technique for modeling probability graphs to Brazilian geochemists and served as a reference to the authors of this article. The authors also thank Lucas Rocha for reviewing section 4 of this article, and the two anonymous reviewers who gave invaluable comments and suggestions to the original manuscript.

9 REFERENCES

AHRENS L.H. A Fundamental Law of Geochemistry. **Nature,** 172(4390):1148-1148, 1953 doi: 10.1038/1721148a0

AHRENS L.H. Lognormal-type distributions—III. **Geochim. Cosmochim. Acta**, 11(4):205-212, 1957. doi: 10.1016/0016-7037(57)90094-7

APOLLARO, C.; DI CURZIO, D.; FUOCO, I.; BUCCIANTI, A.; DINELLI, E.; VESPASIANO, G.; CASTRIGNANÒ, A.; RUSI, S.; BARCA, D.; FIGOLI, A.; GABRIELE, B.; DE ROSA, R. A multivariate non-parametric approach for estimating probability of exceeding the local natural background level of arsenic in the aquifers of Calabria region (Southern Italy). **Sci. of the Tot. Env.** 806, 150345, 2022. doi: 10.1016/j.scitotenv.2021.150345

BENTZEN A.; SINCLAIR A.J. **P-RES – a computer program to aid in the investigation of polymetallic ore reserves.** Tech. Rept. MT-9, Mineral Deposit Research Unit, Dept. of Geological Sciences, University of British Columbia, Vancouver, 1993, 55 pp.

BOUDOIRE, G.; LIUZZO, M.; CAPPUZZO, S.; GIUFFRIDA, G.; COSENZA, P.; DERRIEN, A.; FALCONE, E. E. The SoilExp software: An open-source Graphical User Interface (GUI) for post-processing spatial and temporal soil surveys. **Comput. Geosci.** 142, 104553, 2020 doi: 10.1016/j.cageo.2020.104553

BRUMATTI, M.; ALMEIDA, V.V.; LOPES, A.P.;CAMPOS, F.F.;PERROTTA, M.M.;MENDES, D.; PINTO, L.G.R.; PALMEIRA, L.C.M. **Metalogenia das províncias minerais do Brasil: rochas alcalinas da porção meridional do Cinturão Ribeira, estados de São Paulo e Paraná.** CPRM. Informe de Recursos Minerais. (Série Províncias minerais do Brasil, 6), 2015. 87 p.

CABASSI, J.; VENTURI, S.; DI BENNARDO, F.; NISI, B.; TASSI, F.; MAGI, F.; RICCI, A.; PICCHI, G.; VASELLI, O. Flux measurements of gaseous elemental mercury (GEM) from the geothermal area of "Le Biancane" natural park (Monterotondo Marittimo, Grosseto, Italy): Biogeochemical processes controlling GEM emission. **J. Geochem. Explor.** 228, 106824, 2021 doi: 10.1016/j.gexplo.2021.106824

COPPENS R. **Statistiques élémentaires appliqueés aux sciences de la terre**. Orléans: Institut National Polytechnique de Lorraine, 1977, 219 p.

DOUST J.F.; JOSEPHS H.J. A simple introduction to the use of statistics in telecommunications engineering. **Post Office Eng. Journ**. 34: 36–41,79–84, 139–44, 173–8, 1941

FRISWELL M.I.; MOTIERSHEAD J. E. **Finite Element Model Updating in Structural Dynamics**. Springer Science+Business Media Dordrecht, 1995. 304p. doi: 10.1007/978-94-015-8508-8

GIUSTINI, F.; RUGGIERO, L.; SCIARRA, A.; BEAUBIEN, S. E.; GRAZIANI, S.; GALLI, G.; PIZZINO, L.; TARTARELLO, M. C.; LUCCHETTI, C.; SIRIANNI, P.;

TUCCIMEI, P.; VOLTAGGIO, M.; BIGI, S.; CIOTOLI, G. Radon Hazard in Central Italy: Comparison among Areas with Different Geogenic Radon Potential. **Int. J. Environ. Res. Public Health**, 19, 666, 2022 doi: 10.3390/ijerph19020666

HARDING J.P. The use of probability paper for the graphical analysis of polymodal frequency distributions. **J. Mar. Biol. Assoc. U. K.,** 28(1):141-153, 1949 doi:10.1017/S0025315400055259

HAWKES, H.E.; WEBB, J.S. **Geochemistry in mineral exploration**. New York: Harper & Row. 1962. 415 p.

HAZEN, A. Storage to be provided in impounding reservoirs for municipal water supply. **Proc. Amer. Soc. Civil Eng**., 39:1943–2044, 1913.

LEPELTIER, C. A Simplified statistical treatment of geochemical data by graphical representation. **Econ. Geol**., 64: 538-550, 1969. doi: 10.2113/gsecongeo.64.5.538

MATHERON G. **Traité de géostatistique appliquée**. Paris, Bureau de Recherches Géologiques et Miniéres Mémoire, no. 14, tome 1. 1962

MORADPOURI, F.; HAYATI, M. A copper porphyry promising zones mapping based on the exploratory data, multivariate geochemical analysis and GIS integration. **Appl. Geochem**., 132, 105051, 2021 doi: 10.1016/j.apgeochem.2021.105051

PARSLOW, G.R. Determination of background and threshold in exploration geochemistry. **J. Geochem. Explor.** 3:319-336, 1974. doi: 10.1016/0375-6742(74)90002-8

PEARSON K. III. Contributions to the mathematical theory of evolution. **Philos. Trans. Royal Soc., A**. 185:71–110, 1894. doi: http://doi.org/10.1098/rsta.1894.0003

REIMANN C.; FILZMOSER P.; GARRETT R.G. Background and threshold: critical comparison of methods of determination. **Sci. Total Environ**., 346, 1-16, 2005 doi: 10.1016/j.scitotenv.2004.11.023

RISSIK, H. Probability graph paper and its engineering applications. **Journal of the American Society for Naval Engineers**, 54(1):103-119, 1942. doi.org/10.1111/j.1559-3584.1942.tb05208.x

SEYEDRAHIMI-NIARAQ M.; HEKMATNE-JAD A. The efficiency and accuracy of probability diagram, spatial statistic and fractal methods in the identification of shear zone gold mineralization: a case study of the Saqqez gold ore district, NW Iran. **Acta Geochimica**, 40:78-88, 2020 doi: 10.1007/s11631-020-00413-7

SINCLAIR, A.J. **Some statistical applications to problems in mineral exploration**. Report - Department of Geological Sciences, University of British Columbia 13:18. 1972 https://eurekamag.com/research/020/061/020061599.php

SINCLAIR A.J. Selection of threshold values in geochemical data using probability graphs, **J. Geochem. Explor.** 3:129-149, 1974a. doi: 10.1016/0375-6742(74)90030-2

SINCLAIR A.J. Some considerations regarding grid orientation and sample spacing. In: ELLIOTT, I.L.; FLETCHER, W.K. (Eds) **Geochemical Exploration**, Elsevier Sc. Publ. Co., Spec. Publ. 2: 133-140, 1974b.

SINCLAIR A.J. **Applications of probability graphs in mineral exploration**. The Assoc. of Explor. Geochem, 1976. 95 p. (Spec. Vol. n. 4).

SINCLAIR A.J. Statistical interpretation of soil geochemical data. In: FLETCHER, W.K.; HOFFMAN, S.J.; MEHRTENS, M.B.; SINCLAIR, A.J.; THOMSON, I (Eds). **Exploration Geochemistry: Design and Interpretation of Soil Surveys**. Chelsea: Soc. of Econ. Geol., (Reviews in Economic Geology, v. 3).1986. 180 p., p. 97-115. doi:10.5382/Rev.03.05

SINCLAIR A.J. A fundamental approach to threshold estimation in exploration geochemistry: probability plots revisited. **J. Geochem. Explor**. 41(1-2):1- 22, 1991. doi: 10.1016/0375-6742(91)90071-2

STANLEY C.R. ***PROBPLOT*** **An Interactive Computer Program To Fit Mixtures of Normal (or Log-Normal) Distributions With Maximum Likelihood Optimization Procedures.** Association of Exploration Geochemists Special Volume 14. 1987. Version 1.00 C7

TENNANT C.B.; WHITE H.L. Study of the distribution of some geochemical data. **Economic Geology**, 54(7): 1281-1290, 1959 doi: 10.2113/gsecongeo.54.7.1281

VISTELIUS A. The Skew Frequency Distributions and the Fundamental Law of the Geochemical Processes. **The Journal of Geology**, 68(1):1-22, 1960