EXPLORE
Newsletter for the
Association of Applied Geochemists
www.appliedgeochemists.org
NUMBER 199 JUNE 2023

# The Log Transformation Explained

***Robert G. Garrett***
*Emeritus Scientist, Geological Survey of Canada,601 Booth Street, Ottawa, Ontario, Canada K1A 0E8*
*(robert.garrett@nrcan-rncan.gc.ca)*
https://doi.org/10.70499/WBIO3405

## INTRODUCTION

Numerous papers have been written on the logarithmic distribution of trace elements and ore metals over the years. One of the first was Razumovsky (1940), followed some years later by the influential work of Ahrens, typified by his 1954 paper (Ahrens 1954). Vistelius (1960) argued that lognormal distributions came about naturally on physical grounds due to the processes of rock formation. Similar physical processes, involving repetitively splitting a volume of material into portions containing increased or decreased elemental concentrations, have been studied by DeWijs (1951), Brinck (1976) and Garrett (1986). This process leads to logbinomial distributions. Limpert *et al.* (2001) demonstrated how a similar process leads to the lognormal distributions so common in the physical sciences.

The reality of applied geochemistry is that data are derived from surficial and bedrock environments that are, more often than not, complex and field data sets are rarely symmetrically distributed. The data, as observed, are drawn from the various populations and geochemical processes present in the survey or study area. They do not exhibit 'bell-shaped' distributions and are frequently polymodal. They can be 'tortured' towards normality with tools like the Box-Cox power transform (e.g. Howarth and Earle 1979), of which a logarithmic transform is a special case. Furthermore, such transforms may obscure polymodality that conveys useful information. Alternately, a power transformation that expands the data to a range with maximum spread and/or contrast to provide an improved visualization, may be employed for map presentations and distributional displays (Stanley 2005).

## BACKGROUNDS AND THRESHOLDS

In applied and exploration geochemical surveys, the range of background values must be established for each of the diverse surficial and bedrock environments present. The threshold can be defined as the upper limit of background variation (Reimann and Garrett 2005). These geochemical parameters are best estimated through appropriate orientation surveys and inspection of the data using maps, histograms (Hawkes and Webb 1962), and probability plots (Lepeltier 1969; Sinclair 1976).

If the former did not lead to the choice of a geochemically justified threshold, Hawkes and Webb (1962) proposed that threshold values could be estimated as the mean of the background data plus two standard deviations (SD). An estimate that would, assuming normality, identify the value below which 98% of background data should fall. In any subsequent survey using similar procedures in a geologically and geochemically similar area, applying that threshold would identify 2% of the data for further investigation. These, hopefully, would include any samples related to non-background processes and mineral occurrences of interest. Whether or not this approach is appropriate, and how it should be accomplished, has been the topic of numerous papers, for example, Matschullat *et al.* (2000), Reimann and Filzmoser (2000) and Reimann *et al.* (2005). Methods not requiring normality, non-parametric methods, may be employed. However, even then normality lurks in the background: the median replaces the mean; however, underlying the calculation of the Median Absolute Deviation (MAD), the equivalent of the standard deviation, lies a factor based on the normal distribution. Recently, procedures to unmix complex geochemical data sets have been investigated (e.g. Eschenfelder *et al.* 2023), however, some are based on the assumption of normality (e.g. Lucero-Álvarez *et al.* 2021).

This is further complicated by the fact that geochemical data are compositional, i.e., they sum to a constant, and therefore, as some values increase, others must decrease. The impact of this and the necessity for compositional data analysis procedures have been discussed by Barceló *et al.* (1996), Mateus-Figueras *et al.* (2005) and Buccianti *et al.* (2006), among others.

## BACK TO FIRST PRINCIPLES

Statistical estimates of the background range and threshold are based on an assumption of underlying normality. "Normality assumes that the continuous variables to be used are normally distributed. Normal distributions are symmetric around the center (a.k.a. the mean) and follow a 'bell-shaped' distribution" (Statistics Solutions 2013). This begs the question, what is a continuous variable? "A continuous variable is one which can take on an uncountable set of values. For example, a variable over a non-empty range of the real numbers is continuous, if it can take on any value in that range" (Wikipedia 2019a). So, what is a real number? "A real number is a value of a continuous quantity that can represent a distance along a line. The adjective real in this context was introduced in the 17th century by René Descartes, who distinguished between real and imaginary roots of polynomials. The real numbers include all the rational numbers, such as the integer -5, the fraction 4/3, and all the irrational numbers, such as $\sqrt{2}$" (Wikipedia 2019b).

*continued on page 5*

## The Log Transormation Explained *continued from page 1*

### ANALYTICAL CHEMICAL DATA

Analytical data meet the criteria for being continuous and real. However, they are measured on what McCue (2007) defines as 'ratio scales' that are "numeric and are associated with a true zero – meaning that nothing can be measured. For example, weight is a ratio scale". Furthermore, Mosteller and Tukey (1977) define 'counted fractions' as scales that are bounded by zero and one. Thus 'weight per weight' analytical geochemical data expressed, e.g. in mg/kg, are measured on 'ratio scales' and are 'counted fractions'; they are constrained to vary between zero and 100%, $10^6$ mg/kg, etc., and are bounded.

Thus, data at the extremes, close to zero or the maximum of the ratio scale, can be positively (right) or negatively (left) skewed, respectively, as their possible values cannot fall below zero or exceed the scale maximum. In the central part of the range, the spread of the data may be unconstrained by the bounds and behave like a normal distribution, i.e., following a 'bell-shaped' distribution. Therefore, if parametric statistical procedures are to be applied to the data approaching the scale minima and maxima, they need to be transformed towards normality. Referring specifically to 'Proportions and Percentages', Deacon (2020) offers three procedures:

1. Convert to arcsine values (*see* Holland 2017);
2. A logarithmic transformation; and
3. Converting to probits.

Wilson *et al.* (2010) and Warton and Hui (2011) report that the arcsine transformation is losing popularity, despite its use in the geosciences (Miller and Kahn 1962; Krumbein and Graybill 1965; Holland 2017). If the data are drawn from an underlying Poisson distribution, which is uncommon in geochemistry, the arcsine transform will induce homoscedasticity, i.e. equal spread across the range of the data, a desirable statistical property (Stanley pers. comm. 2023). The logarithmic transform only works for the lower part of the ratio scale as demonstrated below. Converting to probits, though it does cover the full range with reference to the normal distribution, it is more suitable for instances where the values are zero or one and therefore not continuous; it will not be discussed further.

### LOGITS AND THE LOG TRANSFORM

The reality is that analytical data are measured on ratio scales and are counted fractions. What is required is a transform that breaks the bounds of counted fractions and permits values to occupy the complete range of real numbers, i.e. $-\infty$ to $+\infty$. Such a transform is the logit (Berkson 1944; Holland 2017; Wikipedia 2020), the log of the odds for some probability $p$.

$$\text{logit}\,(p) = \log\left(\frac{p}{1 - p}\right)$$

As a mechanism by which to transform a zero-to-one counted fraction, or any concentration that can be rendered zero-to-one through division by the scale maximum, to a real number, the logit transformation suffices. It matters not whether a Naperian logarithm to the base e, or a logarithm to the base 10 is employed; here the former is applied.

The relationship of the logit to the zero-to-one proportion scale is shown in Figure 1 (left). When the proportion is plotted with logarithmic scaling (Fig. 1, right) the relationship between logit and log(base 10) proportion appears to be linear between low proportions and 0.1 (i.e. 10%). The estimated linear (Pearson) correlation coefficient is >0.9999 between proportions equivalent to 1 µg/kg (ppb) and 10%. Clearly, there is an operational equivalency between the logit of a proportion, counted fraction, or concentration and its logarithm up to concentrations of 10%.

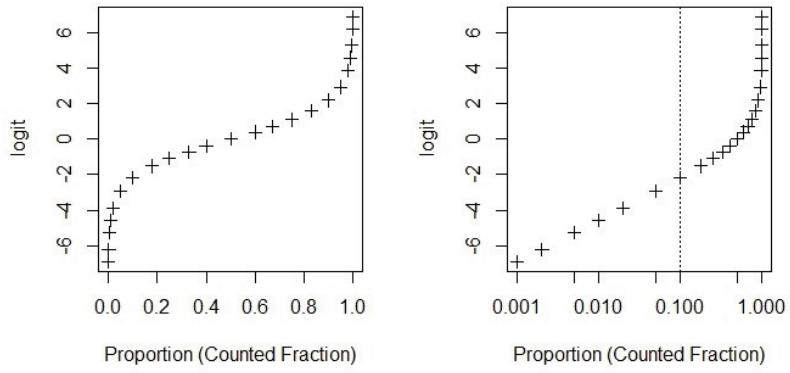


*Fig. 1. The logit function versus proportion (left), with logarithmic scaling (right).*

### LOGIT APPLIED TO GEOCHEMICAL DATA

From Figure 1 (left) the logit transformation should be effective for reducing skewness across all concentrations, i.e., for both distributions exhibiting negative skews as they approach the scale maxima, and positive skews as they approach scale minima. In both instances, the logit transformation releases the data from their bounding restrictions. The silica data in Figure 2 range from 59 to 95% and a negative skew is apparent. The application of a logit transformation (Fig. 2, right) opens the data as it approaches 100% and leads to a more symmetric (normal) distribution.

*continued on page 7*

## The Log Transormation Explained

*continued from page 5*

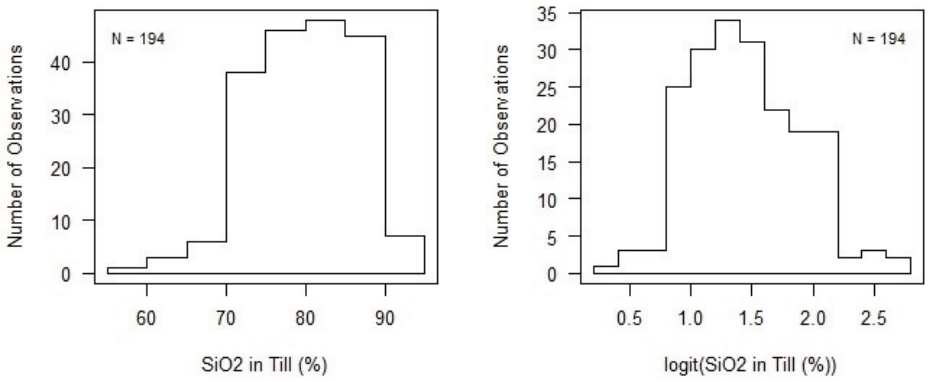


*Fig. 2. Silica concentration ($SiO_2$, Li-metaborate fusion) in <63 µm till (left), and with a logit transformation (right).*

The applicability of the logit transform across a wide range of concentrations is demonstrated with a set of soil organic carbon data ranging from 0.5 to 77% (Fig. 3, left) exhibiting extreme positive skew.

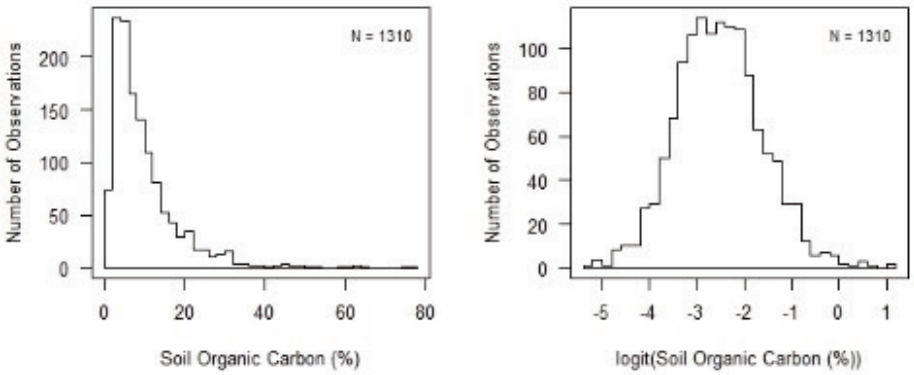


*Fig. 3. Organic carbon concentration (Loss-on-Ignition) in <2 mm soil (left), and with a logit transformation right).*

The logit transform effectively removes the positive skew and leads to a symmetric, more normal, distribution suitable for the application of parametric (normality-based) statistical methods.

An example of an extreme positive skew across almost three orders of magnitude, with data ranging from 0.2 to 96 mg/kg, familiar in trace element studies, is shown in Figure 4 (left). Again, the logit transform is effective in leading to a more symmetric distribution (Fig. 4, right), although still with outliers due to contamination from an anthropogenic sources in the study area.

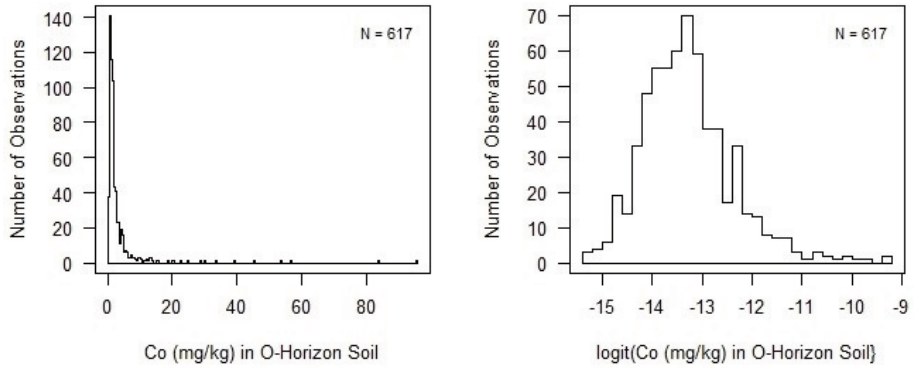


*Fig. 4. Cobalt concentration ($HNO_3$ digestion) in <2 mm O-horizon soil (left), and with a logit transformation (right).*

As demonstrated in Figure 1 (right) at levels below 10% (100,000 mg/kg) the logarithmic and logit transforms are equivalent. Figure 5 provides a visual comparison with the Co data exhibited in Figure 4, where plotting Co concentrations with logarithmic scaling is equivalent to logit transforming the data.

In multivariate data analysis, full compositional data analysis procedures, i.e. log-ratios, are required. However, bivariate displays fall between univariate and multivariate and may benefit from logarithmic scaling. If the data span more than one-and-a-half to two orders of magnitude, they probably display a lack of homogeneity of variance. This feature, also known as heteroscedasticity, is visually expressed by the data points spreading out in an increasingly broader ‘fan’

*continued on page 9*

## The Log Transformation Explained

*continued from page 7*

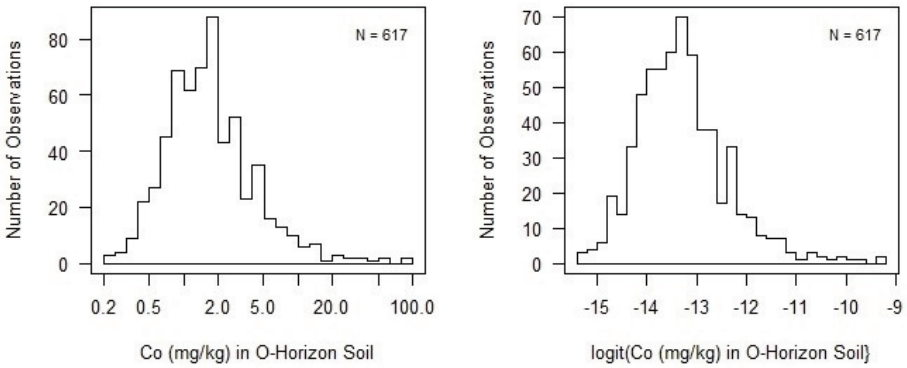


*Fig. 5. Histograms for cobalt concentration ($HNO_3$ digestion) in <2 mm O-horizon soil, with logarithmic (left) and logit (right) transformations.*

with increasing concentration when plotted on the original scale (Fig. 6, left). Plotting geochemical data with logarithmic scaling provides a quick graphical check for heteroscedasticity. If it is present, the data plot as a band of equal spread with increasing concentration, *see* Figure 6 (right). A statistical assumption of regression-line (Ordinary Least Squares) fitting is that across the range of the data the variances, or spreads, of the data are independent of concentrations, i.e. they do not 'fan out'.

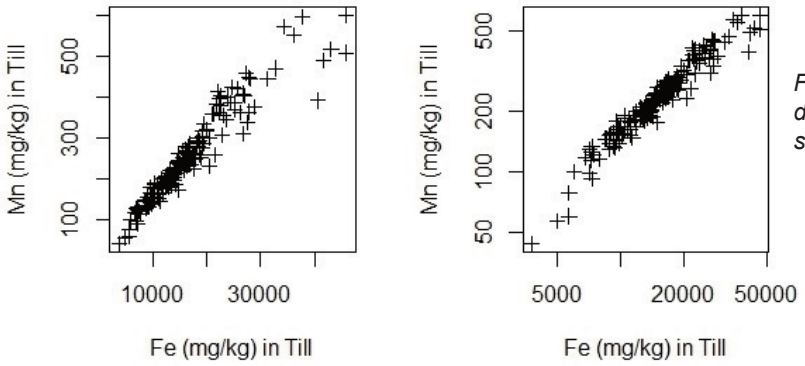


*Fig. 6. Plots of manganese vs. iron concentrations (four-acid digestion) in <63 µm till, without (left) and with logarithmic scaling (right).*

The procedure applied by statisticians in undertaking analyses based on squared differences, e.g. regression modeling and Analysis of Variance, is to logarithmically transform the data (Bartlett 1947; Weissberg 1980, and others).

However, Figure 6 does not tell the whole story, as the Fe and Mn, as well as being counted fractions individually, are members of an even larger 'counted fraction', the overall chemical composition of the sample.

The solution to this problem is the use of log-ratios (*see*, for example, Aitchison 1984 and Pawlowsky-Glahn *et al.* 2015). The simplest approach is to use an arithmetic log-ratio, dividing the elements by another member of the composition and taking the logarithm, or simply plotting the ratios with log-scaling, as is familiar in petrochemical studies (*see* Pearce 1968). Again, taking the logarithm of a ratio turns it into a real number. Figure 7 (left) displays the same Fe and Mn data as ratios to Al, a major component in the overall composition, plotted with logarithmic scaling. The relatively 'tight' band in Figure 6 (right) has been broadened as a result of the recognition that the data are compositional; as some components (parts) increase others must decrease. The data do not 'fan out' and homogeneity of variance is maintained, though the spread has increased due to taking account of at least one of the other elements in the composition.

*continued on page 10*

***Note:*** *This EXPLORE article has been extracted from the original EXPLORE Newsletter. Therefore, page numbers may not be continuous and any advertisement has been masked.*

## The Log Transformation Explained

*continued from page 9*

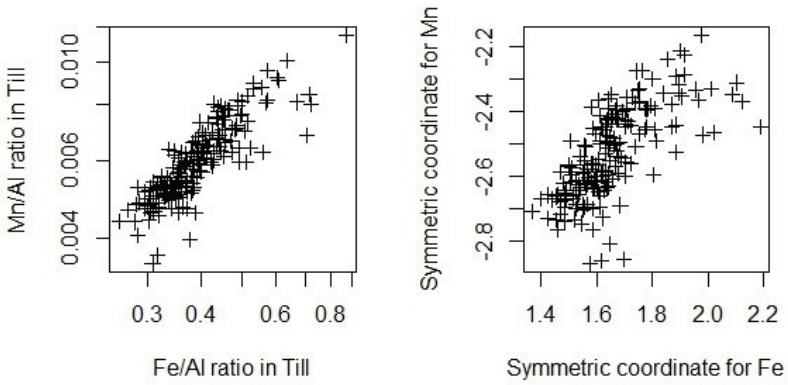


*Fig. 7. Iron and manganese concentrations for <63 µm till, plotted as a log-ratios to aluminum (left), and as symmetric coordinates in the context of all available major and minor components (right). All data after a four-acid digestion.*

The use of symmetric coordinates (Garrett *et al.* 2017; Kynčlová *et al.* 2017), a complex log-ratio, allows all the major and minor element concentrations to be included. The concentration data for Figures 6 and 7 were determined following a four-acid ($HF$-$HClO_4$-$HNO_3$-$HCl$) near-total digestion, Si was not determined. The major and minor elements included in the calculations of symmetric coordinates were Al, Ca, Mg, Na, K, Fe, Mn, Ti, Cr and P. The result of including the nine major and minor elements beyond Al is displayed in Figure 7 (right). Homogeneity of variance has been maintained, but the spread (uncertainty in the inter-element relationship) has been further increased as a result of taking the almost-complete suite of major and minor elements into account.

**DISCUSSION**

It has been shown that logarithmic distributions can come about due to rock-forming processes. Furthermore, the very nature of analytical data as counted fractions, i.e. relative data, requires a transformation to better visualize their distributions and is necessary if statistical procedures that assume normality are to be employed. For example, if there is no prior knowledge of the threshold for an exploration program, or it cannot be derived by graphical inspection or analysis, an option is to select some percentile of the data (e.g. 98$^{th}$ percentile) or employ statistical estimation. The nature of trace element geochemical data requires a transformation. Without transformation, estimates for the upper limits of the background values (i.e. thresholds), by median+2*MAD, or mean+2*SD, may exceed the upper bound of the data; alternatively, a lower bound of less than zero may be estimated. These thresholds are impossible and their presence is a reminder of the need for a transformation. Trace element data visualization benefits from logarithmic scaling: simple calculations should be undertaken following a logarithmic transform, and the results back-transformed to the original scaling. An advantage of visualization with logarithmic scaling is that differences are appreciated as ratios, conforming to the way applied geochemists consider their data, i.e. levels are twice, or half, etc., some other value, not as absolute arithmetic differences.

At higher concentrations, especially when approaching scale maxima, visualizations may benefit from a logit transformation (e.g. Figs. 2 and 3). In the mid ranges, it may not be necessary to undertake any transformation. Webster and Oliver (1990) state that for the arcsine transformation, “When the observed values fall in the range 30–70 percent, there is very little to be gained by the transformation, and it is unlikely that there will be much gain when only a small proportion of the observations fall outside this range”. This statement applies equally to the logit transformation. Prudent investigators will study their data visually and determine if a lack of symmetry or homogeneity of variance requires a data transformation before proceeding further.

For multivariate data analysis, log-ratio transformations (e.g.

*continued on page 11*

## The Log Transformation Explained *continued from page 10*

centred and isometric log-ratios) are required to reveal true inter-element relationships independent of closure. For bivariate relationships the arithmetic log-ratio transformation, as discussed above, is sufficient. A common thread in all these procedures is that a logarithmic transformation is used to turn a zero lower bounded ratio into a real number.

With the availability of machine learning methods and other advanced or non-parametric methods, data transformations to reduce skewness in data prior to analysis may not be necessary. However, for many visualization tasks, a transformation will assist geochemists in their interpretational tasks.

### CONCLUSIONS

Geochemical data are not real numbers in mathematical terms, they are counted fractions lying between bounds, and they can neither fall below, nor rise above, those bounds. To convert the counted fractions to real numbers suitable for statistical analysis, a logit transformation is sufficient. At concentrations below 10%, logit and logarithmic transformations are equivalent. Thus, when trace element data are plotted with logarithmic scaling the values become real, the positive skew is reduced, and the data appear to be more symmetrical and normally distributed. Furthermore, many of the geological processes controlling the distribution of elements in nature are multiplicative, leading to logbinomial or lognormal distributions.

The logarithmic transformation is relevant and useful for two reasons. Firstly, it effectively converts trace element geochemical counted fractions to real numbers and improves data visualization by ‘decompression’ at low concentrations. Secondly, if the assumptions that underlie parametric statistical methods, the estimation of means, variances (standard deviations), and other procedures that are based on squared differences, are to be met, the data should approach normality and variances need to be independent of concentration, i.e. homoscedastic. A logarithmic transformation of trace element data meets both these requirements.

### ACKNOWLEDGEMENTS

The author thanks his colleagues, Clemens Reimann and Peter Filzmoser, for their comments and suggestions on an earlier draft of this article, Patrice de Caritat and Cliff Stanley for their constructive reviews, and Beth McClenaghan for editorial assistance.

### NOTE

All calculations and plot preparation was undertaken with R 3.4.3 (R-Project 2020) and package ‘rgr’ version 1.1.16 (Garrett 2013).

### REFERENCES

Ahrens, L.H. 1954. The lognormal distribution of trace elements (A fundamental law of geochemistry and its subsidiary). Geochimica et Cosmochimica Acta, 5(2), 37–93.

Aitchison, J. 1984. The statistical analysis of geochemical compositions. Mathematical Geology, 16(6), 531–564.

Barceló, C., Pawlowsky, V. and Grunsky, E. 1996. Some aspects of compositional data and the identification of outliers. Mathematical Geology, 28(4), 501–518.

Bartlett, M.S. 1947. The use of transformations. Biometrics, 3(1), 39–52.

Berkson, J. 1944. Application of the logistic function to bioassay. Journal of the American Statistical Association, 39(227), 357–365.

Brinck, J.W. 1976. Critical parameters for the production, depletion and substitution of mineral resources. Geologie en Mijnbouw, 5(3-4), 185–194.

Buccianti, A., Mateau-Figueras, G. and Pawlovsky-Glahn, V. 2006. Frequency distributions and natural laws in geochemistry. Geological Society of London, Special Publication, 264, 175–189.

Deacon, J. 2020. The Really Easy Statistics Site. (accessed 2023/5/15) http://archive.bio.ed.ac.uk/jdeacon/statistics/tress4.html#Transformationofdata

DeWijs, H.J. 1951. Statistics of ore distribution, Part 1. Journal of the Royal Netherlands Geological and Mineralogical Society, 13(11), 365–375.

*continued on page 12*

## The Log Transformation Explained *continued from page 11*

Eschenfelder, J., Lipp, A.G. and Roberts, G.G. 2023. Quantifying excess heavy metal concentrations in drainage basins using conservative mixing models. Journal of Geochemical Exploration, 248, 107178.

Garrett, R.G. 1986. Geochemical abundance models – An update, 1975–1984. In: Cargill, S.M. and Green, S.B. (eds.) Prospects for Mineral Resource Assessments on Public Lands, Proceedings of the Leesburg Workshop. U.S. Geological Survey Circular 980, 207–220.

Garrett, R.G. 2013. The 'rgr' package for the R Open Source statistical computing and graphics environment – a tool to support geochemical data interpretation. Geochemistry: Exploration, Environment, Analysis, 13(4), 355–378.

Garrett, R.G., Reimann, C., Hron, K., Kynčlová, P. and Filzmoser, P. 2017. Finally, a correlation coefficient that tells the geochemical truth. EXPLORE, 176, 1, 5-6, 8–10.

Hawkes, H.E. and Webb, J.S. 1962. Geochemistry in Mineral Exploration. Harper and Row.

Holland, S. 2017. Data Analysis in the Geosciences. (accessed 2023/5/15) http://strata.uga.edu/8370/rtips/proportions.html

Howarth, R.J. and Earle, S.A.M. 1979. Application of a generalized power transformation to geochemical data. Mathematical Geology, 11(1), 45–62.

Krumbein, W.C. and Graygill, F.A. 1965. An Introduction to Statistical Models in Geology. McGraw-Hill Inc.

Kynčlová, P., Hron, K. and Filzmoser, P. 2017. Correlation between compositional parts based on symmetric balances. Mathematical Geosciences, 49(6), 777–796.

Lepeltier, C. 1969. A simplified treatment of geochemical data by graphical representation. Economic Geology, 6(5), 538–550.

Limpert, E., Stahel, W.A. and Abbt, M. 2001. Log-normal distributions across the sciences: keys and clues. Biosciences, 51(5), 341–352.

Lucero-Álvarez, J., Acosta-Rodríguez, B.F., Araiza-González, A.E., Espejel-Garcia, V.V., Villalobos-Aragón, A. and Franco-Gallegos, L.I. 2021. Interpretation of geochemical anomalies and domains using Gaussian mixture models. Applied Geochemistry, 135, 105110

Mateau-Figueras, G., Pawlovsky-Glahn, V. and Barceló-Vial, C. 2005. The additive logistic skew-normal distribution on the simplex. Stochastic Environmental Research and Risk Assessment, 19(3), 205–214.

Matschullat, J., Ottenstein, R. and Reimann. C. 2000. Geochemical background – can we calculate it? Environmental Geology, 39(9), 990–1000.

McCue, C. 2007. Data Mining in Predictive Analysis: Intelligence Gathering and Crime Analysis. Elsevier.

Miller, R.L. and Kahn, J.S. 1962. Statistical Analysis in the Geological Sciences. John Wiley and Sons, Inc.

Mosteller, F. and Tukey, J.W. 1977. Data analysis and regression: a second course in statistics. Addison-Wesley Publishing Co.

Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. 2015. Modeling and analysis of compositional data. Wiley, Chichester, U.K.

Pearce, T.H. 1968. A contribution to the theory of variation diagrams. Contributions to Mineralogy and Petrology, 19(2), 142–157.

R-Project 2020. The R Project for Statistical Computing. (accessed 2023/5/15). http://www.r-project.org

Razumovsky, K. 1940. Distribution of metal values in ore deposits. Comptes Rendus (Doklady) de l'Académie des Sciences de l'URSS, 28(9), 814–816.

*continued on page 13*

## The Log Transformation Explained *continued from page 12*

Reimann, C. and Filzmoser, P. 2000. Normal and lognormal distribution in geochemistry: death of a myth: Consequences for the statistical treatment of geochemical and environmental data. Environmental Geology, 39(9), 1001–1014.

Reimann, C., Filzmoser, P. and Garrett, R.G. 2005. Background and threshold: critical comparison of methods of determination. Science of the Total Environment, 346(1/3), 1–16.

Reimann, C. and Garrett, R.G. 2005. Geochemical background – Concept and reality. Science of the Total Environment, 350(1/3), 12–27.

Sinclair, A.J. 1976. Applications of Probability Graphs in Mineral Exploration. Association of Exploration Geochemistry, Special Volume 4.

Stanley, C.R. 2005. Numerical transformation of geochemical data: 1. Maximizing geochemical contrast to facilitate information extraction and improve data presentation. Geochemistry: Exploration, Environment, Analysis, 5(1), 1–11.

Statistics Solutions 2013. Common Assumptions in Statistics. (accessed 2023/5/15) https://www.statisticssolutions.com/common-assumptions-in-statistics/3

Vistelius, A.B. 1960. The skew distribution and the fundamental law of the geochemical processes. Journal of Geology, 68(1), 1–22.

Warton, D.J. and Hui, F.K. 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology, 92(1), 3–10.

Webster, R. and Oliver, M.A. 1990. Statistical Methods in Soil and Resource Survey. Oxford University Press.

Weissberg, S. 1980. Applied Linear Regression. John Wiley and Sons, Inc.

Wikipedia 2019a. Continuous or discrete variable. (accessed 2023/5/15) https://en.wikipedia.org/wiki/Continuous_or_discrete_variable

Wikipedia 2019b. Real number. (accessed 2023/5/15) https://en.wikipedia.org/wiki/Real_number

Wikipedia 2020. Logit. (accessed 2023/5/15) https://en.wikipedia.org/wiki/Logit

Wilson, E., Underwood, M., Pukrin, O., Letto, K., Doyle, R., Caravan, H., Camus, and Bassett, K. 2010. The Arcsine Transformation: Has the time come for retirement? (accessed 2022/12/15) https://www.mun.ca/faculty/dschneider/b7932/B7932Final10Dec2010.pdf