

A Visual Basic Spreadsheet Macro for Geochemical Background Analysis

by Zoran Nakić¹, Kristijan Posavec², and Andrea Bačani²

Abstract

A Visual Basic macro entitled BACKGROUND calculates geochemical background values of chemical parameters and estimates threshold values separating background data from anomalies. The macro uses two statistical methods, the *iterative 2- σ technique* and the *calculated distribution function*, and integrates these model-based objective methods into a widely accessible platform (i.e., MS Excel). The macro offers the possibility for automated processing of geochemical data and enables an automated generation of background range and threshold values for chemical parameters.

Introduction

Human impacts on ground water quality are growing, causing a significant threat for ground water use in the future. Under such conditions, it is particularly important to realistically evaluate and separate anomalies or outliers from the background concentrations of a chemical element or compound.

Geochemical background was defined by Hawkes and Webb (1962) as the normal abundance of an element in barren earth material. Natural background levels reflect natural processes unaffected by human activities, but it can be argued that such a background no longer exists due to human influence on the whole planet. An option is to define an ambient background under slightly altered conditions, when elevated levels of element concentrations in soil or water result from long-term human impact such as agriculture, industry, and urbanization and are no longer natural (Reimann and Garrett 2005). In 1993, the term geochemical baseline was officially introduced in the context of the International Geological Correlation

Programme (IGCP) 360 project Global Geochemical Baselines (Salminen and Tarvainen 1997). Since then, many authors used this term as a synonym for natural background concentrations to describe natural variations in element concentration in the surficial environment (Salminen and Tarvainen 1997; Hernandez-Garcia and Custodio 2004; Bech et al. 2005), although some authors such as Reimann and Garrett (2005) disagree with this concept.

Geochemical background is very often incorrectly regarded as a fixed value (mean or median) that represents a hypothetical background concentration without taking into account natural variability (Matschullat et al. 2000). However, it changes both regionally with the basic geology and locally with the type and genesis of the overburden. It is more realistic to view it as a range of values rather than as an absolute value. Quite often background values that are used to compute a pollution factor in water or sediments are simply taken from numerous compilations (e.g., Turekian and Wedepohl 1961; Forstner and Wittmann 1981), from preindustrial core results or pristine water in some distant areas (Jaquet et al. 1982; Kilchmann et al. 2004), or from deep ground water, free of anthropogenic influence (Hernandez-Garcia and Custodio 2004).

The use of regionally constant background levels of some elements when there are variable geochemical conditions could yield an underestimation of the contamination in the affected part of the area and an overestimation elsewhere. In fact, most of the scientific and technical problems associated with the assessment of environmental impact can ultimately be attributed to natural variability inherent in ground water chemistry (Kwiatkowski

¹Corresponding author: Faculty of Mining, Geology and Petroleum Engineering, Department of Geology and Geological Engineering, University of Zagreb, Pierottijeva 6, Zagreb, 10000; znakic@rgn.hr

²Faculty of Mining, Geology and Petroleum Engineering, Department of Geology and Geological Engineering, University of Zagreb, Pierottijeva 6.

Received August 2006, accepted February 2007.

Copyright © 2007 The Author(s)

Journal compilation © 2007 National Ground Water Association.

doi: 10.1111/j.1745-6584.2007.00325.x

1991). Large data sets should be reduced to consider statistically homogeneous zones or strata prior to making any conclusions about the data distribution and natural variability.

Separating the data into homogeneous groups provides advantages, such as improving the chance of detecting an environmental problem. Model-based objective methods for the calculation of threshold values between background and anomalies can be used. These methods include probability graph approaches and depend on the conceptual geochemical model, which recognizes the need for reducing large data sets to statistically homogeneous groups.

A source of variation of an element or compound is related to the heterogeneity of the geologic materials and also to anthropogenic activity. Sometimes, even in statistically homogeneous areas, it is very difficult to clearly identify the samples related to the contamination anomaly as opposed to those solely reflecting background processes. Very often two populations overlap. It is expected that the probability density functions for particular elements are different in anomalous and background samples, but the problem is to recognize these differences with confidence (Sinclair 1991). The outlier concept is a special example of the two populations model, in which the anomalous population amounts to a very small proportion of the total data. Outliers are statistically defined (Hampel et al. 1986; Barnett and Lewis 1994; Reimann et al. 2005) as values belonging to a different population because they originate from another process or source; that is, they could be contaminants. Outliers are generally observations resulting from a secondary process and not extreme values from the background distribution.

This paper describes a spreadsheet macro entitled BACKGROUND, which can be used to calculate geochemical background levels of chemical parameters and to estimate threshold values separating background data from anomalous values. Similar software was introduced by Stanley (1987), who developed the MS-DOS software program Probability Plot (PROBPLOT), an interactive software tool, which allows a user merely to compare the actual cumulative frequency distribution with a theoretical frequency distribution model. Although there is an ever-increasing demand for common statistical tools to distinguish between natural and anthropogenically induced concentrations in ground water, there are few computer codes that integrate model-based objective methods into a widely accessible and user-friendly platform. The macro BACKGROUND was written in Visual Basic (VB) for Applications for use in a widely accessible platform, MS Excel, using algorithms that incorporate two statistical methods described by Matschullat et al. (2000), that is, *the iterative 2- σ technique* and *the calculated distribution function*.

Model-Based Objective Methods for the Calculation of Background Values

Traditionally, model-based subjective methods of threshold determination, including mean plus two standard deviations, were frequently used by some regulators to determine the threshold values for defining action levels or cleanup goals. These methods include some type

of a formal statistical or mathematical model to a set of selected geochemical values, making no assumptions about the form of the data distribution. The problem arises from the fact that regional distribution of an element is often influenced by more than one process, resulting in a multimodal distribution, which could be superimposed, and the use of the mean will give an overestimate of the location for the main body of data (Reimann and Filzmoser 2000).

Model-based objective methods include the probability graph approaches and differ from subjective methods, as thresholds are defined by the data rather than by an arbitrary decision of the researcher (Sinclair 1991). Fundamentally, the probability graph approaches split the overall distribution into distinct components and, in so doing, identify threshold values (Panno et al. 2006). Confidence is increased by a high degree of understanding of data and the processes they have been exposed to, so that the identification of threshold values and multiple populations could be facilitated. Ideally, each population corresponds to a relevant process with its underlying probability density function. Background populations in geological environments have a characteristic probability density function that results from the summation of the natural processes. These background populations can be closely approximated by normal or lognormal density functions (Sinclair 1991). Since the anthropogenic influence occurs as contamination, that is, as a positive anomaly, the search for the geochemical background is reduced to the recognition of the relevant data subcollective and its quantitative description (Matschullat et al. 2000).

Sinclair (1991) and Panno et al. (2006) presented the use of cumulative probability plots, the most cited model-based objective method. They divided data sets into two or more populations separated by inflection points on cumulative probability graphs. Threshold values obtained separate single populations, not taking into account the range of background values with a predetermined confidence level, which represents the natural variation of the background population. Important limitations to this approach are that a minimum of 100 values is needed and the determination of the threshold must be viewed as an estimation procedure in a statistical sense, subject to random and systematic error (Sinclair 1991; Panno et al. 2006).

The statistical methods described in this paper, the iterative 2- σ technique and the calculated distribution function, belong to model-based objective methods. Both methods aim at defining the background and threshold by approaching a normal range. These methods take an actual set of measured data and process the data (i.e., remove values) until a normal distribution is obtained. Hence, what is left in the normally distributed data is background and what was removed are the contaminated ground water or nonbackground values. It is important that background values are estimated based on data derived from a population, which amounts to a large proportion of the total data set. The anomalous population or populations, shown in a distribution function as positive asymmetry of a normal curve, which are represented by higher values in the total data set, need to amount to a small proportion of the total data set.

The iterative $2\text{-}\sigma$ technique constructs an approximated normal distribution around the mode value of the original data set (Matschullat et al. 2000). It is particularly suited for the calculation of the threshold value as the outer limit of background variation and takes into account that both the low as well as high values are used to define anomalies in element concentration. The advantage of this method over the calculated distribution function method is its ability to determine the outliers below the lower limit of normal background fluctuation for some chemical parameter (e.g., low values of dissolved oxygen as indicators of extreme oxygen consumption in an aquifer).

The calculated distribution function specifically aims at defining the upper limit of normal background concentrations. It is convenient for use if anthropogenic activities tend to lead to enrichments in natural systems, causing positive anomalies shown in a distribution function as positive asymmetry of a normal curve. The lower values should thus be free from anthropogenic influence (Matschullat et al. 2000).

A successful application of both methods does not require normally or lognormally distributed total (combined) data, and they can be applied to relatively small data sets ($n > 30$). This threshold of sample size separates small-sample statistics from large-sample statistics, where normal distribution can be used as an approximation (Davis 2002). If the number of the remaining data (background data) equals or exceeds four, then a variation devised by Lilliefors (1967) to the use of the Kolmogorov-Smirnov procedure for testing the fit of a background data to a normal distribution can be applied.

The iterative $2\text{-}\sigma$ technique and the calculated distribution function method are applicable to unimodal and skewed distributions. The calculated distribution function method is even applicable to the overlapping polymodal distribution if the data in the background population, which represents lower values in the original data set distribution, are dominant. Both methods can also be applied to scattered distributions if the Lilliefors test statistic T is lower than the critical value of T . The main advantage of these techniques over other methods is that they calculate the normal range of background values (true value of background is within $\text{mean} \pm 2\sigma$ range) with 95% confidence.

The limitations in use are that both methods work well only if the generally predominant background data are used to define background populations. If the number of contaminated samples is greater than or equal to background samples, the Lilliefors test statistic T is greater than the critical value of T and the obtained background range is overestimated.

The disadvantage of the iterative $2\text{-}\sigma$ technique is that it cannot be applied when the distribution is polymodal. If the distribution is, for example, bimodal, the mean might fall between the two peaks of the distribution and the large value of the standard deviation causes an overestimation of the background range.

Although the problem of identifying background values is to a great extent a statistical problem, it is important to consider the presence of anomalous samples,

spatial scale, location, and the kind of sample material, as well as the purpose for which the background values are needed. Geochemical data very often show spatial dependence, as a consequence of an inadequate sampling strategy, and spatially dependent data are correlated and are generally not normally distributed. Prior to attempting to estimate background values by applying described methods, the use of a combination of exploratory data analysis tools (e.g., histogram, box plot, one-dimensional and two-dimensional scattergram, cumulative distribution function) is needed to give an insight into the data structure and possible data errors.

VB Macro Design

The VB macro consists of algorithms, which automate the geochemical data processing using concepts of two statistical methods, the iterative $2\text{-}\sigma$ technique and the calculated distribution function. In addition to algorithms used for geochemical background calculation, spreadsheets also contain an algorithm for creating plots of the frequency distribution of the original data set. This algorithm is intended for use prior to calculating geochemical background in order to examine the frequency distribution of the original data set. The aim is to visually detect possible errors involved in the sampling procedure, sample preparation, and analysis and to estimate whether presented statistical methods are adequate for calculating geochemical background values of the original data set, that is, to establish whether the distribution is unimodal or polymodal.

The first computational step in an algorithm based on the iterative $2\text{-}\sigma$ technique concept is data preparation (Figure 1), which implies the storing of the data set in a new spreadsheet. In the next step ("Calculate Mean and Standard Deviation (σ)" in Figure 1), the macro calculates the mean and standard deviation from the prepared data set. This is followed by calculating the $\text{mean} \pm 2\sigma$ range. The data set is inspected to verify that all values lie within the calculated range. All values that lie outside the range $\text{mean} \pm 2\sigma$ are cleared. These steps are repeated ("Iteration loop" in Figure 1) until all values lie within the $\text{mean} \pm 2\sigma$ range. Calculated background levels ($\text{mean} \pm 2\sigma$ range), threshold limit, and data not considered to be background values are then entered into the spreadsheet. In the next step, the macro calculates bin (bin interval is defined by the user through the VB macro form), frequency, cumulative and cumulative (%) values for the original data set, and calculated background values. These values are later used for the histogram presentation. To test the hypothesis, the macro calculates normal cumulative distribution for the calculated mean and standard deviation. The Lilliefors test (an adaptation of the Kolmogorov-Smirnov test) for level of significance $\alpha = 0.5$ is used for testing goodness of fit to a normal distribution (Lilliefors 1967). In the final step, the macro produces a chart showing histograms for original and calculated background values, cumulative distribution for original and calculated background values, and normal cumulative distribution ("Chart" in Figure 1). The Lilliefors test statistic values (T_{crit} and T) are also entered into the spreadsheet.

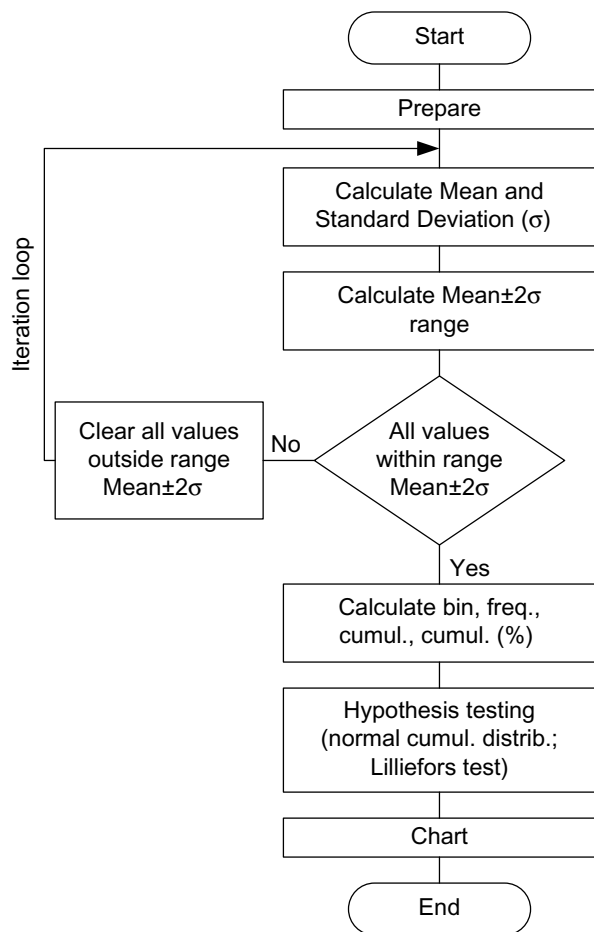


Figure 1. Macro structure for the iterative 2- σ technique.

An algorithm based on the calculated distribution function concept shares the first computational step with the previously described algorithm (“Prepare” in Figure 2). After the data preparation, the macro calculates the median value for the data set and all values greater than the median are then cleared. In the next step, the macro mirrors all remaining values against the calculated median value in the following manner: mirrored value = median – data set value + median. The macro then calculates mean and standard deviation from the prepared data set and the mean $\pm 2\sigma$ range. After calculating the mean $\pm 2\sigma$ range, all values that lie outside the calculated range are cleared (“Clear all values outside range Mean $\pm 2\sigma$ ” in Figure 2). Following the computational steps, from entering the calculated background levels, threshold limit and data not considered to be background values in the spreadsheet are shared with the previously described algorithm. The macro runs on Windows operating systems and accompanied versions of Microsoft Excel.

Use of Methods

To illustrate the use of the iterative 2- σ technique and the calculated distribution function method through an example, the VB macro is used with field data that were sampled to include a higher percentage of data unaffected by human impact. Ambient or present-day

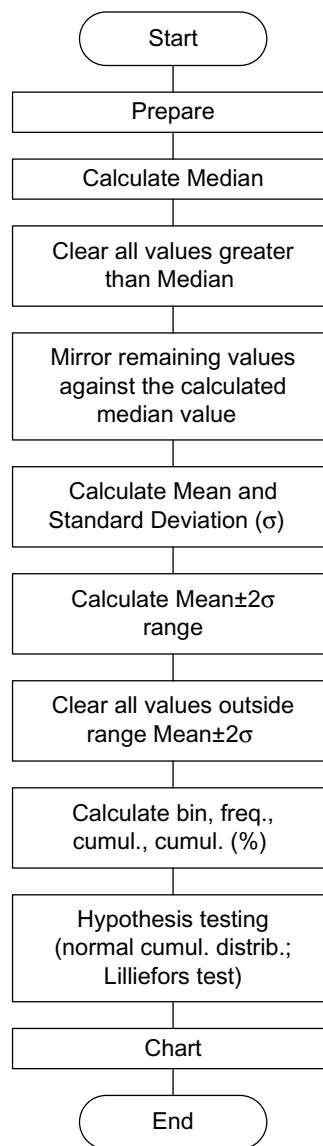


Figure 2. Macro structure for the calculated distribution function.

background concentrations of nitrate in shallow ground water of an alluvial aquifer in the area of Zagreb, the capital of Croatia, calculated by the VB macro (Figures 3 and 4) show that both methods give similar values of background levels and threshold limits.

As a way of assessing the results, the Lielliefors test is used. The test statistic T in both examples is less than the critical value of the Lilliefors test statistic T_{crit} . The identified background values distribution fits the normal distribution well, indicating that the results of both methods under present circumstances are satisfactory.

Conclusions

To distinguish between natural and anthropogenically contaminated concentrations in ground water, geochemical background needs to be calculated. With the VB macro BACKGROUND, it is possible to integrate the model-based objective approach into an automated calculation of geochemical background values of chemical

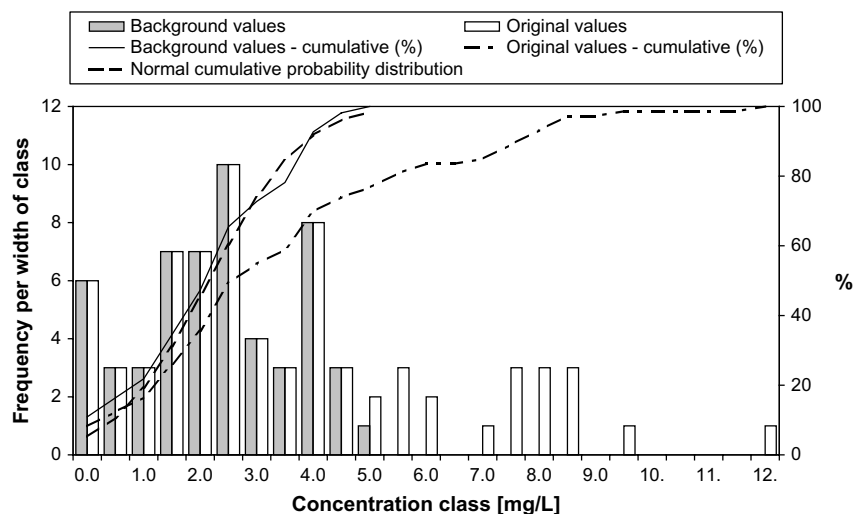


Figure 3. Background concentrations of nitrate calculated using the iterative $2\text{-}\sigma$ technique (background levels: 0/4.8; threshold limit: 4.8; Lilliefors test statistic [$\alpha = 0.5$]: $T_{\text{crit.}} = 0.119$, $T = 0.064$).

parameters and to estimate threshold values dividing background data from anomalies. BACKGROUND uses algorithms that incorporate two statistical methods, the iterative $2\text{-}\sigma$ technique and the calculated distribution function. The iterative $2\text{-}\sigma$ technique is particularly suited for the calculation of the threshold value as the outer limit of background variation and has the ability to determine the outliers below the lower limit of normal background fluctuation for some chemical parameter. The calculated distribution function is convenient for use if anthropogenic activities tend to lead to enrichments in natural systems, causing the distribution function of the original data set to be skewed toward higher values. If data unaffected by human impacts dominate in a sampled data set, the calculated distribution function method could even be applied to a polymodal distribution.

The user-friendly VB macro BACKGROUND provides information on the normal background concentrations with little effort using a widely accessible platform (i.e., MS Excel), which also allows data to be plotted and visualized easily in the same Excel environment that researchers often use to organize and evaluate data. Additionally, it offers the possibility for automated data processing contained in a chemical database, which enables an automated generation of background range and threshold values for chemical parameters.

Software Availability

The Excel spreadsheet with its open access VB macro may be requested charge-free via e-mail from the contributing author of this article (kposavec@rgn.hr).

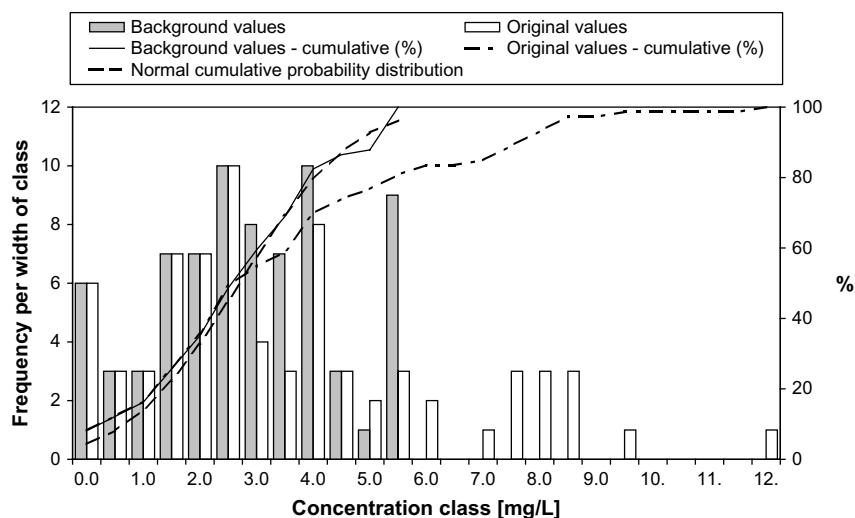


Figure 4. Background concentrations of nitrate calculated using the calculated distribution function (background levels: 0–5.9; threshold limit: 5.9; Lilliefors test statistic [$\alpha = 0.5$]: $T_{\text{crit.}} = 0.103$, $T = 0.049$).

Acknowledgments

The authors are grateful for the helpful review comments by Samuel Panno, Allan Crowe, and the two anonymous reviewers. Thanks also to Mary P. Anderson for editing the paper for English usage and journal format.

References

- Barnett, V., and T. Lewis. 1994. *Outliers in Statistical Data*, 3rd ed. New York: John Wiley & Sons Inc. (Cited by Reimann et al. 2005).
- Bech, J., P. Tume, L. Longan, and F. Reverter. 2005. Baseline concentrations of trace elements in surface soils of the Torreles and Saint Climent municipal districts (Catalonia, Spain). *Environmental Monitoring and Assessment* 108, no. 1–3: 309–322.
- Davis, J.C. 2002. *Statistics and Data Analysis in Geology*, 3rd ed. New York: John Wiley & Sons Inc.
- Forstner, U., and G.T.W. Wittmann. 1981. *Metal Pollution in the Aquatic Environment*. Berlin, Heidelberg, Germany: Springer-Verlag.
- Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw, and W. Stahel. 1986. *Robust Statistics. The Approach Based on Influence Functions*. New York: John Wiley & Sons. (Cited by Reimann et al. 2005).
- Hawkes, H.E., and J.S. Webb. 1962. *Geochemistry in Mineral Exploration*. New York: Harper.
- Hernandez-Garcia, M.E., and E. Custodio. 2004. Natural baseline quality of Madrid Tertiary Detrital Aquifer groundwater (Spain): A basis for aquifer management. *Environmental Geology* 46, no. 2: 173–188.
- Jaquet, J.M., E. Davaud, F. Rapin, and J.P. Vernet. 1982. Basic concepts and associated statistical methodology in the geochemical study of lake sediments. *Hydrobiologia* 91–92, no. 1: 139–146.
- Kilchmann, S., H.N. Waber, A. Parriaux, and M. Bensimon. 2004. Natural tracers in recent groundwaters from different Alpine aquifers. *Hydrogeology Journal* 12, no. 6: 643–661.
- Kwiatkowski, R.E. 1991. Statistical needs in national water quality programs. *Environmental Monitoring and Assessment* 17, no. 2–3: 253–271.
- Lilliefors, H.W. 1967. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association* 62, no. 318: 399–402. (Cited by Davis 2002).
- Matschullat, J., R. Ottenstein, and C. Reimann. 2000. Geochemical background—Can we calculate it? *Environmental Geology* 39, no. 9: 173–188.
- Panno, S.V., W.R. Kelly, A.T. Martinsek, and K.C. Hackley. 2006. Estimating background and threshold nitrate concentrations using probability graphs. *Ground Water* 44, no. 5: 697–709.
- Reimann, C., and P. Filzmoser. 2000. Normal and lognormal data distribution in geochemistry: Death of a myth. Consequences for the statistical treatment of geochemical and environmental data. *Environmental Geology* 39, no. 9: 1001–1014.
- Reimann, C., P. Filzmoser, and R.G. Garrett. 2005. Background and threshold: Critical comparison of methods of determination. *Science of the Total Environment* 346, no. 1–3: 1–16.
- Reimann, C., and R.G. Garrett. 2005. Geochemical background—Concept and reality. *Science of the Total Environment* 350, no. 1–3: 12–27.
- Salminen, R., and T. Tarvainen. 1997. The problem of defining geochemical baselines. A case study of selected elements and geological materials in Finland. *Journal of Geochemical Exploration* 60, no. 1: 91–98.
- Sinclair, A.J. 1991. A fundamental approach to threshold estimation in exploration geochemistry: Probability plot revisited. *Journal of Geochemical Exploration* 41, no. 1–2: 1–22.
- Stanley, C.R. 1987. *PROBPLOT: An Interactive Computer Program to Fit Mixtures of Normal (or Lognormal) Distributions with Maximum Likelihood Procedures Spec. Vol. 14*. Rexdale, Ontario, Canada. Association of Exploration Geochemists. (Cited by Sinclair 1991.)
- Turekian, K., and K.H. Wedepohl. 1961. Distribution of the elements in some major units of the earth's crust. *Geological Society of America Bulletin* 72, no. 2: 175–192. (Cited by Jaquet et al. 1982).