



Background and threshold: critical comparison of methods of determination

Clemens Reimann^{a,*}, Peter Filzmoser^b, Robert G. Garrett^c

^a*Geological Survey of Norway, N-7491 Trondheim, Norway*

^b*Institute of Statistics and Probability Theory, Vienna University of Technology, Wiedner Hauptstr. 8-10, A-1040 Wien, Austria*

^c*Geological Survey of Canada, Natural Resources Canada, 601 Booth Street, Ottawa, Ontario, Canada K1A 0E8*

Received 7 June 2004; received in revised form 1 November 2004; accepted 12 November 2004

Available online 4 February 2005

Abstract

Different procedures to identify data outliers in geochemical data are reviewed and tested. The calculation of [mean \pm 2 standard deviation (sdev)] to estimate threshold values dividing background data from anomalies, still used almost 50 years after its introduction, delivers arbitrary estimates. The boxplot, [median \pm 2 median absolute deviation (MAD)] and empirical cumulative distribution functions are better suited for assisting in the estimation of threshold values and the range of background data. However, all of these can lead to different estimates of threshold. Graphical inspection of the empirical data distribution using a variety of different tools from exploratory data analysis is thus essential prior to estimating threshold values or defining background. There is no good reason to continue to use the [mean \pm 2 sdev] rule, originally proposed as a ‘filter’ to identify approximately 2½% of the data at each extreme for further inspection at a time when computers to do the drudgery of numerical operations were not widely available and no other practical methods existed. Graphical inspection using statistical and geographical displays to isolate sets of background data is far better suited for estimating the range of background variation and thresholds, action levels (e.g., maximum admissible concentrations—MAC values) or clean-up goals in environmental legislation.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Background; Threshold; Mean; Median; Boxplot; Normal distribution; Cumulative probability plot; Outliers

1. Introduction

The detection of data outliers and unusual data behaviour is one of the main tasks in the statistical analysis of geochemical data. As early as 1962, several procedures had been recommended for selecting

* Corresponding author. Tel.: +47 73 904 307; fax: +47 73 921 620.

E-mail addresses: Clemens.Reimann@ngu.no (C. Reimann), P.Filzmoser@tuwien.ac.at (P. Filzmoser), garrett@gsc.NRCan.gc.ca (R.G. Garrett).

threshold levels in order to identify outliers (Hawkes and Webb, 1962):

- (1) Carry out an orientation survey to define a local threshold against which anomalies can be judged;
- (2) Order the data and select the top 2½% of the data for further inspection if no orientation survey results are available; and
- (3) For large data sets, which cannot easily be ordered (at that time—1962; with today's PCs this is no longer a problem), use [mean±2 sdev] (sdev: standard deviation) to identify about 2½% of upper (or lower) extreme values for further inspection.

Recommendation (1) requires additional field-work and resources, and consequently such work is rarely carried out post-survey—by definition orientation surveys should be undertaken prior to any major survey as a necessary part of survey design and protocol development. In the second procedure the 2½% could be questioned: why should there be 2½% outliers and not 5% or 10%, or no outliers at all? However, 2½% is a reasonable number of samples for 'further inspection' and is, as such, a 'practical' approach, and leads to a similar percentage of the data as the third procedure identifies. This third procedure appears the most rigorous and delivers estimates based on a mathematical calculation and an underlying assumption that the data are drawn from a normal distribution. Statisticians use this approach to identify the extreme values in a normal distribution. The calculation will result in about 4.6% of the data belonging to a normal distribution being identified as extreme values, 2.3% at the lower and 2.3% at the upper ends of the distribution. However, are these extreme values defined by a statistical rule really the 'outliers' sought by geochemists and environmental scientists? Will this rule provide reliable enough limits for the background population to be of use in environmental legislation?

Extreme values are of interest in investigations where data are gathered under controlled conditions. A typical question would be "what are the extreme tall and small persons in a group of girls age 12" or "what is the 'normal' range of weight and size in newborns"?

In this case extreme values are defined as: *values in the tails of a statistical distribution.*

In contrast, geochemists are typically interested in outliers as indicators of rare geochemical processes. In such cases, these outliers are not part of one and the same distribution. For example, in exploration geochemistry samples indicating mineralisation are the outliers sought. In environmental geochemistry the recognition of contamination is of interest. Outliers are statistically defined as (Hampel et al., 1986; Barnett and Lewis, 1994): *values belonging to a different population because they originate from another process or source, i.e. they are derived from (a) contaminating distribution(s).* In such a case, the [mean±2 sdev] rule cannot deliver a relevant threshold estimate.

In exploration geochemistry values within the range [mean±2 sdev] were often defined as the 'geochemical background', recognizing that background is a range and not a single value (Hawkes and Webb, 1962). The exact value of mean+2 sdev is still used by some workers as the 'threshold', differentiating background from anomalies (exploration geochemistry), or for defining 'action levels' or 'clean-up goals' (environmental geochemistry). Due to imprecise use of language, the threshold is often also named the background value. Such a threshold is in fact an estimate of the upper limit of background variation. In geochemistry, traditionally, low values, or lower outliers, have not been seen as important as high values; this is incorrect, low values can be important. In exploration geochemistry they may indicate alteration zones (depletion of certain elements) related to nearby mineral accumulations (occurrences). In environmental geochemistry, food micronutrient deficiency related health problems may, on a worldwide scale, be even more important than toxicity.

Interestingly, for many continuous distributions, the [mean±2 sdev] rule results in about 5% of extreme values being identified, even for a *t* or chi-squared distribution (Barnett and Lewis, 1994). However, a problem with the application of this rule is that it is only valid for the true population parameters. In practice, the empirical (calculated) sample mean and standard deviation are used to estimate the population mean and standard deviation. But the empirical estimates are strongly influenced by

extreme values, whether derived from the data distribution or from a second (contaminating) distribution. The implication is that the $[\text{mean} \pm 2 \text{ sdev}]$ rule is not (and never was) valid. To avoid this problem extreme, i.e. ‘obvious’ outliers, are often removed from the data prior to the calculation, which introduces subjectivity. Another method is to first log-transform the data to minimize the influence of the outliers and then do the calculation. However, the problem that thresholds are frequently estimated based on data derived from more than one population remains. Other methods to identify data outliers or to define a threshold are thus needed.

To select better methods to deal with geochemical data, the basic properties of geochemical data sets need to be identified and understood. These include:

- The data are spatially dependent (the closer two sample sites, the higher the probability that the samples show comparable analytical results—however, all classical statistical tests assume independent samples).
- At each sample site a multitude of different processes will have an influence on the measured analytical value (e.g., for soils these include: parent material, topography, vegetation, climate, Fe/Mn-oxyhydroxides, content of organic material, grain size distribution, pH, mineralogy, presence of mineralization or contamination). For most statistical tests, it is necessary that the samples come from the same distribution—this is not possible if different processes influence different samples.
- Geochemical data, like much natural science data, are imprecise, they contain uncertainty unavoidably introduced at the time of sampling, sample preparation and analyses.

Thus starting geochemical data analysis with statistical tests based on assumptions of normality, independence, and identical distribution may not be warranted; do better-suited tools for the treatment of geochemical data exist? Methods that do not strongly build on statistical assumptions should be first choice. For example, the arithmetic mean could be replaced by the median and the standard deviation by the median absolute deviation (MAD or medmed), defined as the median of the absolute deviations from the median of all data (Tukey, 1977). These estimators

are robust against extreme values. Another solution would be to use the boxplot (Fig. 1) for the identification of extreme values (Tukey, 1977). The boxplot divides the ordered values of the data into four ‘equal’ parts, firstly by finding the median (displayed as a line in the box—Fig. 1), and then by doing the same for each of the halves. These upper and lower quartiles, often referred to as upper and lower ‘hinges’, define the central box, which thus contains approximately 50% of the data. The ‘inner fence’ is defined as the box extended by 1.5 times the length of the box towards the maximum and the minimum. The upper and lower ‘whiskers’ are then drawn from each end of the box to the farthest

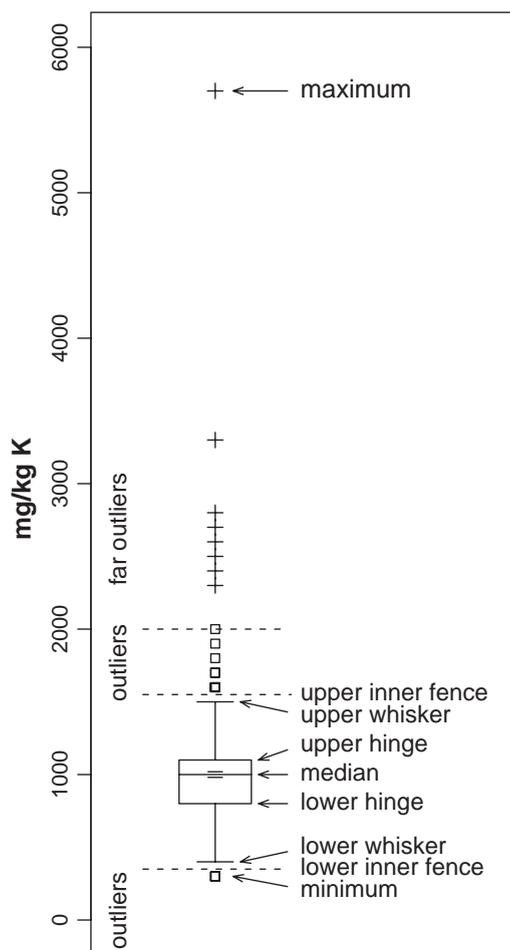


Fig. 1. Tukey boxplot for potassium (K%) concentrations in the O-horizon of podzols from the Kola area (Reimann et al., 1998). For definitions of the different boundaries displayed see text.

observation inside the inner fence. This is defined algebraically, using the upper whisker as an example, as

Upper inner fence (UIF)

$$= \text{upper hinge}(x) + 1.5 \cdot \text{HW}(x) \quad (1)$$

Upper whisker = $\max(x[x < \text{UIF}])$ (2)

where HW (hinge width) is the difference between the hinges (upper hinge–lower hinge), approximately equal to the interquartile range (depending on the sample size), i.e. Q3–Q1 (75th–25th percentile); the square brackets [...] indicate that subset of values that meet the specified criterion. In the instance of lognormal simulations, the fences are calculated using the log values in Eq. (1), and then anti-logged for use in Eq. (2). Any values beyond these whiskers are defined as data outliers and marked by a special symbol (Fig. 1). At a value of the upper hinge plus 3 times the hinge width (lower hinge minus 3 times the hinge width), the symbols used to represent outliers change (e.g., from small squares to large crosses—Fig. 1) to mark ‘far outliers’, i.e. values that are very unusual for the data set as a whole. Because the construction of the box is based on quartiles, it is resistant to up to 25% of data outliers at either end of the distribution. In addition, it is not seriously influenced by widely different data distributions (Hoaglin et al., 2000).

This paper investigates how several different threshold estimation methods perform with true normal and lognormal data distributions. Outliers derived from a second distribution are then superimposed upon these distributions and the tests repeated. Graphical methods to gain better insight into data distribution and the existence and number of outliers in real data sets are presented. Based on the results of these studies, guides for the detection of outliers in geochemical data are proposed.

It should be noted that in the literature on robust statistics, many other approaches for outlier detection have been proposed (e.g., Huber, 1981; Hampel et al., 1986; Rousseeuw and Leroy, 1987; Barnett and Lewis, 1994; Dutter et al., 2003). In this paper we focus on the performance of the methods that are, or have been, routinely used in geochemistry: [mean±2

sdev] rule and the boxplot inner fence. We have added the [median±2 MAD] procedure because, in addition to the boxplot, this is the most easily understood robust approach, is a direct analogy to [mean±2 sdev], and the estimators are usually offered in widely available statistical software packages.

2. Performance of different outlier detection methods

When analysing real data and estimating the limits of background variation using different procedures, the [median±2 MAD] method will usually deliver the lowest threshold and thus identify the highest number of outliers, followed by the boxplot. The boxplot-defined threshold, the inner fence, is in most cases close to, but lower than, the threshold obtained from log-transformed data (Table 1) using the [mean±2 sdev] rule. The examples in Table 1 demonstrate the wide range of thresholds that can be estimated depending on the method and transformation chosen. The question is “which (if any) of these estimates is the most credible?” This question cannot be answered at this point, because the true statistical and spatial distribution of the ‘real world’ data is unknown. However, the question is significant because such statistically derived, and maybe questionable, estimates are used, for example, to guide mineral exploration expenditures and establish criteria for environmental regulation.

To gain insight into the behaviour of the different threshold estimation procedures simulated data from a known distribution were used. For simulated normal distributions (here we considered standard normal distributions with mean 0 and standard deviation 1) with sample sizes 10, 50, 100, 500, 1000, 5000 and 10000, the percentage of detected outliers for each of the three suggested methods ([mean±2 sdev], [median±2 MAD] and boxplot inner fence) were computed. This was replicated for each sample size 1000 times. (For the simulation, the statistics software package R was used, which is freely available at <http://cran.r-project.org/>). Fig. 2A shows that the classical method, [mean±2 sdev], detects 4.6% extreme values for large N —as expected from theory. However, for smaller N ($N < 50$), the percentage declines towards less than 3%. The [median±2

Table 1

Mean, standard deviation (sdev), median, median absolute deviation (MAD), (log-transformed data) and results of the definition of an upper threshold via: [mean \pm 2 sdev], [median \pm 2 MAD], the boxplot (Tukey, 1977) and cumulative probability plots (see Fig. 7) for selected variables for three example data sets: BSS data, agricultural soils from Northern Europe, ploughed (TOP) 0–20 cm layer, <2 mm fraction, $N=750$, 1,800,000 km² (Reimann et al., 2003); Kola, O-horizon of podzol profiles, <2 mm fraction, $N=617$, 180,000 km² (14); and Walchen, B-horizon of forest soils, <0.18 mm fraction, 100 km² (Reimann, 1989)

	Mean		sdev		Median		MAD		[Mean+2 sdev]		[Median+2 MAD]		Upper whisker		98th percentile	Outer limit CDF
	Natural	Log10	Natural	Log10	Natural	Log10	Natural	Log10	Natural	Anti-log	Natural	Anti-log	Natural	Anti-log		
<i>BSS TOP</i>																
As	2.6	0.293	2.4	0.325	1.9	0.279	1.2	0.294	7.4	8.8	4.3	7.4	5.8	11	9	8 (11)
Cu	13	1.01	11	0.307	9.8	0.992	6.6	0.306	36	42	23	40	31	64	43	38 (51)
Ni	11	0.907	8.1	0.301	8	0.904	5.3	0.316	27	35	19	34	36	53	34	45
Pb	18	1.22	8.1	0.172	17	1.22	5.7	0.152	34	37	28	34	32	42	38	37
Zn	42	1.53	30	0.293	33	1.52	20	0.291	101	129	74	128	100	200	121	155
<i>Kola O-horizon</i>																
As	1.6	0.094	2.5	0.245	1.2	0.065	0.46	0.174	6.6	3.8	2.1	2.6	2.5	3.4	6.4	2.1 (9)
Cu	44	1.12	245	0.432	9.7	0.986	5.1	0.267	535	96	20	33	35	76	241	13 (1000)
Ni	51	1.12	119	0.565	9.2	0.963	7.7	0.455	450	177	25	75	54	258	395	10 (500)
Pb	24	1.29	49	0.208	19	1.27	7.4	0.185	122	52	34	44	43	55	48	44 (70)
Zn	48	1.66	18	0.157	46	1.66	15	0.143	84	93	76	89	88	107	93	85
<i>Walchen B-horizon</i>																
As	46	1.47	83	0.368	30	1.48	19	0.294	211	163	69	116	87	179	207	90 (160)
Cu	32	1.42	23	0.291	28	1.45	15	0.23	79	101	58	81	69	110	84	(40) 80
Ni	34	1.43	23	0.291	30	1.48	19	0.294	81	120	69	116	80	153	104	(42) 90
Pb	39	1.41	53	0.368	23	1.36	16	0.32	145	139	56	100	81	202	200	90 (350)
Zn	74	1.81	34	0.253	73	1.86	30	0.171	141	209	132	160	152	200	139	145 (175)

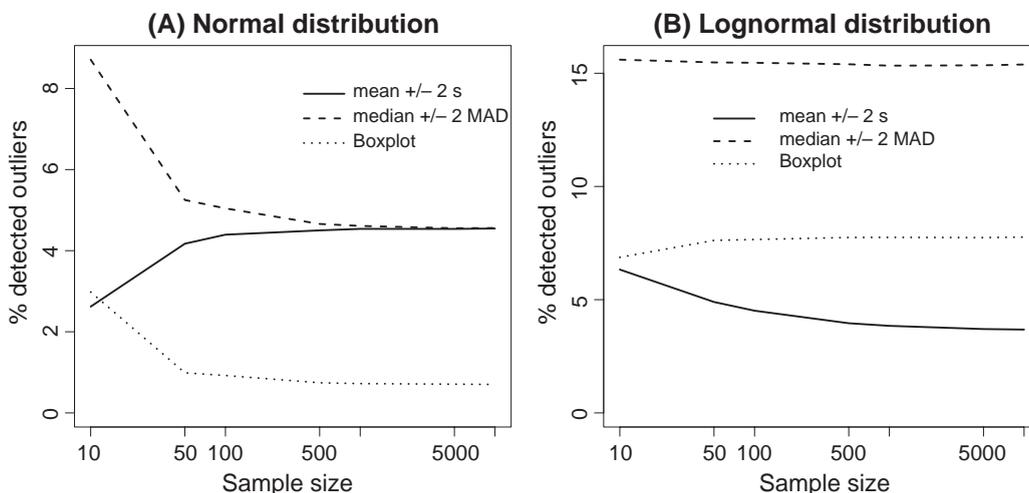


Fig. 2. Average percentage of outliers detected by the rules [mean \pm 2 sdev], [median \pm 2 MAD], and the boxplot method. For several different sample sizes ($N=10$ to 10000), the percentages were computed based on 1000 replications of simulated normally distributed data (A) and of simulated lognormally distributed data (B).

MAD] procedure shows the same behaviour for large N ($N>500$). For smaller N , the percentage is higher and increases dramatically for $N<50$. The boxplot inner fence detects about 3% extreme values for small sample sizes ($N<50$), and at larger sample sizes less than 1% of the extremes (Fig. 2A).

Geochemical data rarely follow a normal distribution (Reimann and Filzmoser, 2000) and many authors claim that they are close to a lognormal distribution (see discussion and references in Reimann and Filzmoser, 2000). Thus the whole exercise was repeated for a lognormal distribution, i.e. the logarithm of the generated values is normally distributed (we considered standard normal distribution). Fig. 2B shows that the classical rule identifies about the same percentage of extreme values as for a normal distribution. The reason for this is that both the mean and standard deviation are inflated due to the presence of high values in the lognormal data, thus resulting in inflated fence values. However, the classical rule identifies a higher number of extreme values for small N ($N<50$). The [median \pm 2 MAD] procedure delivers a completely different outcome. Here about 15.5% of extremes were identified for all sample sizes. The boxplot finds around 7% extreme values, with a slight decline for small sample sizes ($N<50$). The reason for the large proportion of extreme values identified is that the robust estimates of location (median) and

scale (MAD or hinge width) are relatively uninfluenced by the extreme values of the lognormal data distribution. Both location and scale are low relative to the mean and standard deviation, resulting in lower fence values and higher percentages of identified extreme values.

Results from these two simulation exercises explain the empirical observations in Table 1. The [median \pm 2 MAD] procedure always results in the lowest threshold value, the boxplot in the second lowest, and the classical rule in the highest threshold. Because geochemical data are in general right-skewed and often closely resemble a lognormal distribution, the second simulation provides the explanation for the observed behaviour.

Based on the results of the two simulation exercises, one can conclude that the data should approach a symmetrical distribution before any threshold estimation methods are applied. A graphical inspection of geochemical data is thus necessary as an initial step in data analysis. In the case of lognormally distributed data, log-transformation results in a symmetric normal distribution (see discussion above). The percentages of detected extreme values using [mean \pm 2 sdev] or [median \pm 2 MAD] are unrealistically high without a symmetrical distribution. Only percentiles (recommendation (2) of Hawkes and Webb, 1962) will always deliver the same number

of outliers. In the above simulations data were sampled from just one (normal or lognormal) distribution. True outliers, rather than extreme values, will be derived from a different process and not from the normal distributions comprising the majority of the data in these simulations. In the case of a sampled normal distribution, the percentage of outliers should thus approach zero and is independent of the number of extreme values. In this respect, the boxplot performs best as the fences are based on the properties of the middle 50% of the data.

How do the outlier identification procedures perform when a second, outlier, distribution is superimposed on an underlying normal (or lognormal) distribution? To test this more realistic case a fixed sample size of $N=500$ was used, of which different percentages (0–40%) were outliers drawn from a different population (Fig. 3). Note that the cases with high percentages of outliers (>5–10% outliers) are typical for a situation where an orientation survey was suggested as the only realistic solution (Hawkes and Webb, 1962). For the first simulation, both the background data and outlier (contaminating) distributions were normal, with means of 0 and 10 and standard deviation 1. Thus the outliers were clearly divided from the simulated background normal distribution. Such a clear separation is an over-simplifi-

cation for the purpose of demonstration. For each percentage of superimposed outliers, the simulation was replicated 1000 times, and the average percentage of detected outliers for the three methods was computed.

In Fig. 3A, the average percentage of detected outliers should follow the 1:1 line through the plot, indicating equality to the percentage of superimposed outliers. The $[\text{mean} \pm 2 \text{ sdev}]$ rule overestimates the number of outliers for very small (<2%) percentages of superimposed outliers, detects the expected number of outliers to about 10%, and then seriously underestimates the number of outliers. The $[\text{median} \pm 2 \text{ MAD}]$ procedure overestimates the number of outliers to almost 20% and finds the expected number of outliers above this value—theoretically up to 50% superimposed outliers. The boxplot performs best up to 25% contamination, and then breaks down.

Fig. 3B shows the results if lognormal distributions are simulated with means of 0 and 10 and standard deviation 1. Here the $[\text{mean} \pm 2 \text{ sdev}]$ rule performs extremely poorly. The $[\text{median} \pm 2 \text{ MAD}]$ procedure strongly overestimates the number of outliers and approaches correct estimates of contamination only at a very high percentage of outliers (above 35%). The boxplot inner fence increasingly overestimates the number of outliers at low contamination, approaches

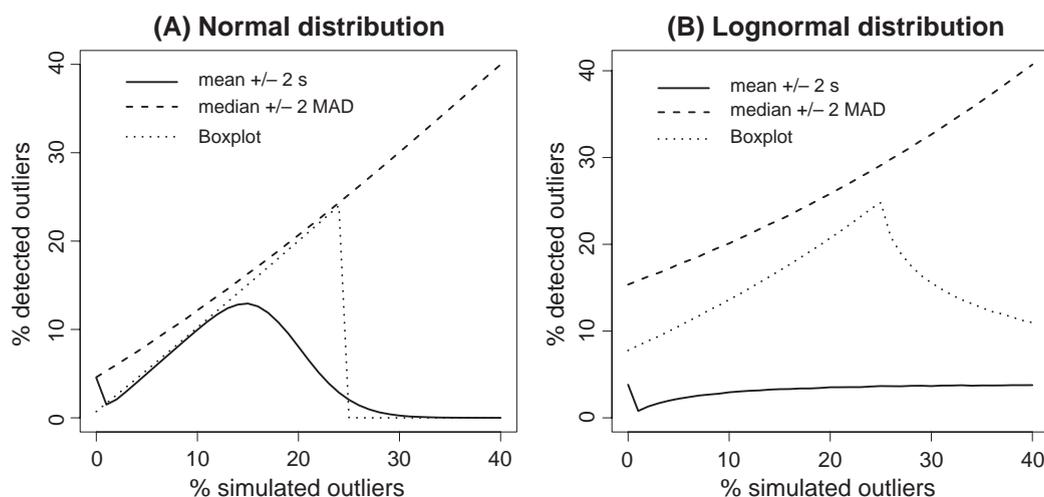


Fig. 3. Average percentage of outliers detected by the rules $[\text{mean} \pm 2 \text{ sdev}]$, $[\text{median} \pm 2 \text{ MAD}]$, and the boxplot method. Simulated standard normally distributed data (A) and simulated standard lognormally distributed data (B) were both contaminated with (log)normally distributed outliers with mean 10 and variance 1, where the percentage of outliers was varied from 0 to 40% for a constant sample size of 500. The computed percentages are based on 1000 replications of the simulation.

the expected number at around 20% and breaks down at 25%.

In a further experiment, both the background data and outlier (contaminating) distributions were normal, with means of 0 and 5, respectively, and standard deviations of 1. Hence, there is no longer a clear separation between background and outliers. The percentage of outliers relative to the 500 background data values was again varied from 0 to 40%, and each simulation was replicated 1000 times. Fig. 4 demonstrates that in this case the results of the $[\text{mean} \pm 2 \text{ sdev}]$ rule are no longer useful. Not surprisingly, only at around 5% of outliers does it give reliable results. The boxplot can be used up to at most 15% outliers. The $[\text{median} \pm 2 \text{ MAD}]$ procedure increasingly overestimates the number of outliers for small percentages and approaches the correct number at 15%, but breaks down at 25%.

Results of the simulations suggest that the boxplot performs best if the percentage of outliers is in the range of 0–10 (max. 15%); above this value the $[\text{median} \pm 2 \text{ MAD}]$ procedure can be applied. The classical $[\text{mean} \pm 2 \text{ sdev}]$ rule only yields reliable results if no outliers exist and the task really lies in the definition of extreme values (i.e. almost never for

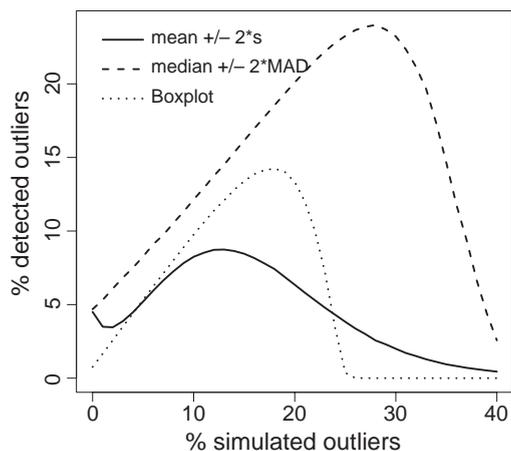


Fig. 4. Average percentage of outliers detected by the rules $[\text{mean} \pm 2 \text{ sdev}]$, $[\text{median} \pm 2 \text{ MAD}]$, and the boxplot method. Simulated standard normally distributed data with mean zero and variance 1 were contaminated with normally distributed outliers with mean 5 and variance 1, where the percentage of outliers was varied from 0 to 40% for a constant sample size of 500. The computed percentages are based on 1000 replications of the simulation.

geochemical data). It is thus necessary to study the empirical distribution of the data to gain a better understanding of the possible existence and number of outliers before any particular procedure is applied. An important difference between the boxplot and the other two procedures is the fact that the outlier limits (fences) of the boxplot are not necessarily symmetric around the centre (median). They are only symmetric if the median is exactly midway between the hinges, the upper and lower quartiles. This difference is more realistic for geochemical data than the assumption of symmetry.

3. Data distribution

There has been a long discussion in geochemistry whether or not data from exploration and environmental geochemistry follow a normal or lognormal distribution (see Reimann and Filzmoser, 2000). The discussion was fuelled by the fact that the $[\text{mean} \pm 2 \text{ sdev}]$ rule was extensively used to define the range of background concentrations and differentiate background from anomalies. Recently, it was again demonstrated that the majority of such data follow neither a normal nor a lognormal distribution (Reimann and Filzmoser, 2000).

In the majority of cases, geochemical distributions for minor and trace elements are closer to lognormal (strong right-skewness) than to normal. When plotting histograms of log-transformed data, they often approach the bell shape of a Gaussian distribution, which is then taken as a sufficient proof of lognormality. Statistical tests, however, indicate that in most cases the data do not pass as drawn from a lognormal distribution (Reimann and Filzmoser, 2000).

As demonstrated, the boxplot (Tukey, 1977) is another possibility for graphically displaying the data distribution (Fig. 1). It provides a graphical data summary relying solely on the inherent data structure and not on any assumptions about the distribution of the data. Besides outliers it shows the centre, scale, skewness and kurtosis of a given data set. It is thus ideally suited to graphically compare different data (sub)sets.

Fig. 5 shows that a combination of histogram, density trace, one-dimensional scattergram and boxplot give a much improved insight to the data

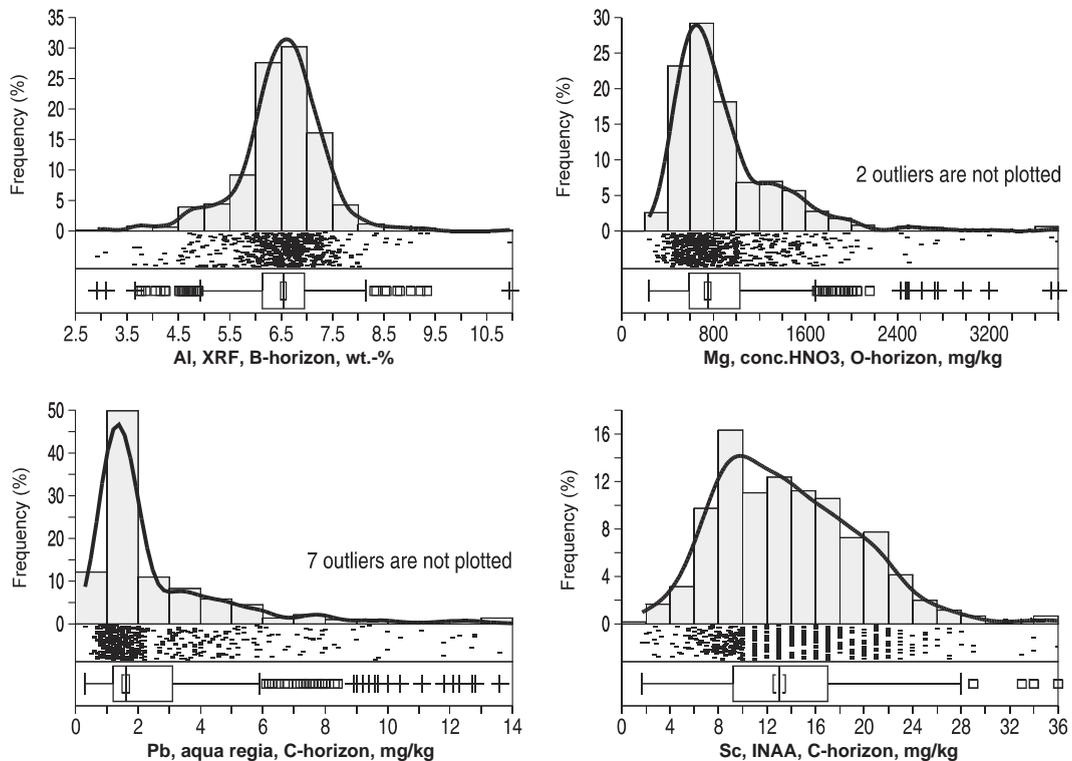


Fig. 5. Combination of histogram, density trace, one-dimensional scattergram and boxplot for the study of the empirical data distribution (data from Reimann et al., 1998).

structure than the histogram alone. The one-dimensional scattergram is a very simple tool where the measured data are displayed as a small horizontal line at an arbitrarily chosen y -scale position at the appropriate position along the x -scale. In contrast to histograms, a combination of density traces, scattergrams and boxplots will at once show any peculiarities in the data, e.g., breaks in the data structure (Fig. 5, Pb) or data discretisation due to severe rounding of analytical results in the laboratory (Fig. 5, Sc).

One of the best graphical displays of geochemical distributions is a cumulative probability plot (CDF diagram), originally introduced to geochemists by Tennant and White (1959), Sinclair (1974, 1976) and others. Fig. 6 shows four forms of such displays. The best choice for the y -axis is often the normal probability scale because it spreads the data out at the extremes, which is where interest usually lies. Also, it permits the direct detection of deviations from normality or lognormality, as normal or lognormally (logarithmic x -scale) distributed data plot as straight

lines. The use of a normal probability model can be criticized, particularly after our criticism of the use of normal models to estimate thresholds, i.e. [mean \pm 2 sdev]. However, in general, geochemical data distributions are symmetrical, or can be brought into symmetry by the use of a logarithmic transform, and the normal model provides a useful, widely known, framework for data display. Alternatively, if interest is in the central part of the data, the linear empirical cumulative distribution function is often more informative as it does not compress the central part of the data range as does a normal probability plot in order to provide greater detail at the extremes (tails). One of the main advantages of CDF diagrams is that each single data value remains visible. The range covered by the data is clearly visible, and extreme outliers are detectable as single values. It is possible to directly count the number of extreme outliers and observe their distance from the core (main mass) of the data. Some data quality issues can be detected, for example, the presence of discontinuous data values at the lower

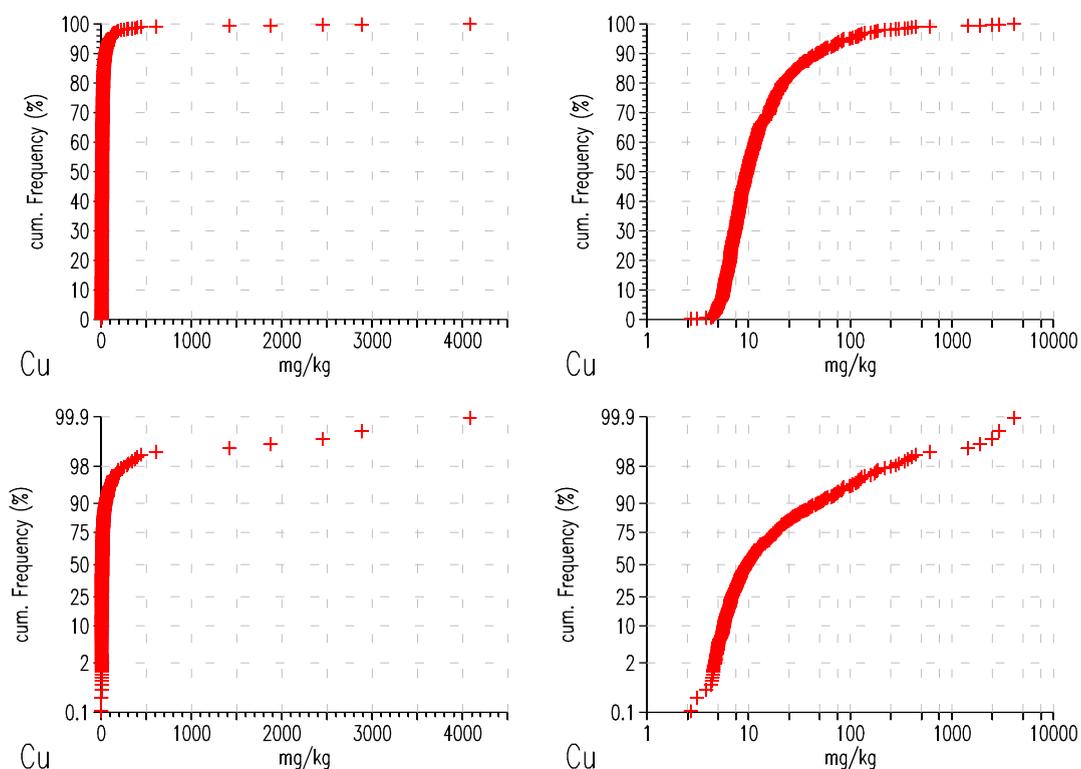


Fig. 6. Four different variations for plotting CDF diagrams. Upper row, empirical cumulative distribution plots; lower row, cumulative probability plots. Left half, data without transformation; lower right, data plotted on a logarithmic scale equivalent to a logarithmic transformation. Example data: Cu (mg/kg) in podzol O-horizons from the Kola area (Reimann et al., 1998).

end of the distribution. Such discontinuous data are often an indication of the method detection limit or too severe rounding, discretisation, of the measured values reported by the laboratory. Thus, values below the detection limit set to some fixed value are visible as a vertical line at the lower end of the plot and the percentage of values below the detection limit can be visually estimated. It should be noted that although reporting of values below detection limit as one value has been officially discouraged in favour of reporting all measurements with their individual uncertainty value (AMC, 2001; see also: http://www.rsc.org/lap/rsccom/amc/amc_techbriefs.htm), most laboratories are still unwilling to deliver values for those results that they consider as “below detection” to their customers. The presence of multiple populations results in slope changes and breaks in the plot.

Identifying the threshold in the cumulative probability plot is, however, still not a trivial task. Fig. 7 displays these plots for selected elements from Table

1, the arrows indicating inflection or break points that most likely reflect the presence of multiple populations and outliers. It is obvious that extreme outliers, if present, can be detected without any problem. In Fig. 6, for example, it can be clearly seen in all the versions of the diagram that the boundary dividing these extreme outliers from the rest of the population is 1000 mg/kg. Searching for breaks and inflection points is largely a graphical task undertaken by the investigator, and, as such, it is subjective and experience plays a major role. However, algorithms are available to partition linear (univariate) data, e.g., Garrett (1974) and Miesch (1981) who used linear clustering and gap test procedures, respectively. In cartography, a procedure known as ‘natural breaks’ that identifies gaps in ordered data to aid isopleth (contour interval) selection is available in some Geographical Information Systems (Slocum, 1999).

However, often the features that need to be investigated are subtle, and in the spirit of Explor-

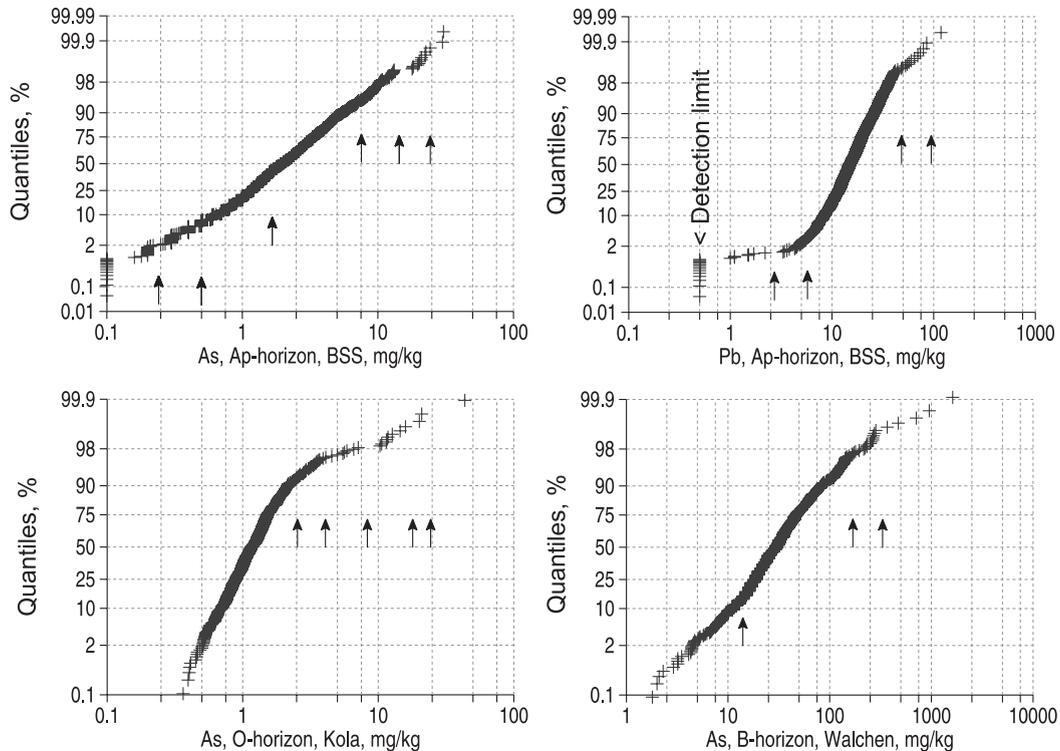


Fig. 7. Four selected cumulative probability plots for the example data from Table 1. Arrows mark some different possible thresholds (compare with Table 1). Example data are taken from (Reimann et al., 1998, 2003; Reimann, 1989). The vertical lines in the lower left corner of the plots for the Baltic Soil Survey (Reimann et al., 2003) data are caused by data below the detection limit (set to half the detection limit for representation).

atory Data Analysis (EDA) the trained eye is often the best tool. On closer inspection of Fig. 6, it is evident that a subtle inflection exists at 13 mg/kg Cu, the 66th percentile, best seen in the empirical cumulative distribution function (upper right). About 34% of all samples are identified as ‘outliers’ if we accept this value as the threshold. This may appear unreasonable at first glance. However, when the data are plotted as a geochemical map (Fig. 8 presents a Tukey boxplot based map for the data in Fig. 6), it becomes clear that values above 18 mg/kg Cu, the 3rd quartile, originate from a second process (contamination from the Kola smelters). The challenge of objectively extracting an outlier boundary or threshold from these plots is probably the reason that they are not more widely used. However, as can be seen in the above example, when they are coupled with maps they become a powerful tool for helping to understand the data.

Results in Table 1 demonstrate that in practice the [mean \pm 2 sdev] rule often delivers surprisingly comparable estimates to the graphical inspection of cumulative probability plots. The reasons for this are discussed later. In most cases, however, all methods still deliver quite different estimates—an unacceptable situation upon which to base environmental regulation.

The uppermost 2%, 2½% or 5% of the data (the uppermost extreme values) are, adapting recommendation (2) of Hawkes and Webb (1962), sometimes arbitrarily defined as ‘outliers’ for further inspection. This will result in the same percentage of outliers for all variables. This approach is not necessarily valid, because the real percentage of outliers could be very different. In a data distribution derived from natural processes there may be no outliers at all, or, in the case of multiple natural background processes, there may appear to be outliers in the context of the main

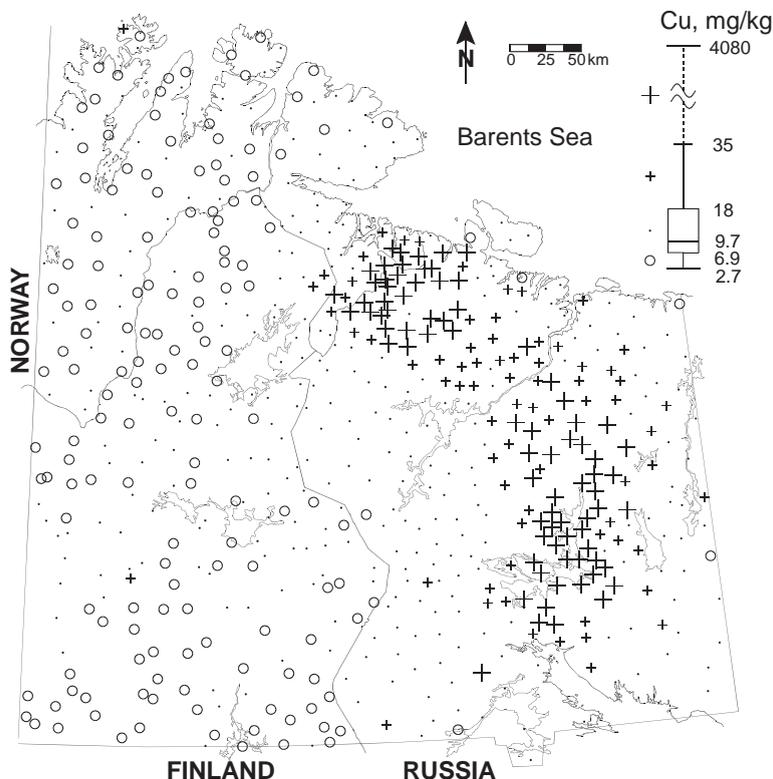


Fig. 8. Regional distribution of Cu (mg/kg) in podzol O-horizons from the Kola area (Reimann et al., 1998). The high values in Russia (large and small crosses) mark the location of the Cu-refinery in Monchegorsk, the Cu-smelter in Nickel and the Cu–Ni ore roasting plant in Zapoljarnij (close to Nickel). The map suggests that practically all sample sites in Russia, and some in Norway and Finland, are contaminated.

mass of the data. However, in practice the percentile approach delivers a number of samples for further inspection that can easily be handled. In some cases, environmental legislators have used the 98th percentile of background data as a more sensible inspection level (threshold) than values calculated by the $[\text{mean} \pm 2 \text{ sdev}]$ rule (e.g., Ontario Ministry of Environment and Energy, 1993).

The cumulative probability, and equivalent Q–Q or Q–Normal plots (quantiles of the data distribution are plotted against the quantiles of a hypothetical distribution like the normal distribution) present in many statistical data analysis packages, delivers a clear and detailed visualisation of the quality and distribution of the data. Deviations from normal or lognormal distributions, and the likely presence of multiple populations, can be easily observed, together with the presence of obvious data outliers. In such instances, which are not rare, these simple diagrams demonstrate

that the $[\text{mean} \pm 2 \text{ sdev}]$ rule is inappropriate for estimating the limits of background (thresholds).

Exploration geochemists have always been interested in the regional spatial distribution of data outliers. One of the main reasons for trying to define a threshold is to be able to map the regional distribution of such outliers because they may indicate the presence of ore deposits. Environmental geochemists may wish to define an inspection or action level, or clean-up goal, in a similar way to geochemical threshold. However, in other instances, regulatory levels for environmental purposes are set externally from geochemical data on the basis of ecotoxicological studies and their extrapolation from the laboratory to the field environment (Janssen et al., 2000). In some instances, this has resulted in regulated levels being set down into the natural background range. This has posed problems when mandatory clean-up is required, and arises from the fact that most

ecotoxicological studies are carried out with soluble salts, whereas geochemical data are often the result near-total strong acid extractions that remove far more of an element from a sample material than is actually bioavailable (see, for example, Chapman and Wang, 2000; Allen, 2002). In addition to the problems cited above, a further problem with defining ‘background’ or ‘threshold’ values is that, independent of the method used, different estimates will result when different regions are mapped or when the size of the area investigated is changed. Thus the main task should always be to first understand the process(es) causing high (or low) values. For this purpose, displaying the spatial data structure on a suitable map is essential.

When mapping the distribution of data outliers they are often displayed as one of the classes on a symbol map. However, the problems of selecting and graphically identifying background levels, thresholds and outliers have led many geochemists to avoid using classes in geochemical mapping. Rather, a continuously variable symbol size, e.g., a dot (black and white mapping), or a continuous colour scale may be used (e.g., Björklund and Gustavsson, 1987).

To display the data structure successfully in a regional map, it is very important to define symbolic or colour classes via a suitable procedure that transfers the data structure into a spatial context. Percentiles, or the boxplot—as proposed almost 20 years ago (Kürzl, 1988)—fulfil this requirement, not, however, arbitrarily chosen classes or a continuously growing symbol. The map displayed in Fig. 8 is based on boxplot classes.

An alternate procedure to identify spatial structure and scale in geochemical maps and to help identify thresholds has been described by Cheng et al. (1994, 1996) and Cheng (1999). In this procedure, the scale of spatial structures is investigated using fractal and multi-fractal models. The procedure is not discussed further here as tools for its application are not generally available; however, those with an interest in this approach are referred to the above citations.

When plotting such maps for large areas (see Reimann et al., 1998, 2003), it becomes obvious that the concepts of ‘background’ and ‘threshold’ are illusive. A number of quite different processes can cause high (or low) data values, not only mineralisation (exploration geochemistry) or contamination (environmental geochemistry). It is thus important to

prepare and inspect these regional distribution maps rather than just defining ‘levels’ via statistical exercises, possibly based on assumptions not suited for geochemical data.

In terms of environmental geochemistry, it would be most interesting to compare such geochemical distributions directly with indicator maps of ecosystem or human health, or uptake into the human system (e.g., the same element measured in blood, hair or urine—Tristán et al., 2000).

Only the combination of statistical graphics with appropriate maps will show where action may be needed—independent of the concentration. The variation in regional distribution is probably the most important message geochemical survey data sets contain, more important than the ability to define ‘background’, ‘threshold’ or ‘inspection levels’. Such properties will be automatically visible in well-constructed maps.

4. Conclusions

Of the three investigated procedures, the boxplot function is most informative if the true number of outliers is below 10%. In practice, the use of the boxplot for preliminary class selection to display spatial data structure in a map has proven to be a powerful tool for identifying the key geochemical processes behind a data distribution. If the proportion of outliers is above 15%, only the [median \pm 2 MAD] procedure will perform adequately, and then up to the point where the outlier population starts to dominate the data set (50%). The continued use of the [mean \pm 2 sdev] rule is based on a misunderstanding. Geochemists want to identify data outliers and not the extreme values of normal (or lognormal) distributions that statisticians are often interested in. Geochemical outliers are not these extreme values for background populations but values that originate from different, often superimposed, distributions associated with processes that are rare in the environment. They can, and often will, be the ‘extreme values’ for the whole data set. This is the reason that the [mean \pm 2 sdev] rule appears to function adequately in some real instances, but breaks down when the proportion of outliers in the data set is large relative to the background population size. The derived values, however, have no statistical

validity. As a consequence, the boxplot inner fences and $[\text{median} \pm 2 \text{ MAD}]$ are all better suited for assisting in the estimation of the background range than $[\text{mean} \pm 2 \text{ sdev}]$. Where percentiles are employed, the use of the 98th has become widespread as a 2%, 1 in 50, rate is deemed acceptable, and it distances the method from the 97.5 percentile, 2½%, 1 in 40 rate associated with the $[\text{mean} \pm 2 \text{ sdev}]$ rule. However, all these procedures usually lead to different estimates of background range. The use of the $[\text{mean} \pm 2 \text{ sdev}]$ rule should be finally discontinued—it was originally suggested to provide a ‘filter’ that would identify approximately 2½% of the data for further inspection at a time when computers to do the drudgery of numerical operations were not widely available and no other practical methods existed. It is definitely not suited to calculate any action levels or clean-up goals in environmental legislation.

The graphical inspection of the empirical data distribution in a cumulative probability (or Q–Q) plot prior to defining the range of background levels or thresholds is an absolute necessity. The extraction of thresholds from such plots, however, involves an element of subjectivity and is based on a priori geochemical knowledge and practical experience. Having extracted these potential population limits and thresholds their regional/spatial data structure must be investigated and their relationship to known natural and anthropogenic processes determined. Only cumulative probability (or Q–Q) plots in combination with spatial representation will provide a clearer answer to the background and threshold question. Given today’s powerful PCs, the construction of these diagrams and maps poses no problems, whatever the size of the data set, and their use should be encouraged.

Appendix A

The following is one heuristic for data inspection and selection of the limits of background variation that has proved to be informative in past studies. The following assumes that the data are in a computer processable form and have been checked for any obvious errors, e.g., in the analytical and locational data; and that the investigator has access to data analysis software. Maps suitable for data inspection can be displayed with data analysis software; however,

availability of a Geographic Information System (GIS) permits the use of such procedures as ‘natural’ breaks, and the preparation of more elaborate displays.

1. Display Empirical Cumulative Distribution Functions (ECDFs) on linear and normal probability scales, and Tukey boxplots. Inspect these for evidence of multiple populations (polymodality), and extreme or outlying values. If there are a few extremely high or low values widely separated from the main mass of the data prepare a subset with these values omitted. Seek an explanation for the presence of these anomalous individuals;
2. Compute the mean and standard deviation (sdev) of the data (sub)set, and then the coefficient of variation (CV%), i.e. $100 \times \text{sdev}/\text{mean}\%$. The CV is a useful guide to non-normality (Koch and Link, 1971). Alternately, the data skewness can be estimated;
3. If the $\text{CV} > 100\%$ plots on a logarithmic scale should be prepared. If the CV is between 70% and 100%, the inspection of logarithmically scaled plots will likely be informative. Another useful guide to deciding on the advisability of inspecting logarithmically scale plots is the ratio of maximum to minimum value, i.e. max/min . If the ratio exceeds 2 orders of magnitude, logarithmic plots will be informative, if the ratio is between 1.5 and 2 orders of magnitude, logarithmically scaled plots will likely be informative;
4. Calculate trial upper ‘fences’, $[\text{median} + 2 \text{ MAD}]$ and Tukey inner fence, see Eqs. (1) and (2). If Step 3 indicates that logarithmic displays would be informative, log-transform the data, repeat the calculations, and anti-log the results to return them to natural numbers;
5. Prepare maps using these fence values, and such ‘natural’ properties of the data such as the medians and quartiles (50th, 25th and 75th percentiles), and if the data are polymodal, boundaries (symbol changes or isopleths) may be set at these natural breaks. Inspect the maps in the light of known natural and anthropogenic processes, e.g., different geological and soil units, and the presence of industrial sites. See if concentration ranges can be linked to different processes;
6. It may be informative to experiment with tentative fence values on the basis of the ECDF plots from

Step 1. Eventually, software will become more widely available to prepare concentration–area plots (Cheng, 1999; Cheng et al., 1994, 1996). These can be informative and assist in selecting data based fences;

7. On the basis of the investigation, propose an upper limit of background variation; and
8. If Step 1 indicates the data are polymodal, a considerable amount of expert judgment and knowledge may be necessary in arriving at a useful upper limit of background variation and understanding what processes are giving rise to the observed data distribution. If the survey area is large or naturally geochemically complex, there may be multiple discrete background populations, in such instances the presentation of the data as maps, i.e. in a spatial context, is essential.

The above heuristic is applicable where the majority of the data reflect natural processes. In some environmental studies, the sampling may focus on a contaminated site, and as a result there is little data representing natural processes. In such cases, the sample sites representing natural background will be at the lower end of the data distribution, and focus of attention and interpretation will be there. In the context of these kinds of investigation, this stresses the importance of extending a survey or study sufficiently far away from the contaminating site to be able to establish the range of natural background. In this respect, the area into which the background survey is extended must be biogeochemically similar to the contaminated site (pre-industrialization) for the background data to be valid. Thus such factors as similarity of soil parent material (geology), soil forming processes, climate and vegetation cover are critical criteria.

References

- Allen HE, editor. Bioavailability of metals in terrestrial ecosystems: importance of partitioning for bioavailability to invertebrates, microbes, and plants. Pensacola, FL: Society of Environmental Toxicology and Chemistry (SETAC); 2002.
- AMC. Analyst 2001;126:256–9.
- Barnett V, Lewis T. Outliers in statistical data. 3rd edition. New York: Wiley & Sons; 1994.
- Björklund A, Gustavsson N. Visualization of geochemical data on maps: new options. J Geochem Explor 1987;29:89–103.
- Chapman P, Wang F. Issues in ecological risk assessments of inorganic metals and metalloids. Hum Ecol Risk Assess 2000; 6(6):965–88.
- Cheng Q. Spatial and scaling modelling for geochemical anomaly separation. J Geochem Explor 1999;65:175–94.
- Cheng Q, Agterberg FP, Ballantyne SB. The separation of geochemical anomalies from background by fractal methods. J Geochem Explor 1994;51:109–30.
- Cheng Q, Agterberg FP, Bonham-Carter GF. A spatial analysis method for geochemical anomaly separation. J Geochem Explor 1996;56:183–95.
- Dutter R, Filzmoser P, Gather U, Rousseeuw P, editors. Developments in robust statistics. International Conference on Robust Statistics 2001. Heidelberg: Physik-Verlag; 2003.
- Garrett RG. Copper and zinc in Proterozoic acid volcanics as a guide to exploration in the Bear Province. In: Elliott IL, Fletcher WK, editors. Geochemical exploration 1974. Developments in Economic Geology, vol. 1. New York: Elsevier Scientific Publishing; 1974.
- Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel W. Robust statistics. The approach based on influence functions. New York: John Wiley & Sons; 1986.
- Hawkes HE, Webb JS. Geochemistry in mineral exploration. New York: Harper; 1962.
- Hoaglin D, Mosteller F, Tukey J. Understanding robust and exploratory data analysis. 2nd edition. New York: Wiley & Sons; 2000.
- Huber PJ. Robust statistics. New York: Wiley & Sons; 1981.
- Janssen CR, De Schampelaere K, Heijerick D, Muysen B, Lock K, Bossuyt B, et al. Uncertainties in the environmental risk assessment for metals. Hum Ecol Risk Assess 2000;6(6):1003–18.
- Koch GS, Link RF. Statistical analysis of geological data, vol 11. New York: Wiley & Sons; 1971.
- Kürzl H. Exploratory data analysis: recent advances for the interpretation of geochemical data. J Geochem Explor 1988; 30:309–22.
- Miesch AT. Estimation of the geochemical threshold and its statistical significance. J Geochem Explor 1981;16:49–76.
- Ontario Ministry of Environment and Energy AT. Ontario typical range of chemical parameters in soil, vegetation, moss bags and snow. Toronto: Ontario Ministry of Environment and Energy; 1993.
- Reimann C. Untersuchungen zur regionalen Schwermetallbelastung in einem Waldgebiet der Steiermark. Graz: Forschungsgesellschaft Joanneum (Hrsg.): Umweltwissenschaftliche Fachtage—Informationsverarbeitung für den Umweltschutz; 1989.
- Reimann C, Filzmoser P. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environ Geol 2000;39/9:1001–14.
- Reimann C, Äyräs M, Chekushin V, Bogatyrev I, Boyd R, de Caritat P, et al. Environmental geochemical atlas of the central barents region. NGU-GTK-CKE Special Publication. Trondheim, Norway: Geological Survey of Norway; 1998.
- Reimann C, Siewers U, Tarvainen T, Bitjukova L, Eriksson J, Gilucis A, et al. Agricultural soils in northern Europe: a geochemical Atlas Geologisches Jahrbuch, Sonderhefte, Reihe

- D Heft SD 5. Stuttgart: Schweizerbart'sche Verlagsbuchhandlung; 2003.
- Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. New York: Wiley & Sons; 1987.
- Sinclair AJ. Selection of threshold values in geochemical data using probability graphs. *J Geochem Explor* 1974;3:129–49.
- Sinclair AJ. Applications of probability graphs in mineral exploration. Spec Vol 4. Assoc. Explor. Geochem 1976 [Toronto].
- Slocum TA. Thematic cartography and visualization. Upper Saddle River, NJ: Prentice Hall; 1999.
- Tennant CB, White ML. Study of the distribution of some geochemical data. *Econ Geol* 54:1959;1281–90.
- Tristán E, Demetriades A, Ramsey MH, Rosenbaum MS, Stravakkis P, Thornton I, et al. Spatially resolved hazard and exposure assessments: an example of lead in soil at Lavrion, Greece. *Environ Res* 2000;82:33–45.
- Tukey JW. Exploratory data analysis. Reading: Addison-Wesley; 1977.