Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data

C. Reimann · P. Filzmoser

Abstract All variables of several large data sets from regional geochemical and environmental surveys were tested for a normal or lognormal data distribution. As a general rule, almost all variables (up to more than 50 analysed chemical elements per data set) show neither a normal or a lognormal data distribution. Even when different transformation methods are used more than 70 % of all variables in every single data set do not approach a normal distribution. Distributions are usually skewed, have outliers and originate from more than one process. When dealing with regional geochemical or environmental data normal and/or lognormal distributions are an exception and not the rule. This observation has serious consequences for the further statistical treatment of geochemical and environmental data. The most widely used statistical methods are all based on the assumption that the studied data show a normal or lognormal distribution. Neglecting that geochemcial and environmental data show neither a normal or lognormal distribution will lead to biased or faulty results when such techniques are used.

Key words Normal distribution · Lognormal distribution · Geochemistry · Exploratory data analysis · Multivariate normal distribution · Robust methods · Non-parametric methods · Median

Received: 21 June 1999 · Accepted: 14 August 1999

C. Reimann (✉)
Geological Survey of Norway, N-7491 Trondheim, Norway
e-mail: Clemens.Reimann@ngu.no

P. Filzmoser
Department of Statistics, Probability Theory and Actuarial Mathematics, Vienna University of Technology,
Wiedner Hauptstr. 8–10, A-1040 Vienna, Austria
e-mail: P. Filzmoser@tuwien.ac.at

Introduction

The first step in data analysis should always be to carefully study the distribution of the measured variables graphically. A combination of different graphics, e.g. the histogram, a density trace, the boxplot and a one-dimensional scattergram, possibly combined with a CDF-diagram (CDF = cumulative distribution function), will give an excellent one-dimensional insight into the data structure (Fig. 1). Why is the data distribution so important that it must be known before doing anything else? Correlation analysis, factor analysis, discriminant analysis and many classical statistical tests, including most calculations of probability levels are based on the assumption of a normal data distribution. In geochemistry and environmental sciences this basic requirement is still widely neglected although a number of papers and even books address the problem (e.g. Philip and Watson 1987; Rock and others 1987; Rock 1988). Ahrens (1953, 1954a, 1954b, 1957) proposed the lognormal type of distribution for geochemical data. Although his ideas encountered immediate criticism (e.g. Aubrey 1954, 1956; Chayes 1954; Miller and Goldberg 1955; Vistelius 1960) the damage was done. Most modern textbooks in geochemistry still assert that geochemical data commonly approach a lognormal distribution. A log-transformation (log10 or ln) is thus most frequently used for data transformation when working with geochemical data. But do the so transformed data really approach a lognormal distribution? This is almost never tested. If it is tested, it is usually found that the data do not follow a lognormal distribution (e.g. McGrath and Loveland 1992). Neglecting this fact in further data analysis has serious consequences. What is so special about geochemical/environmental data? Real world data are rarely as well-behaved as classical statistical tests assume. Geochemical and environmental data show first of all a spatial dependence. Spatially dependent data are not, in general, normally distributed. Furthermore these data are based on rather imprecise measurements. There are many potential sources of error involved in sampling, sample preparation and analysis.

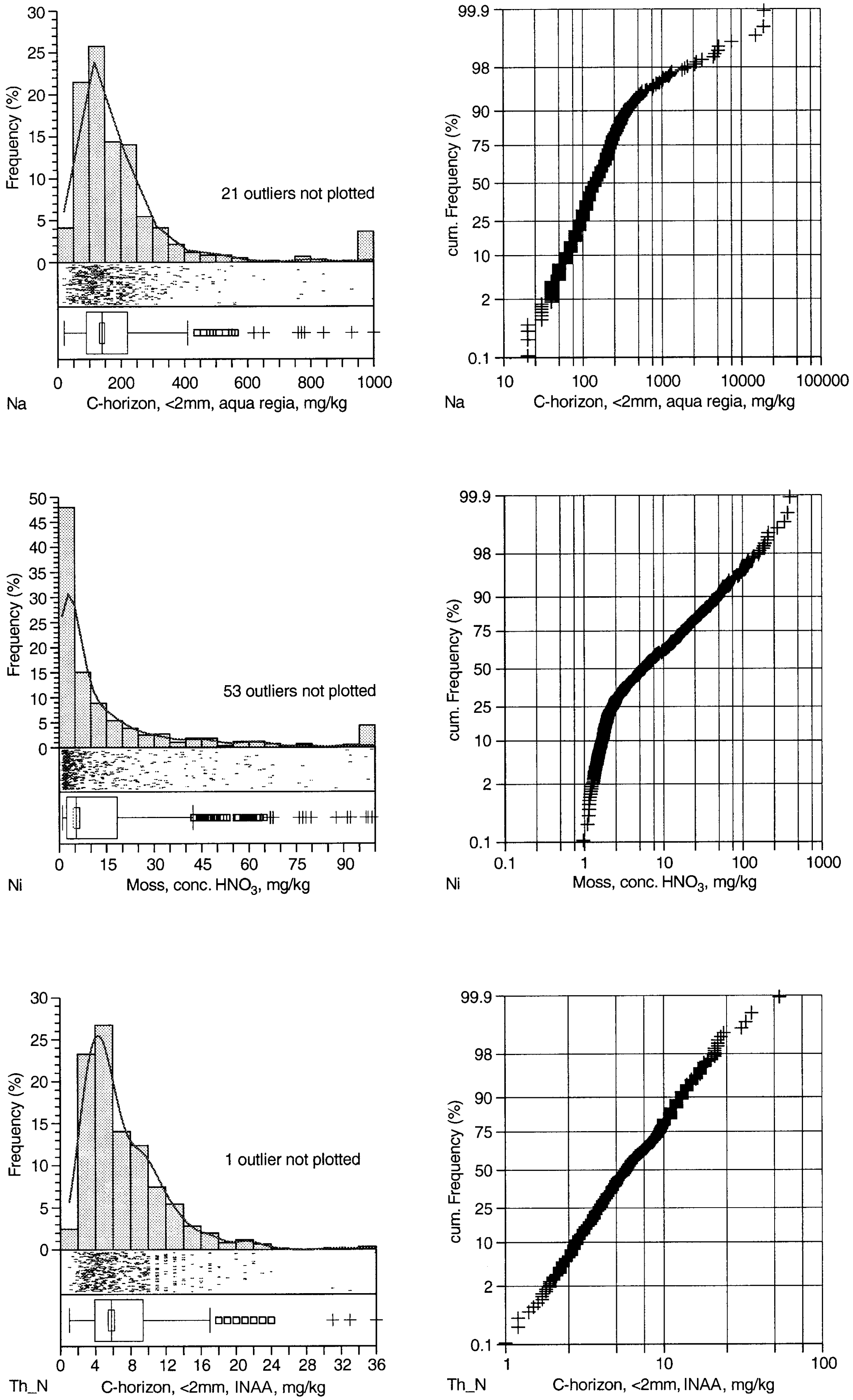

Fig. 1
Combination of histogram, density trace, one-dimensional scattergram and boxplot and a CDF-diagram to give a fast graphical impression of the data distribution

Trace element analyses are often plagued by detection limit problems, i.e. a substantial number of samples are not characterised by a true measured value. In addition, the precision of the measurements changes with element concentration, values are less precise at very low and very high concentrations. The existence of data outliers – in most cases, the existence of some samples with unusually high concentrations – is a very common characteristic of such data sets. They are thus strongly skewed. Even worse, these outliers often originate from another population than the main body of data. Any person trained in statistics will be able to recognise at once that the widely-used, classical statistical methods are likely to fail under such conditions. Different methods would have to be used, e.g. exploratory data analysis (e.g. Tukey 1977; Velleman and Hoaglin 1981), robust methods (Huber 1981; Hampel and others 1986; Rieder 1994), or non-parametric methods (e.g. Noether 1991; Gore and others 1993). Robust and non-parametric methods usually need considerable computing power. Even today they are still not implemented in many standard statistical software packages. They are rarely taught in university courses for earth and environmental scientists. On the other hand graphical, exploratory data analysis is sometimes even defamed as "simple". This may be a reason why this powerful tool is rarely used.

Here a number of large regional geochemical and environmental data sets will be used to demonstrate that geochemical and environmental data, as a rule, show neither a normal nor a lognormal distribution. Statistical techniques that should as a consequence be used when studying geochemical and environmental data are suggested.

Materials and methods

The test data sets

From 1992–1998 the Geological Surveys of Finland (GTK) and Norway (NGU) and Central Kola Expedition (CKE), Russia, carried out a large, international multi-media, multi-element geochemical mapping project, covering 188000 km^2 north of the Arctic Circle. The entire area between 24° and 35.5°E up to the Barents Sea coast was sampled during the summer of 1995. Results of the "Kola Ecogeochemistry" project are documented on a web site (http://www.ngu.no/Kola) and in a geochemical atlas (Reimann and others 1998). The average sample density was one site per 300 km^2. Samples of terrestrial moss, humus (the O-horizon), topsoil (0–5 cm), and the B- and C-horizon of Podzol profiles were taken at more than 600 sites and subsequently analysed for up to more than 50 elements. In many sample materials, elements were analysed with more than one technique or following more than one extraction method (total, i.e. X-ray fluorescence (XRF) and instrumental neutron activation analyses (INAA) vs. partial e.g. aqua regia or ammonium acetate extraction). Details on sampling, sample preparation, analyses and quality control are given in Reimann and others (1998). This represents one of the largest single datasets in geochemistry and environmental sciences in terms of area covered, number of different sample materials and number of elements analysed. It is thus ideally suited for this test.

Some might argue, that the sample density of the Kola data set is unusually low and that this may explain why the data show no normal or lognormal distribution. Thus data from a high density (8 samples per km^2) soil survey of a much smaller area (ca. 100 km^2) in Austria (the "Walchen" dataset) are used here as well. This sample set consists of 772 B-horizon soil samples (forest soils), originally collected for mineral exploration purposes. The samples were sieved to <0.18 mm and analysed for more than 30 elements, mostly using techniques giving total element concentrations (XRF and INAA). The data are of unusually high quality – results of quality control for this dataset are documented in Reimann (1989a).

Some might argue that the sample size in the Kola and Walchen data sets is not sufficient to truly approach a normal distribution (in both cases is $N < 1000$). Thus we have used the data of the stream sediment survey of Austria (Thalmann and others 1989), a survey covering more than 40000 km^2 at an average density of one site per 1.4 km^2. Here 29717 samples have been analysed for a total of 35 elements, giving one of the biggest consistent single data sets that exists in regional geochemistry.

Finally, as some might argue that the areas covered are not big enough to give truly spatially independent analyses, data from a project reporting element concentrations in agricultural soils taken over the whole of northern Europe (the Baltic Soil Survey – BSS; >1500000 km^2) are used. Here large composite samples were taken from the ploughing layer (Ap-horizon, 0–25 cm) and a lower, depth defined layer (50–75 cm). The average sample density is one site per 2500 km^2. The samples were air-dried, sieved to <2 mm and analysed by a variety of methods, giving total and partial element concentrations (only XRF-results used here). This data set has not been officially reported yet.

Treatment of the data sets

As mentioned above, geochemical and environmental data sets are often characterised by a high proportion of samples returning values below detection levels for some of the analysed elements. Such data are very difficult to treat. If there is a high number of values below detection (e.g. $<25\%$) there is no chance that these data will approach a normal or lognormal distribution. Such variables were thus not included in this test. In all other cases, where only a low number of samples returned values below detection, these were set to one half of the detection limit to allow the use of these samples for further statistical analyses.

Test for normality

A large number of tests for normal distribution exist. The easiest method is just to plot a histogram of the distribution and check it for the typical bell shape. This method is often used in geochemical textbooks to "prove" that geochemical data approach a lognormal distribution, probably, because it will quite often show the "wanted" result. Another widely used graphical technique for examining the shape of the distribution of univariate data is the quantile-quantile Q-Q plot (Hazen 1914). These graphical tools give a good first impression of the data distribution. To demonstrate that the data deviate significantly from normal distribution, however, a more formal statistical test should be applied. There are different possibilities for testing for univariate normality. The most popular tests are the Kolmogorov-Smirnov test (Smirnov 1948; Afifi and Azen 1979), the chi-square goodness-of-fit test (Conover 1980), and the Shapiro-Wilk test (Shapiro and Wilk 1965). All these tests compare an independent identically-distributed sample from an unknown univariate distribution with a reference sample with a known distribution (in our case the normal distribution). The tests result in a p-value that can be taken as a decision as to whether the null hypothesis can be rejected. Usually, if $p<0.05$ the null hypothesis of normal distribution is rejected. In general the Shapiro-Wilk test is statistically preferable to the other two tests.

Test for multivariate normal distribution

Some multivariate methods and tests may not only require that each variable entered follows a normal distribution but also that the data set displays in addition a multivariate normal distribution. Everybody knows the bell shape of the one-dimensional normal distribution. The multivariate normal distribution can be envisaged as a "real" 3D (or more dimensions) bell, where any projection as a cut through the z-axis must again result in a one-dimensional normal distribution. This can be tested graphically by a multivariate generalisation of the Q-Q plot (Easton and McCulloch 1990):

Suppose that $x_1,...,x_n$ is a sample from p-dimensional space. Denote by $\bar{x}$ the p-dimensional mean vector and by S the sample variance-covariance matrix. Then

$$r_i^2 = (x_i - \bar{x})^T S^{-1} (x_i - \bar{x})$$

defines the (squared) Mahalanobis distance for each observation $i = 1,...,n$. Since $\bar{x}$ and S are centre and shape of the data the Mahalanobis distance reflects for each data point the "closeness" to the centre. If the data are from a p-dimensional multivariate normal distribution, then the random variables r_i^2 for $i = 1,...,n$ are approximately χ_p^2-distributed. Thus, a plot of the order statistics of the r_i^2's against the expected values of the order statistics of the χ_p^2 distribution is a multivariate extension of the univariate Q-Q plot. If the data indeed follow a multivariate normal distribution, the data points should be arranged along the 45° line in the multivariate Q-Q plot.

Results

As an example Tables 1 and 2 summarise the results for two of the eight data sets tested. The C-horizon soil analyses from the Kola project represent typical regional geochemical data, while the results from terrestrial moss are a typical data set for environmental geochemistry. The tables show that there is a very large deviation between mean and median and standard deviation and mad (median absolute deviation – a measure of dispersion, highly robust to skew and outliers) for practically all variables – a first indication that the data do not exhibit a normal distribution. According to the statistical tests applied, none of the original variables shows a normal distribution ($p < 0.05$ for all variables). Different transformation methods were tested to approach a normal distribution: ln, log, square-root, range and logit. The tables demonstrate that as a general rule, these transformations do not result in normal distributions. With regards to the three different tests for a normal distribution, while differences for single variables were observed, the general result, however, is the same. For comparison, results of the Shapiro-Wilk test, the Kolmogorov-Smirnov test and the chi-square test are all three shown in Tables 1 and 2 for the log-transformed (ln) data.

For the C-horizon, 20 out of 57 variables show p-values > 0.05 for the Shapiro-Wilk test after log-transformation (ln; Table 2). This is actually the highest proportion of lognormally distributed variables for all data sets. A total of 17 additional available variables were not even tested for normal distribution because more than 25 % of all the data were below the detection levels. In these cases normal distribution cannot be approached. For the moss samples the situation is even worse, after transformation only 5 out of 31 elements approximate a normal distribution (Table 1). The other six data sets display a similar behaviour. There is not one data set in which a normal distribution can be approached for more than 30 % of all reported variables. The fact that geochemical/environmental data as a rule obviously do not approach normal or lognormal distribution has serious consequences for the further statistical treatment of geochemical/environmental data, consequences that are all too often neglected by the majority of scientists working in these fields. For example, given these conditions the median will probably represent a better estimate of location than the mean, although it could be argued that the mean gives a better estimate of the location even for skewed populations if the outliers truly belong to this population. Do the high values in geochemical and environmental data sets, however, belong to the same population? In most cases probably not. In geochemical data sets they may be indicative of unusual rock types occurring in an area or even of an ore deposit. In environmental data sets they will most likely be an indication of a pollution source. The distributions displayed as examples in Fig. 1 show a very common characteristic of geochemical data. In many cases the regional distribution of elements is influenced by more than just one process/source, resulting in multi-modal, skewed distributions.

Table 1
Moss, Kola data (Reimann and others 1998) – elements analysed, analytical technique used (ICP-MS inductively coupled plasma mass spectrometer, ICP-AES inductively coupled plasma atomic emission spectrometer, CV-AAS cold vapour atomic absorption spectrometer), detection limit (DL), samples below detection in %, minimum, maximum, mean, median, standard deviation, mad (medmed) (all data in mg/kg) and p-values for a Shapiro-Wilk test for normal distribution of the untransformed (orig) data, and transformed (ln, log, square root, range and logit) data. For the log-transformed (ln) data p-values for three different tests are given: _S-W: Shapiro-Wilk test, _K-S: Kolmogorov-Smirnov test, and _chi^2: chi-square test. Not included: Be (0.03 mg/kg, 89.3 % < DL), La (0.7 mg/kg, 85.5 %), Sc (0.1 mg/kg, 90.8 %), Se (0.8 mg/kg, 99.7 %), and Y (0.1 mg/kg, 74.4 %)

Element	Technique	DL	% <DL	Min	Max	Mean	Median	Sdev	Mad	Orig	ln_ S-W	ln_ K-S	ln_ chi^2	Log	Sqrt	Range	Logit
Ag	ICP-MS	0,01	1,5	<0.01	0,824	0,05	0,033	0,061	0,019	0	0	0	0	0	0	0	0
Al	ICP-MS	0,2	0	33,9	4850	300	193	458	90	0	0	0	0	0	0	0	0
As	ICP-MS	0,02	0	0,037	3,42	0,26	0,173	0,301	0,085	0	0	0	0	0	0	0	0
B	ICP-MS	0,5	3,2	<0.5	21,6	2,17	1,76	1,737	1,097	0	0	0,05	0,302	0	0	0	0
Ba	ICP-MS	0,05	0	6,71	175	21,40	19	12,046	6,227	0	0	0	0,002	0	0	0	0
Bi	ICP-MS	0,004	4,3	<0.004	0,544	0,027	0,018	0,041	0,012	0	0	0	0	0	0	0	0
Ca	ICP-AES	20	0	1680	9320	2740	2620	681	415	0	0	0	0	0	0	0	0
Cd	ICP-MS	0,01	0	0,023	1,23	0,12	0,09	0,111	0,036	0	0	0	0	0	0	0	0
Co	ICP-MS	0,03	0	0,11	13,2	0,92	0,40	1,478	0,304	0	0	0	0	0	0	0	0
Cr	ICP-MS	0,2	1,8	<0.2	14,4	0,9	0,6	1,13	0,33	0	0	0	0	0	0	0	0
Cu	ICP-MS	0,01	0	2,63	355	17,0	7,2	28,41	4,65	0	0	0	0	0	0	0	0
Fe	ICP-AES	10	0	46,5	5140	386	212	545	128	0	0	0	0	0	0	0	0
Hg	CV-AAS	0,01	0	0,023	0,155	0,06	0,05	0,02	0,016	0	0	0	0,023	0	0	0	0
K	ICP-AES	200	0	2260	8590	4360	4220	895	756	0	0,217	0	0,007	0,214	0	0	0,207
Mg	ICP-AES	10	0	518	2380	1132	1090	282	260	0	0,166	0,5	0,004	0,166	0	0	0,164
Mn	ICP-AES	1	0	28,5	1170	444	433	204,3	213,5	0	0	0	0	0	0,004	0	0
Mo	ICP-MS	0,01	0	0,016	1,08	0,11	0,08	0,096	0,037	0	0	0	0	0	0	0	0
Na	ICP-AES	10	0,7	<10	918	107	72	98,4	53,9	0	0,028	0,02	0,003	0,028	0	0	0,028
Ni	ICP-MS	0,3	0	0,96	396	19,5	5,4	40,76	5,43	0	0	0	0	0	0	0	0
P	ICP-AES	15	0	511	3800	1260	1265	287	245	0	0	0	0	0	0,012	0	0,001
Pb	ICP-MS	0,04	0	0,84	29,4	3,34	2,98	2,058	1,127	0	0,001	0,05	0,857	0,001	0	0	0,001
Rb	ICP-MS	0,5	0	1,39	33,5	11,9	11,5	5,61	5,58	0	0	0	0	0	0,101	0	0
S	ICP-AES	15	0	543	2090	888	863	154	119	0	0	0	0	0	0	0	0
Sb	ICP-MS	0,01	0	0,011	0,623	0,052	0,041	0,045	0,018	0	0	0	0	0	0	0	0
Si	ICP-AES	20	0	24,9	983	213	197	107,9	86,0	0	0,42	0,01	0,061	0,42	0	0	0,421
Sr	ICP-MS	0,2	0	2,47	435	15,6	9,4	29,38	5,32	0	0	0	0	0	0	0	0
Th	ICP-MS	0,004	1,3	<0.004	1,14	0,038	0,023	0,07	0,014	0	0	0	0	0	0	0	0
Tl	ICP-MS	0,004	0,2	<0.004	0,35	0,032	0,023	0,032	0,016	0	0,187	0	0,059	0,187	0	0	0,189
U	ICP-MS	0,004	5,4	<0.004	0,451	0,02	0,011	0,037	0,007	0	0	0	0	0	0	0	0
V	ICP-MS	0,02	0	0,28	83,8	2,58	1,60	4,575	0,912	0	0	0	0	0	0	0	0
Zn	ICP-AES	1	0	11,7	81,9	33,7	32,2	10,79	9,42	0	0,587	0,5	0,584	0,586	0	0	0,587

For example, the distribution of Na in the C-horizon of the Kola data set shows a clear break in the distribution due to the occurrence of alkaline rocks in the survey area (Fig. 1 – CDF-diagram, for maps see Reimann and others 1998). The distribution of Ni in moss is strongly influenced by the emissions of the Russian nickel industry in the survey area (see Fig. 1 – CDF-diagram). In both cases two distributions (at least) are superimposed on one another and the use of the mean will clearly give a far too high estimate of location for the underlying main body of data, although it represents the average element concentration in the survey area. The general decision that has to be taken before mean or median are used is thus whether or not secondary processes should be allowed to have a major influence on the estimate of location. In most cases it will be better to first ignore the secondary process, because in a later step of data analysis these will then be much easier to detect. This is probably the main task in regional and environmental geochemistry. In this case the median, as a robust estimator of location, is far superior to the mean.

The standard deviation is based on the squared differences of each observation from the mean. Since the mean is already a bad estimator of location the standard deviation will give a unrealistic estimate of data spread. It is very strongly influenced by high values from a second population or by a few high data outliers. The median absolute deviation (mad) is robust against a high number of outliers. For the use of median and mad the data do not need to follow any model. In most cases they will thus give much more realistic values for location and spread. Fig. 2 shows the relative deviation of mean and standard deviation from median and mad for all eight investigated data sets. Even a cursory glance will show the big differences between the classical and widely used estimators and the better, much less used median and mad. A large difference between mean and median and standard deviation and mad is again a clear indication that the data do not show a normal distribution.

It may be argued that there are two easy solutions to the above problem. One could calculate the mean and standard deviation for the log-transformed values to then

Table 2

C-horizon, Kola data (Reimann and others 1998) – elements analysed, analytical technique (see Table 1 for method abbreviations+XRF X-ray fluorescence, INAA instrumental neutron activation analysis, GF-AAS graphite furnace atomic absorption spectrometer), detection limit, samples below detection in %, minimum, maximum, mean, median, standard deviation, mad (medmed) (all data in mg/kg) and p-values for a Shapiro-Wilk test for normal distribution of the untransformed (orig) data and transformed (ln, log, square root, range and logit) data. For the log-transformed (ln) data p-values for three different tests are given: _S-W: Shapiro-Wilk test, _K-S: Kolmogorov-Smirnov test, and _chi^2: chi-square test. Not included: As_INAA (0.5 mg/kg, 71.9 % < DL), Au_INAA (0.002 mg/kg, 72.4 %), B (3 mg/kg, 89.4 %), Cs_INAA (1 mg/kg, 61.4 %), Hg (0.02, 56 %), Ir (0.005 mg/kg, 100 %), Mo (0.2 mg/kg, 76.5 %), Mo_INAA (1 mg/kg, 84 %), Ni_INAA (20 mg/kg, 86.3 %), Sb_INAA (0.1 mg/kg, 75.1 %), Se_INAA (3 mg/kg, 99 %), Sr_INAA (500 mg/kg, 86.5 %), Ta_INAA (0.5 mg/kg, 78.4 %), Tb_INAA (0.5 mg/kg, 72.4 %), U_INAA (0.5 mg/kg, 48.8 %), W_INAA (1 mg/kg, 97.9 %), Zn_INAA (50 mg/kg, 39.1 %)

Element	Technique	Unit	DL	%<DL	Min	Max	Mean	Median	Sdev	Mad	Orig	ln_S-W	ln_K-S	ln_chi^2	Log	Sqrt	Range	Logit
Ag	GF-AAS	mg/kg	0,001	0,2	0	0,119	0,011	0,008	0,011	0,004	0	0,058	0	0,009	0,058	0	0	0,058
Al	ICP-AES	mg/kg	10	0	1840	85900	12665	9910	9814	5834	0	0,043	0	0,044	0,043	0	0	0,026
Al_XRF	XRF	wt.- %	300	0	2,92	12,08	7,34	7,38	0,969	0,667	0	0	0	0	0	0	0	0
As	GF-AAS	mg/kg	0,1	1,7	<0.1	30,7	1,25	0,5	2,349	0,445	0	0	0	0	0	0	0	0
Ba	ICP-AES	mg/kg	0,5	0	4,7	1300	60,148	43,5	74,33	28,91	0	0,746	0,06	0,164	0,745	0	0	0,746
Ba_INAA	INAA	mg/kg	50	0	210	3000	600	575	223,8	170,5	0	0,72	0,05	0,002	0,719	0	0	0,716
Be	ICP-AES	mg/kg	0,05	0	0,06	14	0,442	0,235	1,06	0,141	0	0	0	0	0	0	0	0
Bi	GF-AAS	mg/kg	0,005	2,5	<0.005	3,89	0,049	0,026	0,164	0,021	0	0,008	0	0,002	0,008	0	0	0,008
Br_INAA	INAA	mg/kg	0,5	25	<0.5	56	4,8	3,7	5,37	3,71	0	0	0	0	0	0	0	0
Ca	ICP-AES	mg/kg	3	0	110	41700	2279	1905	2383	1075	0	0	0	0	0	0	0	0
Ca_XRF	XRF	wt.- %	0,005	0	0,03	6,76	2,133	2,17	0,899	0,801	0,003	0	0	0	0	0	0,003	0,003
Cd	GF-AAS	mg/kg	0,001	0	0,007	0,221	0,029	0,024	0,02	0,01	0	0	0	0	0	0	0	0
Ce_INAA	INAA	mg/kg	3	0	12	500	59	45	53,2	23,7	0	0	0	0,036	0	0	0	0
Co	ICP-AES	mg/kg	0,2	0	1,2	44,3	8,2	7	5,03	3,71	0	0,765	0,5	0,085	0,766	0	0	0,767
Co_INAA	INAA	mg/kg	1	0,2	<1	57	14,3	13	6,72	5,93	0	0	0	0	0	0	0	0
Cr	ICP-AES	mg/kg	0,5	0	2,2	471	36,2	28,4	35,09	16,23	0	0,662	0,03	0,137	0,663	0	0	0,662
Cr_INAA	INAA	mg/kg	5	5	11	910	116	99	87,5	46,0	0	0,015	0	0,001	0,015	0	0	0,015
Cu	ICP-AES	mg/kg	0,5	0	2	149	22,0	16,2	18,44	10,82	0	0,134	0,5	0,307	0,134	0	0	0,134
Eu_INAA	INAA	mg/kg	0,2	0	0,3	14,3	1,24	1,05	1,006	0,371	0	0	0	0	0	0	0	0
Fe	ICP-AES	mg/kg	10	0	3310	79200	17236	14700	10189	7154	0	0,315	0,5	0,56	0,316	0	0	0,271
Fe_XRF	XRF	wt.- %	0,02	0	0,59	12,35	3,605	3,43	1,4	1,342	0	0,059	0,5	0,533	0,059	0,82	0	0
Hf_INAA	INAA	mg/kg	1	0	2	120	6,5	6,0	6,59	1,48	0	0	0	0	0	0	0	0
K	ICP-AES	mg/kg	200	0,5	<200	11000	1478	1100	1295	741	0	0,138	0	0,006	0,138	0	0	0,135
K_XRF	XRF	wt.- %	0,04	0	0,36	5,24	1,558	1,41	0,593	0,482	0	0,236	0	0,017	0,237	0	0	0
La	ICP-AES	mg/kg	0,5	0	3,5	203	17,9	12,8	20,96	6,45	0	0	0	0	0	0	0	0
La_INAA	INAA	mg/kg	0,5	0	6,1	310	30,6	24	29,10	13,34	0	0	0	0,063	0	0	0	0
Li	ICP-AES	mg/kg	0,5	0	1,7	70,9	9,1	7,2	6,88	4,30	0	0,002	0,05	0,521	0,002	0	0	0,002
Lu_INAA	INAA	mg/kg	0,05	0	0,05	2,67	0,37	0,30	0,263	0,163	0	0,151	0,04	0,42	0,152	0	0	0,152
Mg	ICP-AES	mg/kg	5	0	370	70500	4741	3720	4815	2002	0	0,526	0	0,113	0,525	0	0	0,444
Mg_XRF	XRF	wt.- %	0,02	0	0,12	7,32	1,271	1,15	0,677	0,526	0	0,046	0	0,042	0,046	0	0	0
Mn	ICP-AES	mg/kg	0,5	0	33,8	2140	185,2	128,5	179,53	65,83	0	0	0	0	0	0	0	0
Mn_XRF	XRF	wt.- %	0,008	0	0,015	0,356	0,059	0,054	0,031	0,022	0	0	0	0	0	0	0	0
Na	ICP-AES	mg/kg	15	0	20	19400	338	140	1368	89	0	0	0	0	0	0	0	0
Na_XRF	XRF	wt.- %	0,02	0	0,08	4,87	2,26	2,45	0,678	0,504	0	0	0	0	0	0	0	0
Nd_INAA	INAA	mg/kg	5	1,3	<5	220	22,4	18	19,44	8,90	0	0	0	0	0	0	0	0
Ni	ICP-AES	mg/kg	1	0	1,2	228	23,4	18,7	21,09	11,56	0	0,854	0,5	0,274	0,854	0	0	0
P	ICP-AES	mg/kg	7	0	59	7170	446	393	368	185	0	0,658	0,01	0,133	0,656	0	0	0
P_XRF	XRF	wt.- %	0,004	0	0,004	0,589	0,045	0,039	0,032	0,019	0	0,68	0	0	0,679	0	0	0
Pb	GF-AAS	mg/kg	0,2	0	0,3	45,3	2,7	1,6	3,33	0,74	0	0	0	0	0	0	0	0
Rb	INAA	mg/kg	15	6,3	7,5	270	60	54	33,6	26,7	0	0	0	0	0	0	0	0
S	ICP-AES	mg/kg	5	0,5	<5	531	41	30	42,6	17,8	0	0	0	0,025	0	0	0	0
Sc	ICP-AES	mg/kg	0,1	0,2	<0.1	15,4	2,8	2,3	1,81	1,19	0	0,162	0	0	0,162	0	0	0

Table 2
Continued

Element	Technique	Unit	DL	%<DL	Min	Max	Mean	Median	Sdev	Mad	Orig	ln_S-W	ln_K-S	ln_chi^2	Log	Sqrt	Range	Logit
Sc_INAA	INAA	mg/kg	0,1	0	1,7	36	13,6	13,0	5,70	5,93	0	0	0	0	0	0,046	0	0
Se	GF-AAS	mg/kg	0,01	4,1	<0.01	1,15	0,067	0,046	0,086	0,034	0	0	0,07	0,809	0	0	0	0
Si	ICP-AES	mg/kg	10	0	50	590	154,182	140	64,9	44,5	0	0	0	0,002	0	0	0	0
Si_XRF	XRF	wt.-%	0,23	0	17,05	40,27	31,46	31,74	2,579	2,216	0	0	0	0	0	0	0	0
Sm_INAA	INAA	mg/kg	0,1	0	0,9	37	4,0	3,4	3,11	1,63	0	0	0,04	0,044	0	0	0	0
Sr	ICP-AES	mg/kg	0,5	0	1,6	1040	25,3	7,7	98,23	3,78	0	0	0	0	0	0	0	0
Te	GF-AAS	mg/kg	0,003	22,6	<0.003	0,271	0,011	0,008	0,015	0,007	0	0	0	0	0	0	0	0
Th	ICP-AES	mg/kg	3	6,1	<3	66	7,9	6,5	6,16	3,71	0	0	0	0	0	0	0	0
Th_INAA	INAA	mg/kg	0,2	0	1	54	7,2	5,8	4,95	3,41	0	0,652	0,02	0,008	0,651	0	0	0
Ti	ICP-AES	mg/kg	0,5	0	48,8	5730	895	807	515,2	405,5	0	0	0	0	0	0,099	0	0
Ti_XRF	XRF	wt.-%	0,003	0	0,053	1,9	0,362	0,347	0,16	0,151	0	0,068	0	0,001	0,068	0	0	0
V	ICP-AES	mg/kg	0,5	0	4,5	183	35,0	30,9	19,65	15,72	0	0,983	0,5	0,087	0,983	0	0	0
Y	ICP-AES	mg/kg	0,5	0	0,9	169	6,4	4,4	10,97	2,37	0	0	0	0	0	0	0	0
Yb_INAA	INAA	mg/kg	0,2	0	0,3	19,9	2,4	1,9	1,84	1,04	0	0,037	0	0	0,037	0	0	0
Zn	ICP-AES	mg/kg	0,5	0	3,7	348	27,4	20,9	24,17	12,45	0	0,622	0,1	0,144	0,622	0	0	0

transform these back. Using this approach one should get much more realistic values for mean and standard deviation even for skewed data. This approach is far from ideal for several reasons (1) it is awkward, (2) the above tests demonstrated that, for the majority of variables, a log-transformation does not result in a normal distribution, and (3) there are better and easier methods (e.g. using median and mad). Another solution could be to first remove the outliers and then calculate mean and standard deviation. However, the problem is then the definition of outliers – which values should already be removed and which still included for the calculations. And why go to such lengths when easier and better methods exist?

Fig. 3 shows the same plot as Fig. 2. Here the data were first log-transformed (ln), mean and standard deviation and median and mad were calculated and then transformed back. Fig. 3 shows that there are still big differences between mean and median and standard deviation and mad. Thus a log-transformation of the data, as suggested in the vast majority of textbooks in geochemistry, is no real solution for approaching a normal distribution when working with geochemical or environmental data. Their real characteristic is that they are skewed (Vistelius 1960!) and the frequent occurrence of outliers, originating from another, superimposed population.

The vast majority of advanced statistical methods and tests are based on the assumption of an underlying normal distribution of the data. How then should one continue with data analysis for geochemical or environmental samples? Exploratory data analysis (EDA – Tukey 1977) was especially developed to deal with such situations. It provides a large number of simple graphical techniques for study of the data in detail prior to use of any advanced statistical technique (e.g. Velleman and Hoaglin 1981; Dutter and others 1992). Fig. 1 gives just one example of some of these graphics. The CDF-diagrams could directly be used to detect more than one population in a data set (see discussion above). For Th the one-dimensional scattergram beneath the histogram shows that the data were reported in 0.1-mg/kg-steps up to 10 mg/kg and then, suddenly in 1-mg/kg-steps, leading to fragmented data at the upper end of the distribution (Fig. 1). The boxplot is another prominent example of one of these graphics. It can be used in combination with the histogram as in Fig. 1. In combination with intelligently chosen data subgroups it is an even more powerful tool for detection of the important information in a data set (Reimann and Wurzer 1986; Reimann 1987, 1989b; Reimann and others 1988) in a simple, graphical way without any assumptions on data behaviour. EDA should in general be the first step in the analysis of geochemical and environmental data.

There are quite a number of further techniques that do not make any assumptions about the data distribution. These run in general under the name of nonparametric methods (e.g. Afifi and Azen 1979; Conover 1980; Puri and Sen 1985; Noether, G.E. 1991; Gore and others 1993). Robust methods (Huber 1981; Hampel and others 1986)

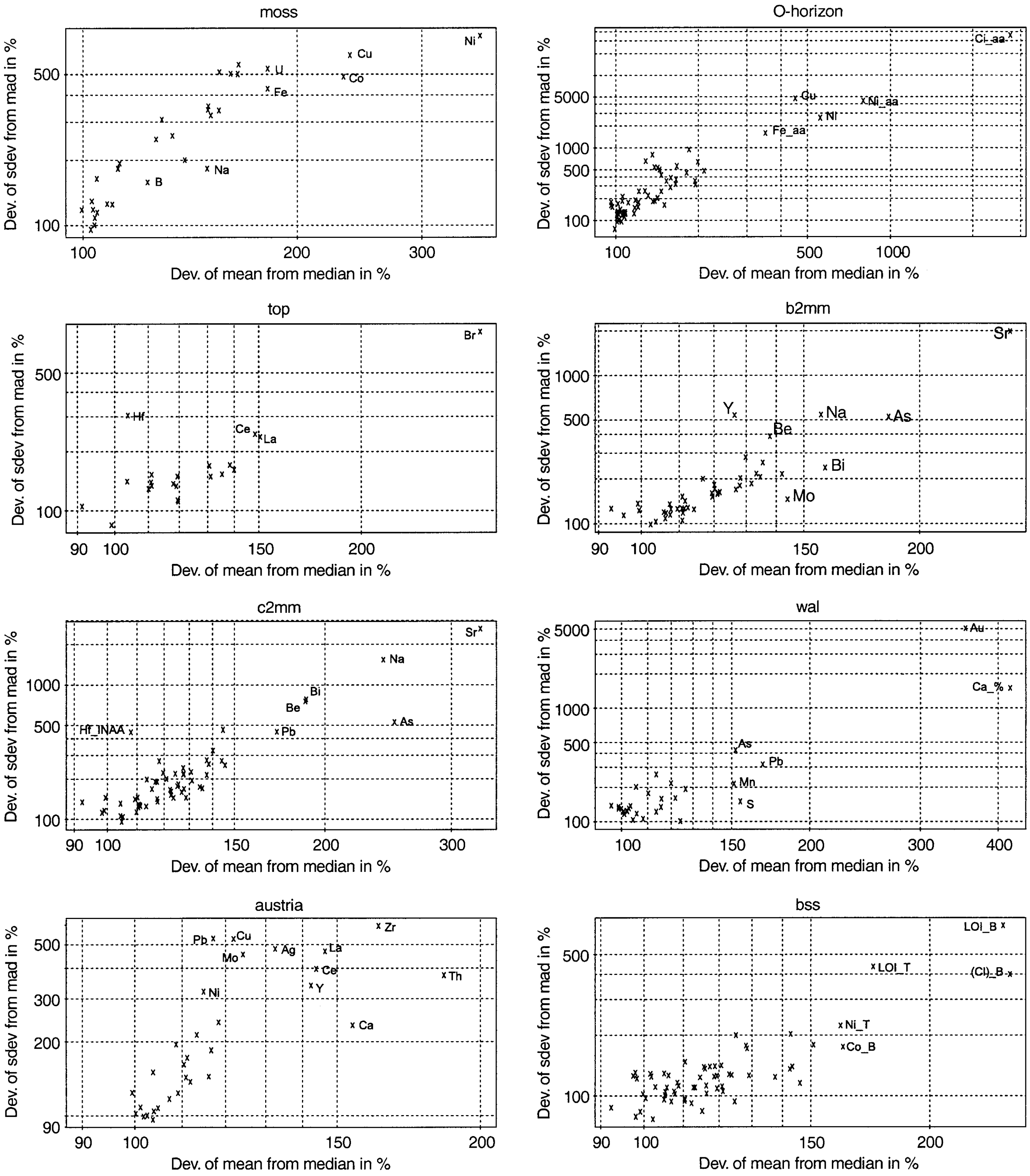

Fig. 2
Plot of the deviation between median and mean and standard deviation and mad, expressed in % for the original data in 8 selected data sets

take care against data outliers. These methods should be the first choice when dealing with geochemical and environmental data.

Other classical statistical techniques, e.g. factor analysis by the maximum likelihood method, not only require that each variable entered shows a normal distribution

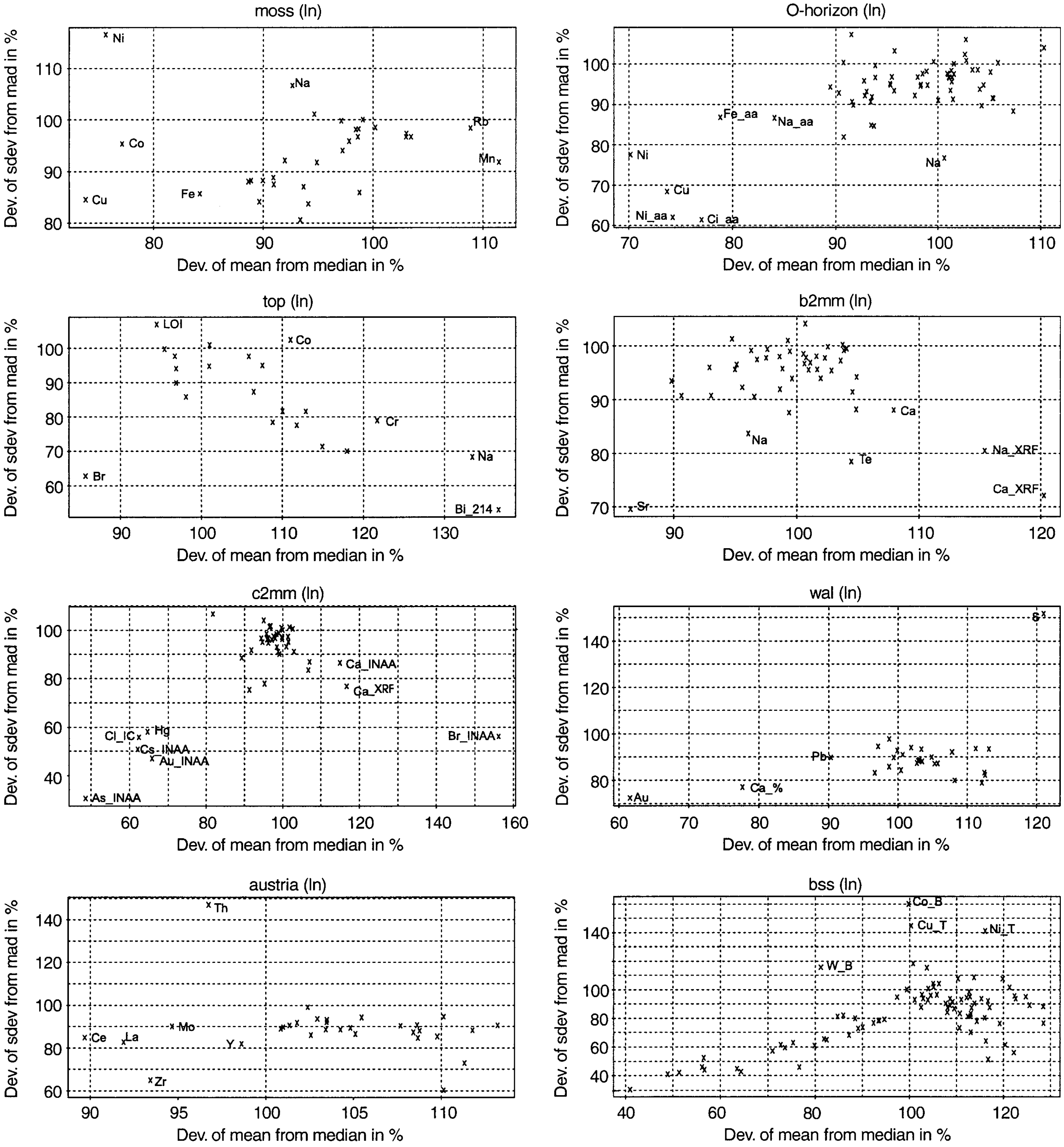

Fig. 3
Plot of the deviation between median and mean and standard deviation and mad, expressed in % in 8 selected data sets. Data were first log-transformed (ln) and the estimators were then back transformed

but in addition that the whole data set shows a multivariate normal distribution. None of the test data sets comes even close to a multivariate normal distribution, neither the original data (Fig.4) nor the log-transformed (ln) data (Fig.5). Such methods should thus be avoided in treating geochemical and environmental data.

What influence has the absence of normal and lognormal distributions in geochemistry on one of the most frequently used techniques: correlation analysis? An easy

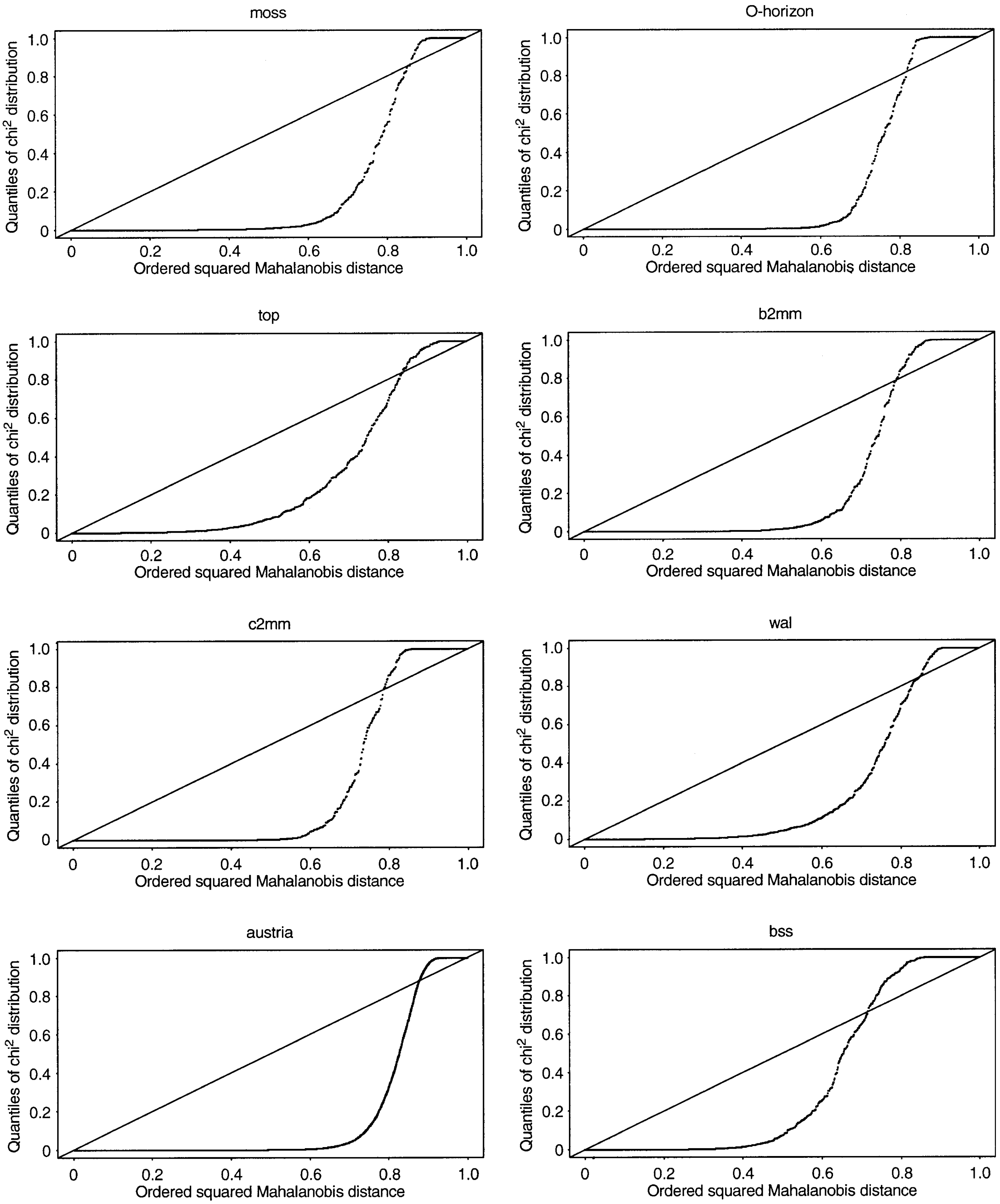

Fig. 4
Deviation of the test data sets from a multivariate normal distribution, which should follow the straight 45°-line

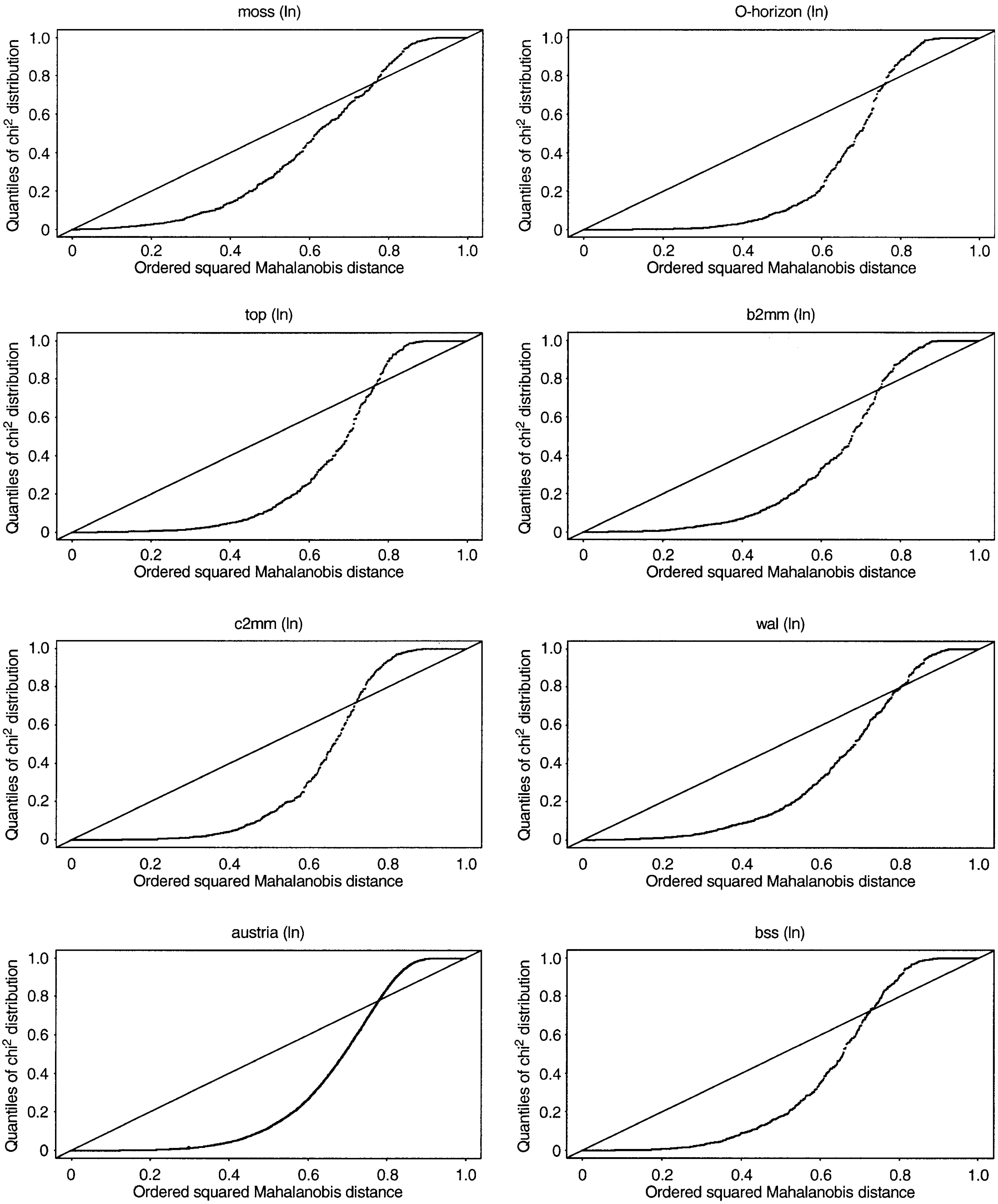

Fig. 5
Deviation of the log-transformed (ln) test data sets from a multivariate normal distribution, which should follow the straight 45°-line

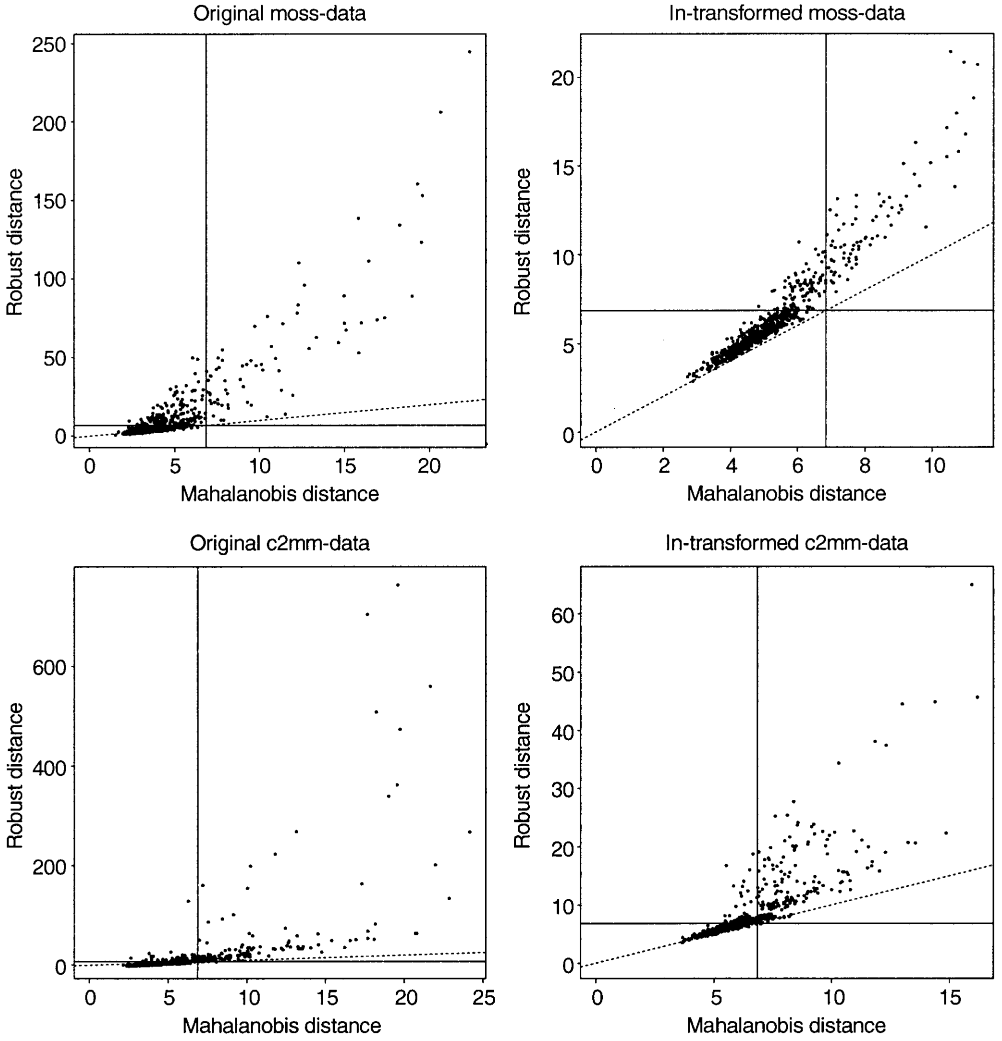

Fig. 6
Distance–distance plots for the moss and C-horizon test data sets, demonstrating the high number of outliers and their influence of outliers on correlation analysis. Normally distributed data would follow the *stippled line*. A log-transformation (ln) of the data results in a slight improvement only

way to visualise the sensibility of correlation to outliers is the distance-distance (D-D) plot (Rousseeuw and Van Driessen 1999) which is, for example, implemented in the software package S-PLUS (Venables and Ripley 1997). Here the Mahalanobis distance is drawn against the robust distance. Robust distances can be obtained by introducing robust counterparts to the arithmetic mean and the sample variance-covariance matrix into the formula for the Mahalanobis distance. A fast and stable possibility for estimation of mean and covariance in a robust way is to use the minimum covariance determinant (MCD) estimator (Rousseeuw 1985). The objective is to find those h out of n observations, typically $h=0.75$ n, for which the classical covariance matrix has the lowest determinant. The MCD estimate of location is then the average of these h points, whereas the MCD estimate of scatter is their covariance matrix. On both axes of the D-D plot the cutoff value $\sqrt{\chi^2_{p;0.975}}$ separates outlying observations. All points in the D-D plot should plot near the stippled line if outliers did not corrupt the data. Deviations from this line indicate that the classical estimators for mean and covariance, and hence also the correlation matrix, are biased.

Fig. 6 shows, for two examples, the moss and the C-horizon data-sets of the Kola project, that large differences have to be expected between the results of a robust correlation analysis and correlation analysis performed with the original data. The situation is improved when log-transformed (ln) data are entered into correlation analysis (Fig. 6). But even after log-transformation (ln) of the original data there is still a large number of outliers (Fig. 6). This outcome is a clear indication that all non-robust correlation-based methods will deliver distorted results with geochemical and environmental data. A robust correlation matrix, which can be obtained by the MCD-estimator, should be used as a foundation for all correlation based methods (e.g. principal component analysis, factor analysis). Another solution would be to first use techniques that are well suited to detect data outliers (e.g. EDA), remove these and then continue with more advanced methods. When, for example, uncritically entering a factor analysis even with the original (log-transformed) data the results will be governed by the process(es) causing the "high" values. The result is thus biased and could be easily predicted by much simpler methods.

Table 3
Frequently used statistical parameters, tests and multivariate methods and their suitability for regional geochemical and environmental data which neither show a normal or lognormal distribution

Location and spread	
Arithmetic mean	Should only be used in special cases
Geometric mean	Can be used, but may be problematic in some cases
Median	Can be used, should be first choice as location estimator
Robust mean (Hampel or Huber)	Can be used
Standard deviation	Should not be used if data outliers exist
Mad (medmed)	Can be used
Hinge spread	Can be used
Robust spread	Can be used
Tests for comparability of means/variances	
t-test	Should not be used
F-test	Should not be used
"Notches" in boxplot	Can be used, very easy and fast
Nonparametric tests	Can be used
Robust tests	Can be used
Multivariate methods	
Correlation analysis	Should not be used with the original (untransformed) data OK in graphical form (e.g. draftman's display)
Regression analysis	Should not be used with the original (untransformed) data
Robust regression analysis	Can be used, preferably on log-transformed data
Nonparametric regression analysis	Can be used, preferably on log-transformed data
Principal component analysis (PCA)	Very sensible to outlying observations, should not be used
Robust PCA	Can be used, preferably with log-transformed data
Factor analysis	Very sensible to outlying observations, should not be used
Robust factor analysis	Can be used, preferably with log-transformed data
Discriminant analysis	Very sensible to outlying observations, should not be used
Correspondence analysis	Very sensible to outlying observations, should not be used
Cluster analysis	Can be used
Partial least squares (PLS)	Very sensible to outlying observations, should not be used
Robust PLS	Can be used
ANOVA	Very sensible to outlying observations, should not be used
Robust ANOVA	Can be used

Table 3 gives a collection of the most frequently used statistical parameters, tests, and methods in geochemistry and environmental sciences together with an estimation of their vulnerability to non-normally distributed data.

Conclusions

It has been suggested (Ahrens 1953) that geochemical data as a law show a lognormal distribution. This is, however, a rare exception in geochemistry (e.g. when several analyses are carried out on the same samples or when all samples where taken from one outcrop or one rock unit over a very limited area – Vistelius 1960). Regional geochemical and environmental data almost never follow a normal distribution. In the majority of cases a data transformation (e.g. log, ln, logit, square root or range) will not result in a normal distribution. This observation has serious consequences for the further statistical treatment of geochemical and environmental data that are widely neglected.

Mean and standard deviation, which are the best estimators of location and spread for data that follow a normal distribution, are far from ideal when used for regional geochemical or environmental data. The reason for the strong skew in data sets from geochemistry and environmental sciences is often that the samples represent more than one population/process. In most cases the best measure of location for such data is the median. The geometric mean may be an acceptable alternative (but has a number of other associated dangers – see discussion in Rock 1988). As a measure of spread, the median absolute deviation (mad) or the hinge-spread (Tukey 1977) should be used instead of the standard deviation which is very vulnerable to the existence of data outliers. Due to the very different information that mean and median represent for skewed data, it may be justified to present both in data tables.

The vast majority of classical statistical methods are based on the assumption of a normal distribution in the data entered. If using them with non-normally distributed data one should be very aware that this could give biased or even faulty results. Data outliers do not influence robust methods. Non-parametric methods are not based on model assumptions. These are thus preferable to the classical methods. In any case, a thorough univariate analysis and documentation of geochemical and environmental data sets is an absolute necessity before using more advanced statistical methods. Some multivariate methods and statistical tests require not only that each variable shows a normal distribution but also a multivariate normal distribution. None of the test data sets came even close to a multivariate normal distribution. A log-transformation (ln) of the data resulted only in a slight improvement, not however, in multivariate normal distributions. Methods requiring a multivariate normal distribution are especially vulnerable when used with geochemical and environmental data and will often deliver unstable and faulty results.

Geochemists and environmental scientists should realise that in very many cases they are actually presenting biased and faulty results by still believing in the lognormal law of distribution of their data. It is high time that they stop to uncritically use techniques that were not made for such situations. Today there exist a multitude of statistical techniques giving correct results. Already the simple study of distributions in graphics will often give more important geochemical insights than very advanced statistical methods – as suggested 40 years ago by Vistelius (1960).

Acknowledgements Discussions on the existence of normal and lognormal data distributions in geochemistry and environmental sciences over nearly two decades with R. Garrett, R. Dutter, S.P. McGrath, F. Koller, and J. Matschullat were appreciated. Discussions with G. Kashulina on the significance of different distributions for the recognition of the underlying geochemical processes were very valuable. The Austrian Geological Survey is thanked for the permission to use the data of the stream sediment geochemical atlas of Austria. An improvement of the English presentation by R. Boyd is appreciated.

References

Afifi AA, Azen SP (1979) Statistical analysis: a computer oriented approach. Academic Press, New York

Ahrens LH (1953) A fundamental law of geochemistry. Nature 172 : 1148

Ahrens LH (1954a) The lognormal distribution of the elements (a fundamental law of geochemistry and its subsidiary). Geochim Cosmochim Acta 5 : 49–74

Ahrens LH (1954b) The lognormal distribution of elements. II. Geochim Cosmochim Acta 6 : 121–132

Ahrens LH (1957) Lognormal-type distribution. III. Geochim Cosmochim Acta 11 : 205–213

Aubrey KV (1954) Frequency distribution of the concentrations of elements in rocks. Nature 174 : 141–142

Aubrey KV (1956) Frequency distributions of elements in igneous rocks. Geochim Cosmochim Acta 9 : 83–90

Chayes F (1954) The lognormal distribution of elements: a discussion. Geochim Cosmochim Acta 6 : 119–121

Conover WJ (1980) Practical non-parametric statistics, 2nd edn. J Wiley, New York

Dutter R, Leitner T, Reimann C, Wurzer F (1992) Grafische und geostatistische Analyse am PC. (Beiträge zur Umweltstatistik) Schriftenreihe der Technischen Universität Wien, Bd 29, pp 78–88

Easton GS, McCulloch RE (1990) A multivariate generalization of quantile-quantile plots. J Am Stat Assoc 85 : 376–386

Gore AP, Deshpande JV, Shanubhogue A (1993) Statistical analysis of non-normal data. J Wiley, New York

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics. The approach based on influence functions. J Wiley, New York

Hazen A (1914) Storage to be provided in impounding reservoirs for municipal water supply. Trans Am Soc Civil Eng 77 : 1529–1669

Huber PJ (1981) Robust Statistics, J Wiley, New York

McGrath SP, Loveland, PJ (1992) The soil geochemical atlas of England and Wales. Blackie Academic, London

Miller RL, Goldberg ED (1955) The normal distribution in geochemistry. Geochim Cosmochim Acta 8 : 53–62

Noether GE (1991) Introduction to statistics: the nonparametric way. Springer, New York Berlin Heidelberg

Philip GM, Watson DF (1987) Probabilism in geological data analysis. Geol Mag 124 : 577–583

Puri ML, Sen PK (1985) Nonparametric methods in general linear models. J Wiley, New York

Reimann C (1987) Aussagekraft der geochemischen Basisaufnahme – Mineralogische, geochemische und statistische Detailuntersuchungen an Bachsedimenten im alpinen Bereich. (Ber GBA Bd10) Verlag der geologischen Bundesanstalt, Wien

Reimann C (1989a) Reliability of geochemical analyses: recent experiences. – Trans Inst Min Metal 98 : B123-B130

Reimann, C. (1989b) Untersuchungen zur regionalen Schwermetallbelastung in einem Waldgebiet der Steiermark. In: Forschungsgesellschaft Joanneum (ed) Umweltwissenschaftliche Fachtage – Informationsverarbeitung für den Umweltschutz, Graz 1989; pp 39–50

Reimann C, Wurzer F (1986) Monitoring accuracy and precision – improvements by introducing robust and resistant statistics. Mikrochim Acta 2 : 31–42

Reimann C, Kürzl H, Wurzer F (1988) Applications of exploratory data analysis to regional geochemical mapping. In: Thornton I. (ed) Monograph series environmental geochemistry and health: Geochemistry and health. Proceedings of the Second International Symposium London 1987. Science Reviews Ltd., Northwood, pp 21–27

Reimann C, Äyräs M, Chekushin VA, Bogatyrev I, Boyd R, Caritat P de, Dutter R, Finne TE, Halleraker JH, Jæger Ø, Kashulina G, Niskavaara H, Lehto O, Pavlov V, Räisänen M L, Strand T, Volden T (1998) Environmental geochemical atlas of the Central Barents Region. NGU–GTK–CKE spec publ. Geol Surv Norway, Trondheim

Rieder H (1994) Robust asymptotic statistics. Springer, New York Berlin Heidelberg

Rock NMS, Webb JA, McNaughton NJ, Bell G (1987) Nonparametric estimation of averages and errors for small datasets in isotope geoscience: a proposal. IsotGeosci 66 : 163–177

Rock NMS (1988) Numerical geology. (Lecture Notes in Earth Sciences 18) Springer Verlag, New York Berlin Heidelberg

Rousseeuw PJ (1985) Multivariate estimation with high breakdown point. In: Grossmann W, Pflug G, Vincze I, Wertz W (eds) Mathematical statistics and applications, vol B. Reidel, Dordrecht pp 283–297

Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41 : 212–223

Shapiro SS, Wilk MB (1965) An analysis of variance test for normality. Biometrika 52 : 591–611

Smirnov NV (1948) Table for estimating the goodness of fit of empirical distributions. Annal Math Stat 19 : 279–281

Thalman F, Scherman O, Schroll E, Hausberger G (1989) Geochemischer Atlas der Republik Österreich, Textteil. Geologische Bundesanstalt Wien, Österreich

Tukey JW (1977) Exploratory Data Analysis. Addison-Wesley, Reading

Velleman PF, Hoaglin DC (1981) Applications, basics and computing of exploratory data analysis. Duxbury Press, Boston, Mass

Venables WN, Ripley BD (1997) Modern applied statistics with S-PLUS, 2nd edn. Springer, New York Berlin Heidelberg

Vistelius AB (1960) The skew frequency distributions and the fundamental law of the geochemical processes. J Geol 68 : 1–22