Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Geochemistry 20 (2005) 1857–1874
Applied Geochemistry
www.elsevier.com/locate/apgeochem

Statistical characterization of a large geochemical database and effect of sample size

Chaosheng Zhang [a,*], Frank T. Manheim [b], John Hinde [c], Jeffrey N. Grossman [b]

[a] *Department of Geography, National University of Ireland, Galway, Ireland*
[b] *U.S. Geological Survey, MS 954, 12201 Sunrise Valley Dr., Reston, VA 20192, USA*
[c] *Department of Mathematics, National University of Ireland, Galway, Ireland*

Received 25 April 2004; accepted 4 June 2005
Editorial handling by Asa Danielsson
Available online 19 August 2005

Abstract

The authors investigated statistical distributions for concentrations of chemical elements from the National Geochemical Survey (NGS) database of the U.S. Geological Survey. At the time of this study, the NGS data set encompasses 48,544 stream sediment and soil samples from the conterminous United States analyzed by ICP-AES following a 4-acid near-total digestion. This report includes 27 elements: Al, Ca, Fe, K, Mg, Na, P, Ti, Ba, Ce, Co, Cr, Cu, Ga, La, Li, Mn, Nb, Nd, Ni, Pb, Sc, Sr, Th, V, Y and Zn. The goal and challenge for the statistical overview was to delineate chemical distributions in a complex, heterogeneous data set spanning a large geographic range (the conterminous United States), and many different geological provinces and rock types. After declustering to create a uniform spatial sample distribution with 16,511 samples, histograms and quantile–quantile (Q–Q) plots were employed to delineate subpopulations that have coherent chemical and mineral affinities.

Probability groupings are discerned by changes in slope (kinks) on the plots. Major rock-forming elements, e.g., Al, Ca, K and Na, tend to display linear segments on normal Q–Q plots. These segments can commonly be linked to petrologic or mineralogical associations. For example, linear segments on K and Na plots reflect dilution of clay minerals by quartz sand (low in K and Na). Minor and trace element relationships are best displayed on lognormal Q–Q plots. These sensitively reflect discrete relationships in subpopulations within the wide range of the data. For example, small but distinctly log-linear subpopulations for Pb, Cu, Zn and Ag are interpreted to represent ore-grade enrichment of naturally occurring minerals such as sulfides.

None of the 27 chemical elements could pass the test for either normal or lognormal distribution on the declustered data set. Part of the reasons relate to the presence of mixtures of subpopulations and outliers. Random samples of the data set with successively smaller numbers of data points showed that few elements passed standard statistical tests for normality or log-normality until sample size decreased to a few hundred data points. Large sample size enhances the power of statistical tests, and leads to rejection of most statistical hypotheses for real data sets. For large sample sizes (e.g., $n > 1000$), graphical methods such as histogram, stem-and-leaf, and probability plots are recommended for rough judgement of probability distribution if needed.

* Corresponding author. Fax: +353 91 525700.
E-mail address: Chaosheng.Zhang@nuigalway.ie (C. Zhang).

0883-2927/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apgeochem.2005.06.006

1. Introduction

Like other geochemical survey projects of large scope (e.g., Darnley et al., 1995; Darnley, 1997; Plant et al., 2001), the National Geochemical Survey (NGS) of U.S. Geological Survey had its origin in either baseline investigation or mineral resource surveys. In recent years the NGS data set has been further augmented (USGS, 2004), better quality-controlled, and has targeted new applications, including regional characterization of toxic elements and other environmental management issues.

The purposes of this study were to: (1) create a uniformly gridded set of the NGS database; (2) provide a brief overview of the elemental distribution and descriptive statistics for the latter set; and (3) attempt to delineate the presence of normal and lognormal subpopulations by display of probability plots. A challenge for a concise statistical overview lies in the fact that the NGS spans the entire conterminous United States and reflects many different geological substrates, as well as transport and depositional processes. It includes elemental enrichments from both mineralized areas as well as sites reflecting anthropogenic influences.

Because of the severe deviation from normality to be expected in data from such spatially extended, heterogeneous data (Reimann and Filzmoser, 2000) it was concluded that it would be best to use graphical, descriptive methods to visualize the elemental distributions. The authors attempt to "let the data speak" in ways that delineate meaningful relationships. Probability plots are powerful tools for this purpose because they are able to compress large numbers of data representing broad ranges of concentration and complex relationships into relatively simple curves whose variability in shape can reflect geochemical relationships in systematic but nuanced ways. Quantile–quantile (Q–Q) plots yield significant detail on the behavior of subpopulations at both the high and low end of the frequency spectrum, as well as those having a central tendency. The applicability of Q–Q plots to delineation of clustered properties in even very large data sets has been recently illustrated by its use in data mining in space research (e.g., Manku et al., 1999).

Major elements in rocks were already well analyzed and linked to their mineral associations by the early 20th century (Clarke, 1924). Led by the pioneering Norwegian geochemist, V.M. Goldschmidt (1954 and references cited therein), a first-order understanding of the role and distribution of minor and trace elements was achieved by the middle of the 20th century. Statistics were first applied about this time, in large part to provide better prediction and quantification (mapping) of the distribution of mineral deposits (Krige, 1951, 1960, 1966; Sichel, 1952; Miesch and Riley, 1961; and references in Ahrens, 1965). Ahrens (1954) stimulated debate among geochemists by asserting a geochemical "law" that postulated the lognormal distribution of trace elements in igneous minerals and rocks about a geometric mean. "Dispersion" of the data was described by the standard deviation of the logarithms of values. Geochemists soon presented objections to the universality of the lognormal scientific law. For example, Aubrey (1956) found that the abundance and distribution of host minerals for trace elements, as well as mode of classification of rocks could be key factors in determining dispersion of the elements. Vistelius (1960) proposed, instead of the lognormal law, a "fundamental law of geochemical processes". This was defined as "The joint probability distribution function of the concentration of the minor chemical element deposited by natural chemical reactions has a large positive skewness", i.e., a long tail toward high concentrations that reflects rarer but higher degrees of enrichment. Zhang and Selinus (1998) considered the lognormal distribution as a special case of skewed distributions in studies of trace elements in a suite of tills from Sweden. Post World War II contributions by leading Soviet earth scientists like A.B. Vistelius (Vistelius, 1960) and A.P. Vinogradov (Vinogradov, 1959) introduced western geochemists to extensive geochemical surveys conducted earlier in the Soviet Union (e.g., Razumovsky, 1940) but obscured by language and political barriers.

In the 1960s and 70s the advent of mainframe computers stimulated more elaborate analyses of statistical relationships in the chemical composition of sediment and rock samples. These analyses included use of probability plots that facilitated visualization of normal and lognormal distributions as straight-line segments (Sinclair, 1974, 1976; Chambers et al., 1983). Mandelbrot (1982) introduced the concept of fractal distributions (plots of frequency/log concentration) as a descriptor of a wide range of earth properties. Mandelbrot's work stimulated study of fractal distributions in describing elemental enrichments in earth materials (Bölviken et al., 1992; Cheng et al., 1994, and references cited therein). The advent of computers greatly expanded use of statistical techniques to define genetic affinities in geochemical populations. However, Reimann and Filzmoser (2000) point out that some of the most popular techniques, like correlation analysis, principal component analysis, and factor analysis, as well as ANOVA methods, are based on the assumption of normality. They criticize use of such methods for all but especially heterogeneous data sets because of the almost invariable presence of outliers and hence absence of normality in the data sets.

Major advances were made possible since the 1980s by availability of ever more powerful desktop computers and statistical and mapping software. These tools made working with and displaying larger data populations much easier than in earlier periods. Additional stimuli for the compilation of large regional data sets were inter-

est in toxic elements in surface sediment and water, as mentioned previously, and advent of powerful new analytical tools for trace element analyses. Recently, the probability distribution of trace elements in very large numbers of ground-water samples in the STORET database (U.S. Environmental Protection Agency, 2002) was investigated by Newcomb and Rimstidt (2002). In spite of the convergence of enabling factors, preparation and documentation of large regional sediment data sets remains a large task, and such data sets are only beginning to become readily accessible. Their statistical treatment remains an area that includes many unresolved problems.

2. Basic data and methods

2.1. Sampling and analyses

The NGS program of the U.S. Geological Survey (USGS, 2004) comprises chemical analyses of samples collected over a period of about 30 a. The majority of the samples were collected during the National Uranium Resource Evaluation (NURE) Program, sponsored by the Department of Energy as a search for U resources. As part of the NURE program, the Hydrogeochemical Stream Sediment Reconnaissance (HSSR) collected several hundred thousand samples, principally stream sediments, from about 2/3 of the area of the US, and analyzed them for trace elements by a wide variety of methods; most of these samples were collected between 1976 and 1980. Approximately 40,000 of these NURE samples, spanning the full areal extent of the program, were selected for reanalysis by modern (post-1998) methods for the NGS. The NGS also contains data for several thousand samples of the same types, collected by the USGS between 1970 and 1990. Finally, several thousand new samples were collected in areas not covered by the NURE or USGS sample archives. Approximately 2% of the samples were analyzed in duplicate.

Because the NGS was assembled from multiple sources and covers the diverse terrains and sample media (stream sediments, lake sediments, soils, etc.) present in the US, sample collection protocols were not the same for all samples. The majority of samples, about 80%, represent stream sediments sieved to below 100 mesh (<150 μm grain size). Approximately 12% of the samples are soils, principally from areas where stream sediments were difficult to obtain; these were also sieved to 100 mesh. The remainder of the samples were lake, pond, playa, and spring sediments, mostly representing geographic areas where those sample media were dominant.

The NGS sampling density (samples/km^2) is somewhat variable across the US as the result of the differing protocols of the component studies responsible for collecting the samples. The original NURE HSSR program collected sediment samples at an average density of 1 sample per 17 km^2. The NGS subsampled the NURE collection for reanalysis by superimposing a grid of 17×17 km cells, and choosing one NURE sample at random from each cell; thus, the NGS sample density is generally 1/289 km^2. In some study areas, the NURE collection was sampled more intensively, resulting in densities of up to 1/100 km^2 (e.g., New England and the northern Great Basin; Fig. 1(a)), and the entire NURE suite was reanalyzed in southeastern Kentucky (Fig. 1(a)). Non-NURE sampling areas in the NGS have densities ranging from 1/100 km^2 (e.g., Florida and Mississippi) to 1/289 km^2 (e.g., Michigan); a few small areas were sampled at high densities (e.g., 1/2.5 km^2 around Tallahassee, Florida).

All NGS samples were analyzed by a consistent set of techniques, mostly at a single laboratory facility, XRAL Laboratories (Canada). Among these methods is one using inductively coupled plasma-atomic emission spectrometry (ICP-AES) after acid digestion for which statistics are calculated in this report. Sample aliquots of 200 mg were decomposed using a mixture of HCl, HNO_3, $HClO_4$, and HF at low temperature. The digested samples are aspirated into the ICP-AES discharge where the elemental emission signal is measured simultaneously for 40 elements. Calibration is performed by standardizing with digested rock reference materials and a series of multi-element solution standards. Analytical data for an experimental run are deemed acceptable based on results determined for two in-house quality-control standards: recovery for all elements present at >5 times the detection limit must be within 15% of the certified value, and the calculated relative standard deviation of duplicate samples is no greater than 15%.

2.2. Descriptive statistical summary

Table 1 provides a summary of the chemical data for the original data set. Of note are several elements with variable detection limits (e.g., Cd, Nb) and wide ranges of concentrations, and thus calculation for the percentiles that are below detection limits is arbitrary. It should be noted that other NURE data sets exist, performed by alternative analytical methods, in which lower detection limits permit much larger proportions of trace metals like Ag, As, Cd, etc. to be quantitatively estimated. The current study is limited to the largest coherent database and to those elements for which sufficient data above the limit of detection are available. Improved and more uniform analytical quality was a goal of reanalysis of the present data set by common techniques (see next section).

The NGS data set encompasses much of the conterminous United States and many heterogeneous rock and sediment distributions. Besides heterogeneity in

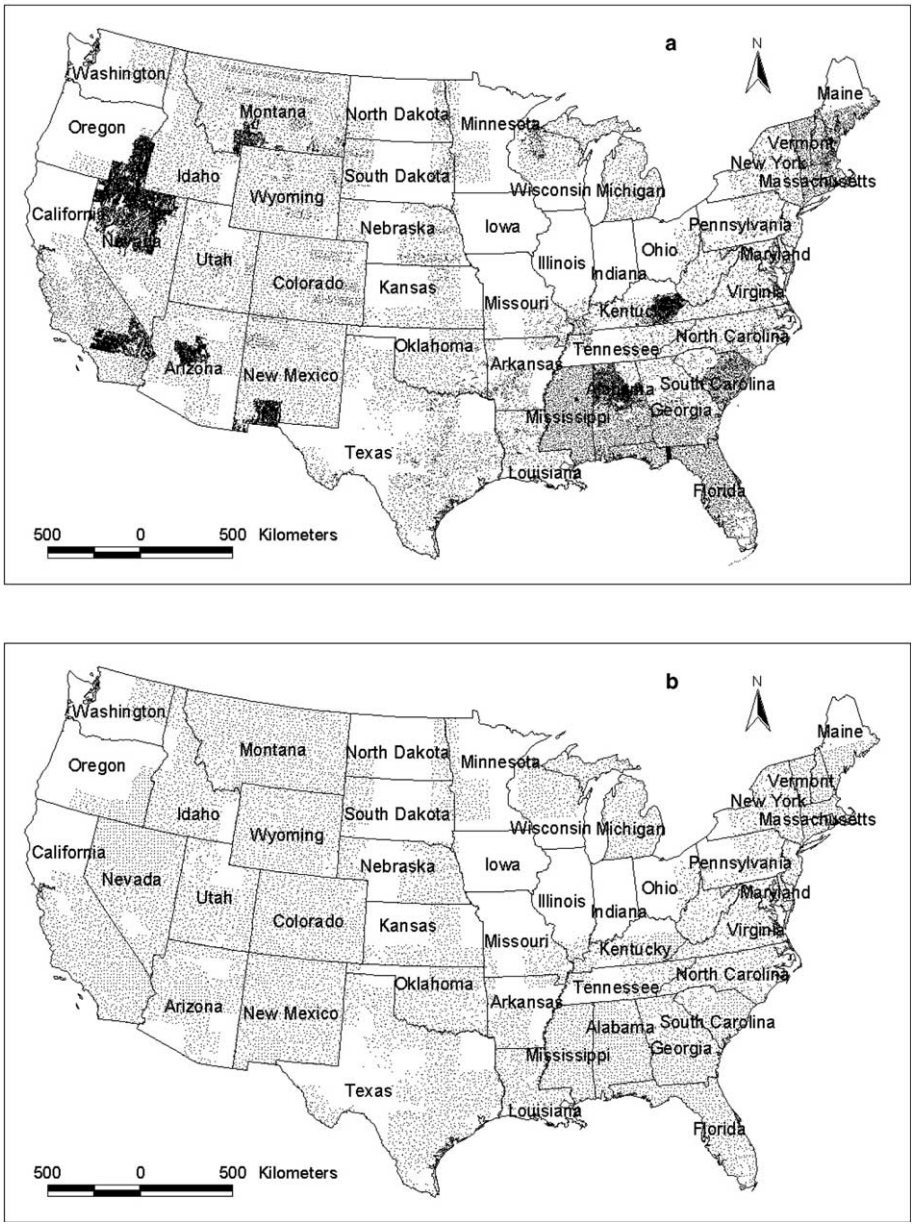

Fig. 1. Sampling locations in the conterminous US for the NGS program of the U.S. Geological Survey: (a) original samples ($n = 48{,}544$); (b) re-sampled samples based on a 17×17 km grid system ($n = 16{,}511$).

the NGS data set, potential sources of variability that affect probability distribution include sampling methods, the precision and accuracy of analytical methods, and samples below detection limits. Approximations can be used to estimate means in the presence of values below detection limit (Helsel and Hirsch, 1992). However, in this study medians are used rather than means in the descriptive statistical table. For the purpose of calculating graphical display the authors have used the common convention of assigning the value of concentrations below detection limits to half the detection limits.

2.3. Declustering by subsampling

Due to the uneven spatial distribution of the sampling locations, the dataset were subsampled to create a more uniform distribution across the whole conterminous US. Since the majority of the areas were sampled at a density of one sample per 17×17 km grid

Table 1
Percentiles of 40 elements analyzed by ICP-AES following a 4-acid near-total digestion for the original data set ($n = 48{,}544$; Al–Ti are reported in wt%; Ag–Zn units are μg/g)

		$n >$ DL	Min	5%	10%	25%	Median	75%	90%	95%	Max
Al	48,544	48,539	<0.005	0.45	0.99	3.11	5.40	7.10	7.99	8.36	17.0
Ca	48,520	48,284	<0.005	0.02	0.04	0.31	1.58	2.86	4.59	6.52	39.6
Fe	48,544	48,375	<0.005	0.30	0.60	1.47	2.40	3.50	4.90	6.20	64.4
K	48,544	47,895	<0.005	0.05	0.17	0.88	1.61	2.05	2.40	2.64	17.0
Mg	48,532	48,100	<0.005	0.02	0.04	0.22	0.62	1.03	1.51	2.00	20.7
Na	48,536	47,083	<0.005	0.01	0.02	0.23	0.92	1.67	2.20	2.50	8.16
P	48,500	47,444	<0.005	0.01	0.01	0.03	0.06	0.09	0.12	0.15	9.00
Ti	48,544	48,538	<0.01	0.13	0.17	0.25	0.35	0.52	0.73	0.91	4.50
Ag	46,895	840	<0.2	<0.5	<2	<2	<2	<2	<2	<2	342
As	46,661	12,412	<5	<10	<10	<10	<10	8	14	20	5870
Au	46,663	102	<4	<8	<8	<8	<8	<8	<8	<8	27
Ba	48,542	48,540	<1	40	91	278	556	750	951	1100	16,000
Be	48,544	29,496	<1	<1	<1	<1	1	2	2	2	53
Bi	48,544	3368	<2	<5	<10	<10	<50	<50	<50	13	816
Cd	48,544	3425	<0.4	<2	<2	<2	<2	<2	<2	0.6	130
Ce	48,544	48,382	<5	23	33	49	65	86	130	183	12,000
Co	48,544	42,624	<2	<2	<2	5	8	13	19	25	695
Cr	48,544	46,646	<2	3	6	15	34	63	111	170	12,000
Cu	48,544	48,544	0.5	1	3	7	13	23	34	47	3800
Eu	43,171	42,200	<2	1	1	1	1	1	1	2	51
Ga	46,895	40,656	<4	<4	<4	8	14	18	21	23	65
Ho	43,183	1611	<4	<4	<4	<4	<4	<4	<4	<4	32
La	48,544	48,362	<2	11	16	25	33	44	67	95	8700
Li	48,544	46,882	<2	3	5	12	20	28	36	43	910
Mn	48,539	48,484	<4	62	121	289	540	830	1130	1399	41,970
Mo	48,544	14,225	<2	<2	<2	<2	<2	2	5	6	431
Nb	48,544	43,113	<2	<4	<4	7	12	17	23	28	468
Nd	44,832	41,902	<4	<9	12	20	28	39	59	84	3200
Ni	48,544	43,918	<2	<3	3	8	15	23	35	49	2420
Pb	48,544	47,374	<20	6	8	13	19	25	33	42	27,394
Sc	48,544	43,595	<2	<2	<2	4	7	11	15	18	53
Sn	46,895	5742	<2	<5	<5	<5	<50	<50	2	17	743
Sr	48,531	48,087	<2	7	15	59	177	330	449	540	7500
Ta	43,183	15	<40	<40	<40	<40	<40	<40	<40	<40	301
Th	48,544	41,093	<2	<6	<6	7	10	15	24	38	1180
U	46,895	269	<1	<10	<100	<100	<100	<100	<100	<100	882
V	48,544	48,418	<2	10	17	38	65	96	140	184	2991
Y	48,544	47,379	<2	4	6	12	17	23	32	41	433
Yb	43,183	35,405	<1	<1	<1	1	2	3	3	4	30
Zn	48,542	48,147	<2	7	13	32	59	82	104	129	27,040

cell, a 17×17 km grid system was created in a GIS layer to cover the whole conterminous US. The closest sample to each grid node was selected and marked. A "join" function in GIS database management was employed to relate the selected samples and the raw data set in order to avoid duplicate selection. In this way, a total of 16,511 samples out of 48,544 samples were selected (Fig. 1(b)). Basic results for the declustered data are shown in Table 2. A caveat should be noted about the declustering step. Geological/regional geochemical data are collected from sample space that may have structure that even detailed field mapping may miss.

2.4. Graphic display

Histograms are time-tested tools for displaying frequency distributions and rely on dividing the data into classes (or bins). The data following a normal distribution will show a symmetric bell-shaped histogram, whereas a tail towards high values implies that the data are positively skewed. Fig. 2 shows the histograms of a computer-generated data set following the standard normal distribution and another following a lognormal distribution.

Offsetting their popularity and value in providing visual approximations to distributions like the bell

Table 2
Percentiles of 40 elements analyzed by ICP-AES following a 4-acid near-total digestion for the re-sampled data set ($n = 16,511$; Al–Ti are reported in wt%; Ag–Zn units are μg/g)

	n	*n* > DL	Min	5%	10%	25%	Median	75%	90%	95%	Max
Al	16,511	16,508	<0.005	0.60	1.25	3.07	4.78	6.16	7.41	8.00	13.0
Ca	16,511	16,446	<0.005	0.03	0.06	0.37	1.33	2.76	4.78	7.23	38.0
Fe	16,511	16,459	<0.005	0.38	0.67	1.38	2.12	2.97	4.16	5.21	64.4
K	16,511	16,319	<0.005	0.08	0.29	0.94	1.59	1.97	2.30	2.58	17.0
Mg	16,511	16,385	<0.005	0.02	0.06	0.24	0.57	0.96	1.45	1.88	19.6
Na	16,510	16,088	<0.005	0.01	0.05	0.25	0.71	1.19	1.81	2.18	8.16
P	16,504	16,176	<0.005	0.01	0.01	0.03	0.05	0.08	0.10	0.13	2.77
Ti	16,511	16,511	0.01	0.13	0.17	0.22	0.30	0.44	0.66	0.84	3.19
Ag	16,359	102	<0.2	<2	<2	<2	<2	<2	<2	<2	53
As	16,329	3822	<5	<10	<10	<10	<10	<10	14	17	5870
Au	16,329	24	<4	<8	<8	<8	<8	<8	<8	<8	16
Ba	16,509	16,508	<1	56	130	300	506	673	836	974	12,330
Be	16,511	8652	<1	<1	<1	<1	1	1	2	2	53
Bi	16,511	1028	<2	<10	<10	<50	<50	<50	<50	12	816
Cd	16,511	427	<0.4	<2	<2	<2	<2	<2	<2	<2	103
Ce	16,511	16,458	<4	22	32	47	64	84	124	177	4880
Co	16,511	14,799	<2	<2	<2	5	8	11	16	20	695
Cr	16,511	15,888	<1	3	5	12	25	45	78	96	4340
Cu	16,511	16,511	1	1	3	6	11	17	26	35	3050
Eu	16,109	15,467	<2	1	1	1	1	1	2	3	51
Ga	16,359	14,225	<4	<4	<4	8	13	17	22	25	65
Ho	16,109	999	<4	<4	<4	<4	<4	<4	<4	4	32
La	16,511	16,453	<2	10	15	23	32	42	63	90	2920
Li	16,511	16,072	<2	3	5	12	20	27	36	43	501
Mn	16,507	16,485	<4	73	133	269	442	705	1050	1357	19,190
Mo	16,511	7037	<2	<2	<2	<2	<2	3	5	6	431
Nb	16,511	14,839	<2	<4	<4	8	12	17	24	30	410
Nd	16,261	15,199	<4	<9	12	19	28	38	56	79	2610
Ni	16,511	14,972	<2	<3	3	7	13	20	30	39	2130
Pb	16,511	16,282	<4	7	9	14	20	26	35	44	24,310
Sc	16,511	14,980	<2	<2	2	4	7	9	13	16	53
Sn	16,359	1281	<2	<5	<5	<50	<50	<50	<50	19	392
Sr	16,510	16,369	<2	10	21	65	142	243	389	489	4700
Ta	16,109	4	<40	<40	<40	<40	<40	<40	<40	<40	301
Th	16,511	13,345	<2	<6	<6	7	10	15	24	38	1180
U	16,359	27	<1	<100	<100	<100	<100	<100	<100	<100	882
V	16,511	16,481	<2	13	19	37	60	84	118	150	782
Y	16,511	16,198	<2	4	7	11	16	21	29	38	433
Yb	16,109	13,539	<1	<1	<1	1	2	2	3	4	30
Zn	16,509	16,395	<2	9	14	30	52	73	97	118	27,040

shape of a normal distribution, histograms have limitations. The number of bins affects the shape of the curve. Too few bins may conceal important subpopulations or create apparent truncation of the low end of the distribution. Too many bins may create ragged plots. In this study, histograms for the raw datasets were plotted using about 20–30 bins, a compromise range chosen for effectiveness in displaying the major features of the data and revealing skewness. A curve representing the normal distribution corresponding to the mean and standard deviation of the data set is superimposed on the histogram for the purpose of comparison. Separated extreme values frequently occur at the high end and may be identified as outliers. In this study, due to the wide range of the raw data, the extremely high values were truncated in order to show the majority of the data better. The limitations of the histograms are offset by probability plots described below.

Normal quantile–quantile (Q–Q) plots, familiar to physical scientists as cumulative probability plots, are created by plotting observed values of a variable against the corresponding normal quantiles. A popular way to display such a plot is to plot the normal quantiles (or

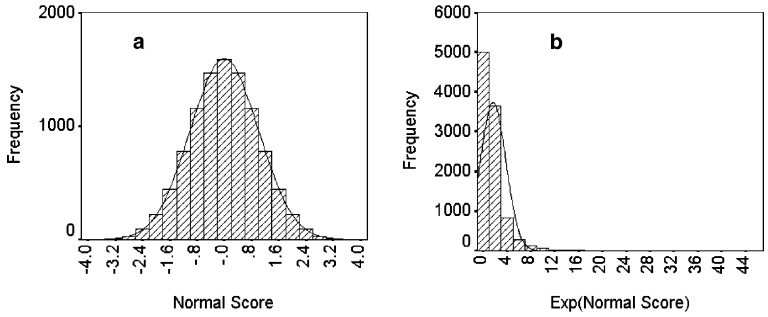

Fig. 2. Histograms for simulated distributions: (a) normal, (b) lognormal ($n = 10{,}000$).

scores) of the standard normal distribution on the x-axis with values usually between -4 to 4, and the variable under study on the y-axis. In this study, the observed values (in μg/g or %) are plotted on the x-axis, and values expected for a normal distribution are plotted on the y-axis because the function was displayed in this manner in the software package available (SPSS® for Windows, version 11.0). If the samples are from a normal distribution, points will cluster along a straight line.

The steps in constructing the normal Q–Q plot used in this study are as follows: First, the raw data were ranked in ascending order (duplicated values were assigned with different ranks as their expected values are different in this plot). The expected normal values are calculated by taking the z-scores of $(i - 0.375)/(n + 0.25)$ cumulative probability (Blom, 1958), where i is the rank in increasing order, and n is the number of samples. Then, the scores are converted to the expected normal values based on the mean (μ) and standard deviation (σ) using the function of $(z \times \sigma + \mu)$. Commonly used normal Q–Q plots do not carry out this conversion, but the conversion is applied in this study to offer visual comparison of the "expected normal value" at the same scale as the original data. It avoids using the abstract concept of "normal scores" on the figures. Curvature in the trend of points indicates departures from normality.

Fig. 3 demonstrates normal Q–Q plots for computer-simulated data sets following standard normal and lognormal distributions. All samples of the normal data set are located on the diagonal line, whereas the lognormal data exhibit a convex shape.

In this study, the normal Q–Q plot is used for both raw and logarithmically transformed data. In the raw data plot normal distributions will follow a straight-line behavior, whereas log-transformed data will display straight-line segments for lognormal distributions. In order to permit comparisons, the normal Q–Q plots for all the elements were plotted on one page.

2.5. Selective random sampling and tests for normality

The effect of changing sample size on standard hypothesis tests for normality were studied, using the Kolmogorov–Smirnov (K–S) test (Chakravarti et al., 1967). The K–S test measures the degree to which a given data set follows a specific theoretical distribution (such as normal, uniform, or Poisson). The test statistic of K–S is based on the largest absolute difference between the observed and the theoretical cumulative distribution functions. The K–S test assumes that the parameters (e.g., mean and standard deviation) of the test distribution are specified in advance, whereas the Lilliefors correction (Lilliefors, 1967) for the K–S test is applied when means and variances are not known and must be estimated from the data.

The K–S test is chosen over others because it is widely applied and considered conservative. The Anderson–Darling test (Stephens, 1974) is an example of an alternative test that gives more weight to the tails than the K–S test which tends to be more sensitive near the center of the distribution. The authors therefore prefer the K–S test to minimize the effect of outlier populations for many of the elements. During comparison studies, discrepant results were obtained for K–S test values performed by SPSS® and S-Plus® statistical software. The problem was resolved when it was found that the K–S results obtained with the S-Plus® application automatically utilized the Lilliefors correction. For comparison, the Lilliefors significance correction for the K–S test using SPSS® was also applied in the following experiments for effect of sample size.

2.6. Computer software[1]

The raw data were stored in a dBASE file (dbf format), and basic calculations were performed using Microsoft Excel®. Most of the statistical calculations were accomplished with SPSS® software (version 11.0). The sampling location maps were produced with ArcView® GIS software (version 3.3).

[1] Mention of commercial products in this paper does not imply endorsement by National University of Ireland, Galway or the U.S. Geological Survey.

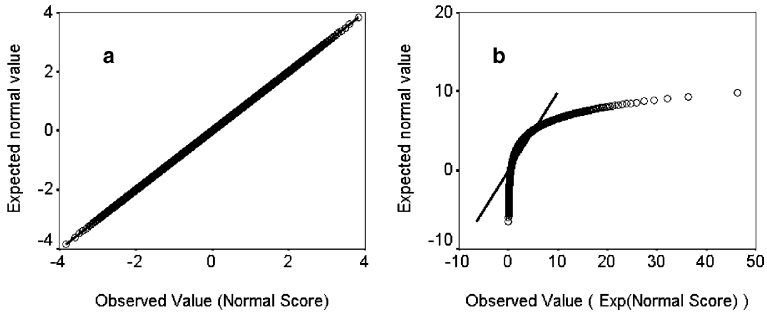

Fig. 3. Normal Q–Q plots for simulated distributions: (a) normal, (b) lognormal ($n = 10{,}000$).

3. Results

3.1. Basic statistics and element selection

The number of samples re-sampled from the NGS database was 16,511. Table 2 shows the estimated percentiles for the elements. It may be mentioned that the NGS database includes some samples that were analyzed by other methods that may have lower detection limits (DLs) for some elements, and provide measurements of constituents not included here (e.g., SiO_2, loss on ignition, etc.).

Table 2 provides information about both element concentrations and data quality. Elements with too few values above detection limits can provide only limited information, e.g., for mineral exploration surveys seeking high values. Inadequate detection limits are observed for 11 elements: Ag, As, Au, Be, Bi, Cd, Ho, Mo, Sn, Ta and U. The 25th percentiles for all these elements are below the detection limits. Some other elements do not have a good precision (either analytical resolution or the use of only one significant figure). This yields many values in low integers, such as Eu and Yb. For example, 5–75th percentiles of Eu are all equal to 1 μg/g, i.e., at least 70% of the values for Eu are 1 μg/g. Accordingly, these 13 elements were regarded as having inadequate data quality, and were not used in the ensuing statistical analyses. Among the remaining elements, the 10th percentile levels for Co, Ga, Nb and Th were below the detection limits. However, their precision appeared much better than those of Eu and Yb, and they were selected. The 10th percentiles for all the other elements were above the detection limits, and thus they were regarded as of adequate quality for this investigation. Altogether, a total of 27 elements were selected for detailed statistical analyses: Al, Ca, Fe, K, Mg, Na, P, Ti, Ba, Ce, Co, Cr, Cu, Ga, La, Li, Mn, Nb, Nd, Ni, Pb, Sc, Sr, Th, V, Y and Zn.

The uppermost ranges of concentration for several elements, including Fe, Cu, Zn, Pb, Ag, Ti, and rare earths, approach ore grade or byproduct metal extraction grade, although high concentrations of elements like Cu, Pb, and Zn can also be influenced by pollution.

3.2. Histograms

Histograms for the raw data of the elements are shown in Fig. 4. On the x-axis of these plots, mid-values of the bins are shown once every two bins to facilitate a clear display for all the elements on one page.

None of the elements show the bell-shaped distribution expected of a normal distribution pattern, but many approach a lognormal distribution as illustrated in Fig. 2. Most elements have high values, which are not well reflected in the histograms because extreme values form only a part of the highest concentration bins. However, the extreme values are especially well visualized on the later Q–Q plots. Many elements have values below (e.g., Co, Ga, Sr and Th) or close to (e.g., Ca, K, Mg, and Na) the detection limits (see also Tables 1 and 2), resulting in the high frequencies for the lowest value group.

Except for Al, the frequency distributions of most of the other 26 elements under study are positively skewed and include some very high values. Mixed populations, that are expected to be prevalent in the current dataset, are likewise not well displayed on histograms, and are discussed under the other graphical methods later.

Some extreme values appear separated from the majority of the samples, i.e., do not appear to be part of a continuous distribution (e.g., Fig. 4, Table 2). The term "outlier" depends on the purpose of a statistical investigation, but these very high samples would normally fit such a description, and in exploration geochemistry might be regarded as evidence of mineralization or other rare processes.

3.3. Q–Q plots

On normal Q–Q plots (Fig. 5), elements that are building blocks of major rock-forming minerals, Al,

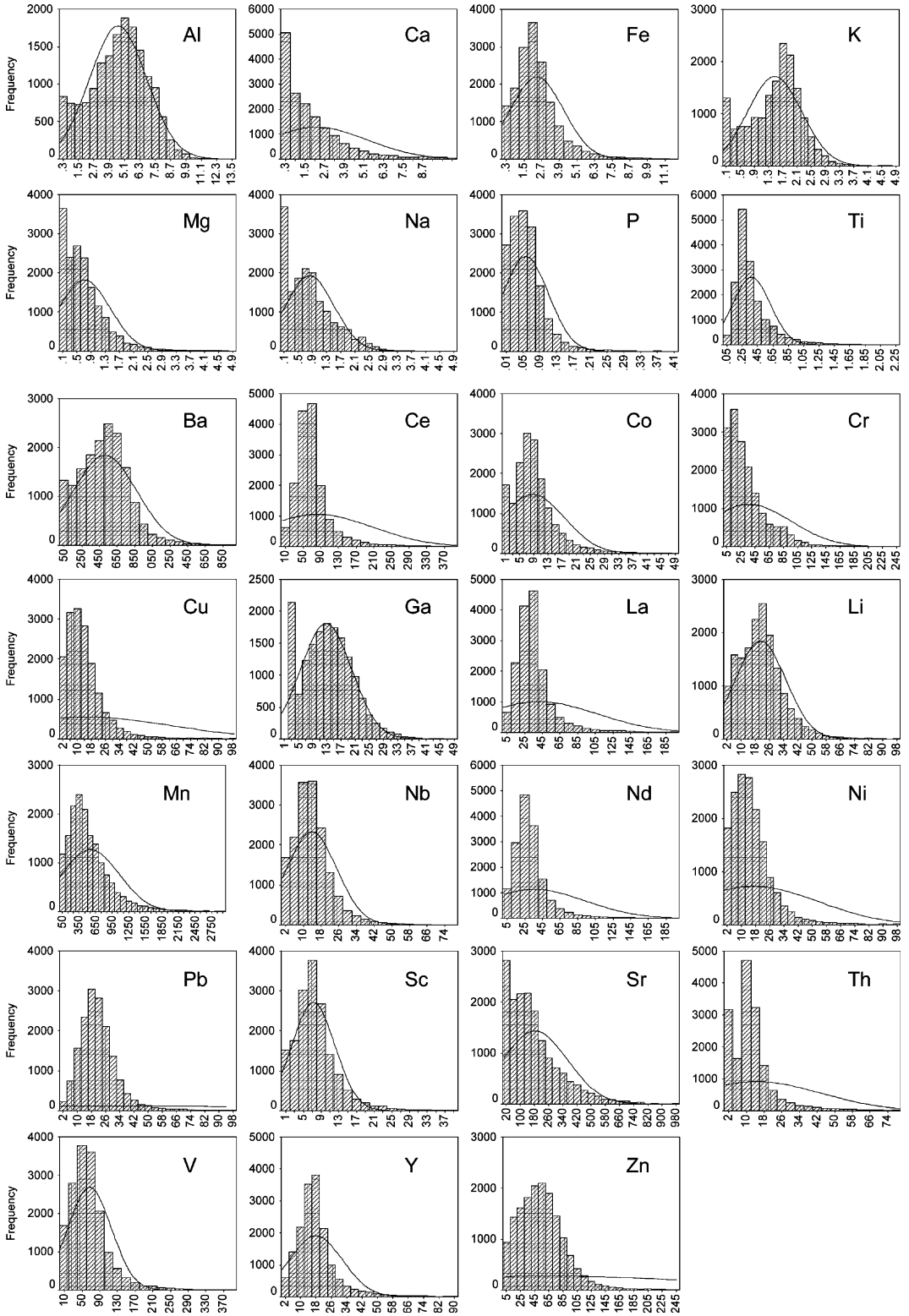

Fig. 4. Histograms of raw data ($n = 16{,}511$, values larger than shown are truncated; units: Al–Ti in wt%; Ba–Zn in μg/g; a normal distribution curve for all the values is superimposed for comparison).

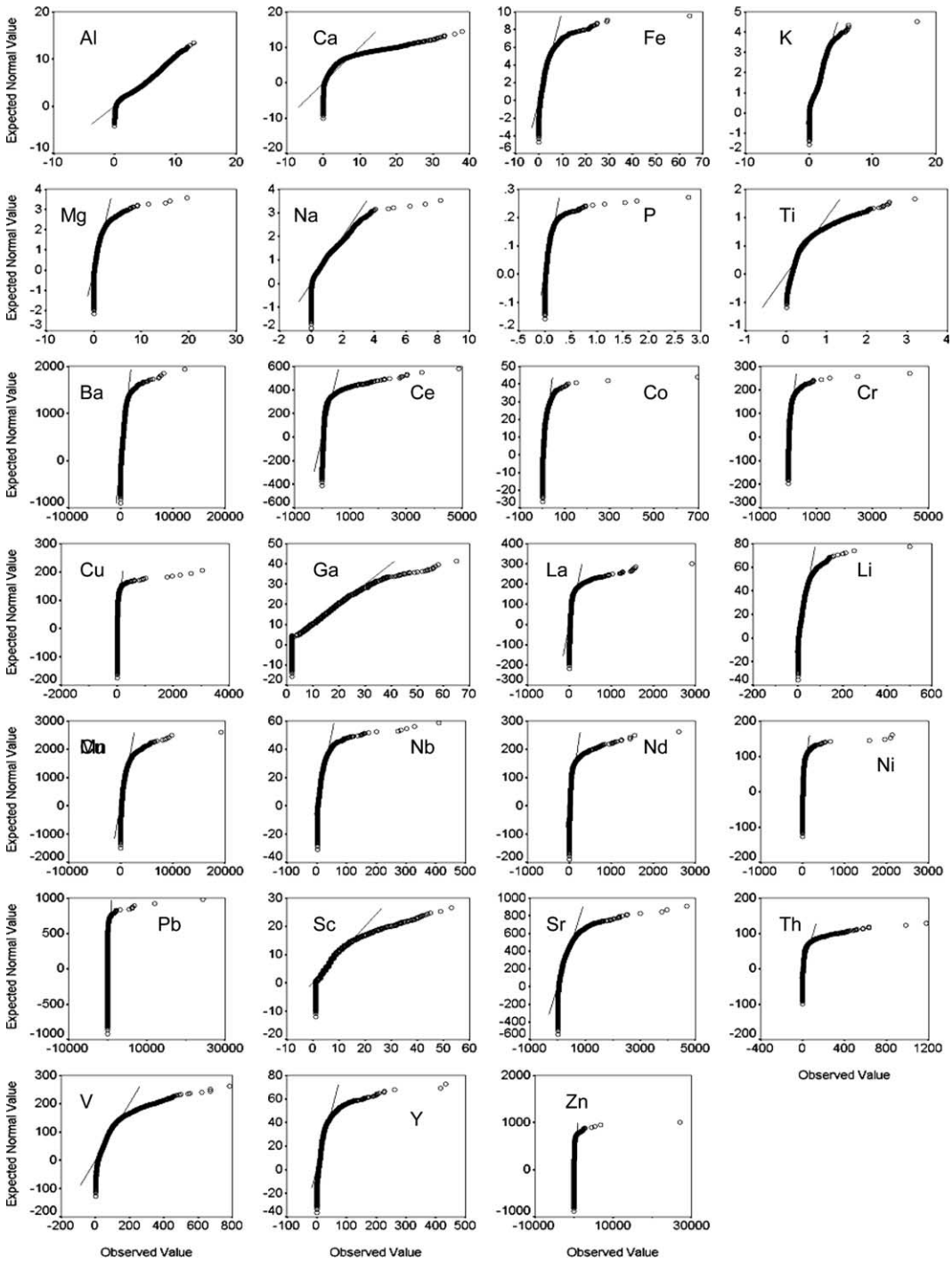

Fig. 5. Normal Q–Q plots for raw data ($n = 16{,}511$; units: Al–Ti in wt%; Ag–Zn in μg/g). Negative values on the x-axis (observed values) are included in order to preserve symmetry for the expected values (y-axis).

Fe, Mg, K, and Na, tend to show straight-line segments. In the case of Fe and Mg straight segments that approximate the modeled normal line are obscured by the steepness of the plots. Straight segments are also observed for some trace elements, including Ga and Sc. Samples below detection limit form the vertical line on the left side of the diagram. For ore-forming elements, departures from normality occur toward the

highest values, located on the right side of the straight normal distribution line. The result is referred to as the positively skewed or convex distribution feature for most elements shown on the histograms. It should be noted that the slopes of the plots cannot be directly compared, as axial scales vary to encompass the wide disparity in concentrations. Most of the trace elements do demonstrate some form of convex shapes as in Fig. 3, showing that they are composed of multiple phases and have closer overall similarities to lognormal distributions.

The plots provide useful information for outlier detection. For example, single samples of Fe and K are separated from the distribution curve. Similar outliers can be identified for other elements.

The plots for log-transformed data sets (Fig. 6) show improved normal (straight-line) probability features for most elements, except Al and K. They also indicate mixtures of populations by identified kinks or changes in slope for many elements including Ca, Fe, Ba, Ce, La, Li, Nd, Ni, Pb, Sr, Th and Zn. The elements P, Ti, Co, Cr, Li, Nb, Ni, V and Y approach lognormal distributions most closely. Owing to the magnification of values in low ranges, the logarithmic transformation renders multiple values due to detection limits as vertically stacked points more clearly than in the normal plots. Nondetects were all assigned values of half the detection limits. Samples recorded as being below detection limits may extend below their assigned values. In cases where lognormal behavior appears or is independently known to be consistent throughout the concentration range, a more realistic distribution of values may be predicted by extending the slope of the straight line.

The logarithmic transformation ''overtransforms'' some elements, changing the overall curve shape from convex to concave, especially for major elements such as Al, K, Mg and Na. This feature is expected where the elements in question have significant populations that tend towards normal distributions, such as elements that are the major components of common rock-forming minerals. The curves for Ca, Ba, and Fe depart from those of other major constituents as discussed later. A tendency to approach the modeled slope on the left side of the plot may be associated with dilution effects likewise detailed in the Discussion section. This feature may be observed for Cu, Pb, P, Ga, Sc, Th, and the rare earth elements, Ce, La, Y, and Nd.

The plots reveal a number of consistent and interesting slope segments. Elements like Cu, Zn and Pb show a sharp reduction in slope (a ''fat tail'') at high concentration values. The ''ferrides'', Co, Cr, Ni, Co and V, show the most consistent tendencies toward lognormality throughout their concentration range.

3.4. Shape parameters and tests for normality

To quantify the probability features of the data sets, the shape parameters of skewness and kurtosis were calculated and are listed in Table 3. The K–S test was performed to test hypotheses of data normality. The significance levels (K–S p) are also listed in the table. In this table, the kurtosis values are calculated relative to a normal distribution that has a kurtosis of 3.

The results from the K–S test show that none of the 27 elements pass the test for either normality or log-normality, in line with previous expectations. However, in this study, raw data for Al and Ga seem to approach a normal distribution as both their skewnesses and kurtoses are close to ''0''. Some of the log-transformed data also approach lognormality based on their shape parameters.

Widely recognized factors related to the non-normality and non-lognormality, such as mixture of populations, outliers, analytical precision and detection limits, can all be identified in this study. However, another problem, large sample size, may also be a factor affecting statistical treatment. To investigate the effect of sample size on probability distribution, 100 randomly chosen independent selections of 50, 100, 200, 500, 1000, 2000, and 5000 samples were taken from the declustered sample set (16,511 samples). The K–S test for normality and log-normality was applied to all these re-sampled data sets. For each sample number and element, a total of 100 significance values of K–S p were obtained. The percentage of p-values higher than the widely used 0.05 significance level provides a measure of the likelihood that the data pass the test for normality or lognormality. Meanwhile, the Lilliefors correction for the K–S test was also applied in this study. To evaluate the re-sampling strategy and for comparison, a computer simulated normal distribution data set with $n = 16{,}511$ was tested using the same procedure. Table 4 shows the results of these experiments.

The results show that when the sample size is 50, departure from normality is seldom detected, and it is easy for a data set to pass the test for normality at the 0.05 significance level. With increase in sample size, the statistical test gains power. When the sample size reaches 2000, only one experiment (for Al) passes the K–S test for normality, and none of the other elements pass. This result is in agreement with the results shown in Table 3, that none of the elements pass the K–S test for normality when all the 16,511 samples were used. The element closest to normal distribution is Al, which is also consistent with the observations from the above methods. Other elements that are close to normality include K, Ba and Ga.

Similar results for the effect of sample size on probability distribution are observed for the log-transformed data. With smaller sample sizes of 100 or less many elements tend to pass the test for lognormality. However,

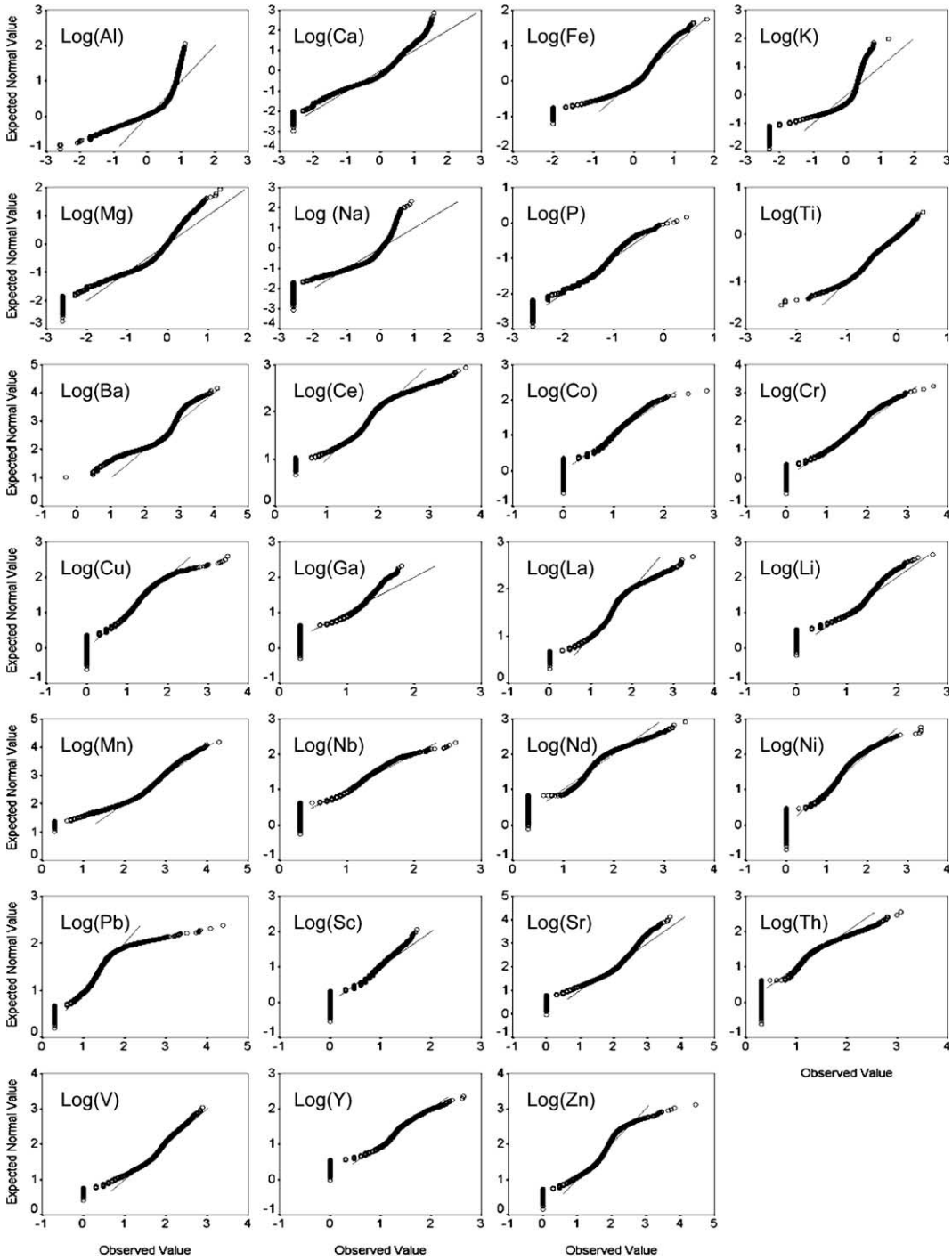

Fig. 6. Normal Q–Q plots for data after logarithmic transformation with the base of 10 ($n = 16{,}511$; units of raw data: Al–Ti in wt%; Ba–Zn in μg/g).

when the sample size reaches 1000, only one experiment, for Ti, passes the K–S test. When the sample size is larger than 2000, none of the elements under study pass the test for lognormality.

The results also show that most of the major elements (Al–Ti) tend toward a normal distribution, whereas trace elements (Ba–Zn) tend to follow a lognormal distribution when the sample size is smaller than 100. When

Table 3
Skewness and kurtosis in the data sets with results of Kolmogorov–Smirnov test

Element	Raw data			Log-transformed data		
	Skewness	Kurtosis	K–S *p*	Skewness	Kurtosis	K–S *p*
Al	−0.13	−0.54	0.00	−2.39	7.65	0.00
Ca	3.93	22.34	0.00	−0.94	0.50	0.00
Fe	5.55	109.07	0.00	−1.84	6.77	0.00
K	0.49	10.39	0.00	−2.58	7.42	0.00
Mg	4.83	66.27	0.00	−1.32	1.80	0.00
Na	1.09	2.11	0.00	−1.54	2.04	0.00
P	13.67	498.08	0.00	−0.89	1.01	0.00
Ti	2.50	10.11	0.00	−0.09	1.69	0.00
Ba	6.37	131.03	0.00	−1.87	4.61	0.00
Ce	14.05	316.36	0.00	0.15	4.80	0.00
Co	31.19	2200.25	0.00	−0.75	0.50	0.00
Cr	31.11	1916.58	0.00	−0.70	0.83	0.00
Cu	41.40	2125.10	0.00	−0.46	1.22	0.00
Ga	0.56	0.91	0.00	−1.01	0.13	0.00
La	15.08	390.02	0.00	0.00	4.82	0.00
Li	4.42	92.40	0.00	−1.28	2.06	0.00
Mn	6.73	137.51	0.00	−1.23	3.60	0.00
Nb	9.00	210.23	0.00	−0.72	0.60	0.00
Nd	15.46	420.92	0.00	−0.95	3.27	0.00
Ni	39.61	2097.99	0.00	−0.81	1.07	0.00
Pb	75.36	6883.48	0.00	0.45	8.55	0.00
Sc	1.82	6.58	0.00	−0.79	0.47	0.00
Sr	4.70	61.70	0.00	−1.19	1.86	0.00
Th	14.28	349.69	0.00	0.03	0.74	0.00
V	2.83	16.84	0.00	−0.99	2.40	0.00
Y	7.22	130.82	0.00	−1.13	3.74	0.00
Zn	92.16	10183.61	0.00	−1.04	3.60	0.00

the sample size is larger than 1000 few samples pass the K–S test for normality or lognormality.

The above results show that sample size heightens the effect of minor deviations from normal or lognormal distribution. To provide a much more lenient measure of the threshold of deviation, a significance level 0.0001 was chosen, and the results are shown in Table 5. Such a low significance level is not generally used in practical applications when a hypothesis is not rejected, but it is used here to demonstrate the effect of sample size when using a statistical significance test. The results are similar to those in Table 4, with only marginally increased opportunities for the data sets to pass the test for normality. When sample size is larger than 2000 no data set except Al passes the test for normality.

The results from Lilliefors correction are generally in line with the K–S test without the correction. The difference is that without Lilliefors correction, the K–S test tends to be conservative, resulting in more datasets passing the test.

The effect of sample size does not apply to the perfectly normal distribution. Almost all the experiments for the normally distributed data set pass the tests within all the sample size ranges. The very small proportion of experiments that does not pass the Lilliefors test is assumed to reflect artifacts of the random re-sampling and testing processes.

4. Discussion

4.1. Skewness and kurtosis

Skewness measures the presence of a tail deviating from normal distribution (0) toward lower values (negative skewness) or higher values (positive skewness). For the raw data sets, only the skewness for Al is negative, and those of K and Ga are close to "0". The skewnesses for all the other elements are positive, many of which are very high, with the maximum value of 75.4 for Pb. This confirms the strongly skewed distributions displayed in the histograms and normal Q–Q plots. Kurtosis refers to "fatness" of the tails in statistical plots, with positive kurtosis denoting peakedness, whereas negative kurtosis denotes a "flat" distribution. The kurtoses for most elements are very high, with the highest value of 6883 for Pb caused by the large number of low value

Table 4
Effect of sample size on results of K–S test: percent of 100 random sub-samples showing normality at $p > 0.05$

n	Raw data							Log-transformed data						
	50	100	200	500	1000	2000	5000	50	100	200	500	1000	2000	5000
Al	100(87)	100(72)	98(70)	89(12)	44(1)	1(0)	0(0)	31(2)	3(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Ca	22(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	83(19)	40(1)	1(0)	0(0)	0(0)	0(0)	0(0)
Fe	86(33)	49(10)	12(0)	0(0)	0(0)	0(0)	0(0)	76(16)	33(6)	5(0)	0(0)	0(0)	0(0)	0(0)
K	97(65)	94(38)	70(5)	6(0)	0(0)	0(0)	0(0)	7(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Mg	80(19)	30(0)	0(0)	0(0)	0(0)	0(0)	0(0)	69(8)	21(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Na	99(44)	83(3)	5(0)	0(0)	0(0)	0(0)	0(0)	39(0)	1(0)	0(0)	0(0)	0(0)	0(0)	0(0)
P	89(39)	66(3)	17(0)	0(0)	0(0)	0(0)	0(0)	91(21)	56(3)	5(0)	0(0)	0(0)	0(0)	0(0)
Ti	52(2)	6(0)	0(0)	0(0)	0(0)	0(0)	0(0)	100(70)	96(51)	96(27)	40(0)	1(0)	0(0)	0(0)
Ba	99(82)	90(57)	74(47)	30(6)	4(0)	0(0)	0(0)	50(7)	7(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Ce	22(5)	2(3)	0(0)	0(0)	0(0)	0(0)	0(0)	91(35)	74(11)	14(1)	0(0)	0(0)	0(0)	0(0)
Co	69(14)	22(1)	0(0)	0(0)	0(0)	0(0)	0(0)	75(10)	11(1)	0(0)	0(0)	0(0)	0(0)	0(0)
Cr	55(7)	9(0)	0(0)	0(0)	0(0)	0(0)	0(0)	99(69)	98(43)	73(8)	0(0)	0(0)	0(0)	0(0)
Cu	56(13)	13(2)	0(0)	0(0)	0(0)	0(0)	0(0)	99(50)	85(9)	24(1)	0(0)	0(0)	0(0)	0(0)
Ga	100(82)	100(62)	93(15)	13(0)	0(0)	0(0)	0(0)	65(4)	10(0)	0(0)	0(0)	0(0)	0(0)	0(0)
La	24(8)	3(2)	0(0)	0(0)	0(0)	0(0)	0(0)	96(28)	74(9)	15(0)	0(0)	0(0)	0(0)	0(0)
Li	90(53)	81(39)	47(12)	2(0)	0(0)	0(0)	0(0)	80(11)	29(1)	0(0)	0(0)	0(0)	0(0)	0(0)
Mn	71(22)	25(4)	2(0)	0(0)	0(0)	0(0)	0(0)	95(47)	71(21)	40(5)	2(0)	0(0)	0(0)	0(0)
Nb	83(43)	51(9)	10(0)	0(0)	0(0)	0(0)	0(0)	88(23)	54(2)	4(0)	0(0)	0(0)	0(0)	0(0)
Nd	30(10)	5(2)	0(0)	0(0)	0(0)	0(0)	0(0)	71(5)	12(1)	0(0)	0(0)	0(0)	0(0)	0(0)
Ni	56(21)	21(2)	0(0)	0(0)	0(0)	0(0)	0(0)	93(22)	53(1)	1(0)	0(0)	0(0)	0(0)	0(0)
Pb	53(23)	20(5)	3(0)	0(0)	0(0)	0(0)	0(0)	96(31)	72(10)	24(0)	0(0)	0(0)	0(0)	0(0)
Sc	78(21)	31(0)	0(0)	0(0)	0(0)	0(0)	0(0)	81(12)	19(0)	1(0)	0(0)	0(0)	0(0)	0(0)
Sr	79(14)	31(1)	1(0)	0(0)	0(0)	0(0)	0(0)	89(23)	45(7)	6(0)	0(0)	0(0)	0(0)	0(0)
Th	16(2)	2(0)	0(0)	0(0)	0(0)	0(0)	0(0)	69(1)	3(0)	0(0)	0(0)	0(0)	0(0)	0(0)
V	77(33)	46(6)	10(0)	0(0)	0(0)	0(0)	0(0)	92(39)	72(14)	25(1)	0(0)	0(0)	0(0)	0(0)
Y	57(17)	13(1)	0(0)	0(0)	0(0)	0(0)	0(0)	79(19)	33(1)	0(0)	0(0)	0(0)	0(0)	0(0)
Zn	77(51)	54(21)	23(3)	0(0)	0(0)	0(0)	0(0)	91(28)	54(7)	12(0)	0(0)	0(0)	0(0)	0(0)
Normal	100(96)	100(94)	100(99)	100(97)	100(94)	100(99)	100(100)	–	–	–	–	–	–	–

Numbers in brackets are calculated after Lilliefors significance correction.

samples compared with the wide data range associated with high outliers. The high positive kurtoses show that very high values in the data are rare, presumably associated with rare processes such as mineralization. Only Al, Na and Ga have fairly small kurtoses, showing that these values are more evenly distributed within the data ranges, and that special concentration of these elements is rare.

Logarithmic transformation tends to move both skewness and kurtosis for most elements towards normality, i.e., closer to 0. However, for the elements Al, K, Na and Ga, the logarithmic transformation moves them further away from normality, i.e., skewness and kurtosis for the transformed data sets diverge further from 0. Another feature of the logarithmic transformation is that most of the transformed data sets have negative skewness values showing that the log-transformation has over-transformed the data. It should be pointed out that the detection limits have some negative effects on these shape parameters.

4.2. Normal Q–Q plots

Probabilistic display delineates discrete populations that have genetic or other affinities as line segments showing changes in slope (i.e., "kinks"). The observed distributions differ significantly from the synthetic distributions modeled by Sinclair (1976) for chemical distributions in rocks. This pioneering author appeared to superimpose different types of distributions having the same concentration range, which yielded curved changes in slope. The present data include elemental distributions with substantial overlap and curvatures. However, notable features are, in fact, the prominent role of line segments showing sharp angular change. Such features require largely discrete chemical subpopulations.

Flat segments on normal Q–Q plots are most prominent for Al, Na, K, Ga and Sc. An approximate normal distribution is expected for major chemical constituents (i.e., Al in aluminosilicates) or elements that are essential components in rock-forming minerals (e.g., K and Na in

Table 5
Effect of sample size on results of K–S test: percent of 100 random sub-samples showing normality at $p > 0.0001$

n	Raw data							Log-transformed data						
	50	100	200	500	1000	2000	5000	50	100	200	500	1000	2000	5000
Al	100(100)	100(98)	100(99)	100(94)	100(60)	97(1)	0(0)	100(31)	85(2)	8(0)	0(0)	0(0)	0(0)	0(0)
Ca	95(22)	40(0)	0(0)	0(0)	0(0)	0(0)	0(0)	100(83)	100(36)	86(1)	1(0)	0(0)	0(0)	0(0)
Fe	99(85)	95(56)	80(9)	6(0)	0(0)	0(0)	0(0)	100(71)	100(34)	72(3)	2(0)	0(0)	0(0)	0(0)
K	100(98)	100(91)	97(67)	93(9)	20(0)	0(0)	0(0)	98(7)	35(0)	0(0)	0(0)	0(0)	0(0)	0(0)
Mg	99(78)	91(41)	59(0)	0(0)	0(0)	0(0)	0(0)	100(71)	100(20)	64(0)	0(0)	0(0)	0(0)	0(0)
Na	100(99)	100(88)	100(4)	1(0)	0(0)	0(0)	0(0)	100(37)	88(2)	14(0)	0(0)	0(0)	0(0)	0(0)
P	99(88)	91(56)	74(18)	13(0)	0(0)	0(0)	0(0)	100(91)	100(42)	96(4)	0(0)	0(0)	0(0)	0(0)
Ti	100(49)	95(9)	11(0)	0(0)	0(0)	0(0)	0(0)	100(100)	100(97)	100(92)	100(43)	95(3)	10(0)	0(0)
Ba	100(99)	96(86)	86(79)	70(36)	33(1)	1(0)	0(0)	100(51)	96(6)	37(0)	0(0)	0(0)	0(0)	0(0)
Ce	75(24)	20(3)	1(0)	0(0)	0(0)	0(0)	0(0)	100(92)	100(68)	99(15)	47(0)	0(0)	0(0)	0(0)
Co	100(66)	95(29)	58(1)	0(0)	0(0)	0(0)	0(0)	100(71)	100(16)	60(0)	0(0)	0(0)	0(0)	0(0)
Cr	98(52)	77(5)	16(0)	0(0)	0(0)	0(0)	0(0)	100(99)	100(93)	100(72)	91(5)	13(0)	0(0)	0(0)
Cu	89(56)	72(18)	24(0)	0(0)	0(0)	0(0)	0(0)	100(98)	100(80)	100(25)	45(0)	0(0)	0(0)	0(0)
Ga	100(100)	100(100)	100(95)	100(17)	67(0)	0(0)	0(0)	100(62)	98(5)	54(0)	0(0)	0(0)	0(0)	0(0)
La	72(22)	19(4)	2(0)	0(0)	0(0)	0(0)	0(0)	100(94)	100(69)	97(16)	33(0)	0(0)	0(0)	0(0)
Li	100(89)	98(77)	91(43)	62(4)	4(0)	0(0)	0(0)	100(73)	100(17)	64(0)	0(0)	0(0)	0(0)	0(0)
Mn	100(70)	92(32)	66(1)	2(0)	0(0)	0(0)	0(0)	100(92)	100(82)	100(42)	58(0)	3(0)	0(0)	0(0)
Nb	98(82)	92(57)	79(11)	6(0)	0(0)	0(0)	0(0)	100(89)	100(43)	96(3)	1(0)	0(0)	0(0)	0(0)
Nd	78(31)	29(10)	5(0)	0(0)	0(0)	0(0)	0(0)	100(67)	100(17)	50(0)	0(0)	0(0)	0(0)	0(0)
Ni	95(55)	72(27)	31(0)	0(0)	0(0)	0(0)	0(0)	100(87)	100(40)	88(3)	1(0)	0(0)	0(0)	0(0)
Pb	79(48)	52(23)	18(3)	0(0)	0(0)	0(0)	0(0)	100(93)	100(72)	99(31)	41(0)	0(0)	0(0)	0(0)
Sc	100(75)	100(27)	65(1)	1(0)	0(0)	0(0)	0(0)	100(76)	100(16)	67(1)	0(0)	0(0)	0(0)	0(0)
Sr	98(80)	94(29)	60(1)	0(0)	0(0)	0(0)	0(0)	100(87)	100(48)	94(8)	5(0)	0(0)	0(0)	0(0)
Th	69(17)	23(2)	0(0)	0(0)	0(0)	0(0)	0(0)	100(72)	99(0)	36(0)	0(0)	0(0)	0(0)	0(0)
V	100(77)	97(46)	78(10)	4(0)	0(0)	0(0)	0(0)	100(91)	100(65)	100(29)	39(0)	0(0)	0(0)	0(0)
Y	95(54)	80(15)	27(0)	0(0)	0(0)	0(0)	0(0)	100(78)	100(37)	73(0)	3(0)	0(0)	0(0)	0(0)
Zn	90(78)	77(54)	50(21)	13(0)	0(0)	0(0)	0(0)	100(89)	100(48)	99(11)	2(0)	0(0)	0(0)	0(0)
Normal	100(100)	100(100)	100(100)	100(100)	100(100)	100(100)	100(100)	–	–	–	–	–	–	–

Numbers in brackets are calculated after Lilliefors significance correction.

feldspars and their weathering products, and Fe in oxides and with Mg in mafic rocks). Straight (normal) segments for Ga (and to a lesser extent, Sc) are explained by the diadochic geochemical behavior of this trace element with Al, e.g., Ga has similar ionic radius and chemical properties as Al (Goldschmidt, 1954 and references cited). Gallium therefore tends to occur in constant ratio with Al, regardless of the mineral phase.

The lognormal Q–Q plots demonstrate that many trace metals tend to approach an S-shape where the lower (left) part is slightly curved, whereas the right part of the curve bends over to form a flatter and often linear slope. If analytical sensitivity is low enough, dispersal in or dilution of an enriched phase by a more common phase or phases can produce a lognormal distribution (flatness) on the left side of log plots. Contaminated sediments enriched in Cu, Zn, Cr, Ag, and other trace metals, and variably diluted with clays and sand in harbours or estuaries demonstrate this kind of relationship (Manheim et al., 1999). Separate enrichment processes, however, may control the right side of the plot. This is displayed by heavy metals like Cu, Zn, Pb, Ag, and rare earths which in this data set are represented by Ce, La, Y and Nd. The slope toward higher concentrations is much steeper than on the left. Between the two discrete populations there will be a curved mixing line. Elements like Calcium, Fe and Ba show "bumpy" or "kinky" irregular curves, consistent with their presence in mixed populations of diverse mineral phases. Calcium, for example, is variably enriched in Ca feldspar in metamorphic rocks, as well as in sedimentary rocks like limestone, dolomite and phosphorite. The extended flat segment of Ca toward the right in the normal Q–Q plots is suggested to be associated with carbonates. This conclusion is supported by examination of an independent NGS data set ("XRF") in USGS (2004), where correlation between high Ca and ignition loss as a surrogate for CO_2 can be observed. Titanium follows the pattern of ferrides (Co, Ni, Cr, V) in higher concentration regions, and of other major constituents in lower concentrations. In a large data set of the current type other mixed populations that have less obvious expression may be present.

Even though multiple kinks can be attributed to multiple populations on these Q–Q plots, it may still be a challenge to properly identify them. One way to do this is to identify samples that belong to a distinct line segment using plotting software that allows identification of individual samples on the slope by brushing with the mouse. Related chemical or other properties of samples associated with the points can then be identified. For example, selecting point data on high-concentration segments on the right side of either the Cu, Pb or Zn curves demonstrates that these points are enriched in all three of these elements, but not in Cr. This pattern suggests sulfide mineralization. Copper, Pb and Zn can be enriched in soils highly contaminated with sewage sludge, but municipal sewage sludge normally contains enhanced Cr as well and does not reach such high metal values. Therefore, the authors suggest sulfide ore mineralization over pollution as the principle source of the metal-enriched samples.

Logarithmic transformation "overtransforms" plots for major and other elements that display flatness in normal Q–Q plots, changing their skewnesses (deviation from the straight line delineating normality) from positive to negative values. Finally, lognormal Q–Q plots delineate detection limits especially sharply in the form of vertical overlapping segments separated from the major body of analyses. Although the present sample set has been reanalyzed to minimize systematic error, another capability of probability plotting, which will not be discussed in this paper, is to suggest both the presence and nature of analytical errors.

4.3. Effect of sample size

The total data set used here would not be expected to show normal distributions because of its heterogeneity as mentioned earlier. However, a pattern of K–S test values signifying significant tendencies toward normality or lognormality emerges rapidly as the number of data points in random re-samplings decreases toward 200 samples or less. At 200 samples the normal sample data set has 4 elements, Al, Ga, Ba, and K, for which 50% of data sets pass the K–S test. At 100 samples, 9 elements pass, and at 50 samples, 22 elements pass. Comparable numbers for the lognormal computations are 2, 12, and 24, respectively.

The large sample size issue may not have received enough attention. Many if not most statistical studies have focused on small sample sizes. However, studies by Cornfield (1966), Morrison and Henkel (1970, and references cited) and Gingerich (1995) suggest that statistical significance level is related to sample size. Generally, at a given significance level (p-value), when the sample size is small, it is hard to reject the null hypothesis; when the sample size is large, it is easy to reject it. Lindley and Scott (1984, p. 3) wrote that "All significance tests are dubious because the interpretation to be placed on the phrase 'significant at 5%' depends on the sample size: it is more indicative of rejection of the falsity of the null hypothesis with a small sample than with a large one". With the increase of sample size, as statistical tests become more powerful, a lower significance level may be chosen. In this study, results of Table 5 were obtained based on a very low significance level of 0.01%. However, even at this level the large data set fails to pass the K–S test for normality as shown in the table.

The effect of sample size on statistical tests provides an explanation for Ahrens' (1954, 1965) lognormal data claims. Ahrens' initial results were based only on about

100 samples. At this small sample size many elements could have shown lognormality even though their parent populations were highly heterogeneous, as demonstrated in this study. Meanwhile, this result also provides an explanation for recent studies demonstrating that geochemical data sets follow neither normal nor lognormal distributions (e.g., Reimann and Filzmoser, 2000), as sample sizes in geochemical data sets get larger. Of course, other factors than sample size, like those related to mixture of populations, existence of outliers, constant detection limits, and two-sided truncation of concentration values (between 0 and 100%), may also play a role.

Differences between real datasets and theoretical distributions always exist. When the sample size is large, such differences can be more easily detected by statistical tests. They lead to rejection of most statistical hypotheses. The hypothesis testing framework is essentially designed for inference from small samples. With large datasets one needs a different mindset: extensive data can be used to explore complex relationships and delineate interesting features. The probability plot approach is recommended for preliminary analysis of such data sets. Meanwhile, probabilistic approaches combined with random sampling such as was utilized in these studies, or employing other partial data selection based on specific criteria, may offer experimental approaches for larger geochemical data sets in the future.

5. Conclusions

Re-sampling of the NGS database ($n = 48{,}544$) of the U.S. Geological Survey yielded 16,511 samples on a 17×17 km grid. This re-sampling removed the effect of clustering of the base data set, providing a more representative data set for the whole area under study. The base data themselves represent a selection of a much larger raw data repository that USGS has subjected to reanalysis by uniform methods in recent years.

Histograms show widely distributed skewness in minor and trace elements in the data set. This skewness is partly due to mixed populations and some highly enriched phases, such as near ore-grade enrichment in Cu, Zn, Pb, Ag, and rare earth elements. High maximum concentrations and lack of association of the latter elements with elements often related to pollutants (e.g., Cr) suggest that these enrichments are due to sulfide mineralization rather than contaminated sediments.

Major chemical constituents of rocks and chemical building blocks of major mineralogical components of rocks (e.g., Na and K in clay minerals) tend to display straight-line segments on normal Q–Q plots. Except for values below detection, Al concentrations conformed most closely to normal, straight-line distribution.

Dispersal and dilution of enriched chemical constituents by phases like quartz sand produces lognormal distributions and straight-line segments on lognormal Q–Q plots. Elements that show flat segments attributed to mineralization processes include Cu, Zn and Pb as sulfides, and rare earth elements, especially Ce. Relatively high detection limits curtailed detailed observations of contaminant metals and elements. Elements whose curves show more complex kinks include Ca, Fe and Ba, attributable to enrichment in multiple mineral phases.

Randomly selected samples ranging in size from 50 to 5000 were taken from the parent population of 16,511 used in this study. At small sample sizes all elements showed test scores indicating either normal or lognormal behavior, whereas at sample sizes exceeding 1000 few elements passed the K–S and Lilliefors tests for statistical normality at the 0.05 significance level. Statistical tests are designed for relatively small sample sizes. With large sample sizes alternative graphical methods are recommended.

The probability plots permitted useful characterization of outlier populations, allowing visual comparison among elemental groups, as well as quick assessment of quantitative relationships for element-rock associations. The combination of concentration and cumulative frequency in the form of normal Q–Q plots offer useful features for displaying large data sets encompassing complex mixtures and a wide range of concentrations. Extremes, both on the low concentration side, as well as on the high side, as in heavy metals, are well displayed. Outlier subpopulations and their genetic affinities can be identified, and analytical factors and limitations like detection limits are conveniently observable without distorting the display.

Acknowledgments

The authors are grateful for valuable comments and suggestions provided by reviewers Gilpin R. Robinson, Bob Garrett, Nils Gustavsson, Suzanne Nicholson, Larry Drew and one anonymous reviewer. Helpful discussions with Jerome Seahan, John Newell, Ruben Roa Ureta, Jeffrey D. Blume, Isobel Clark, Ned Levine, Thies Dose, Chris Hlavka and Donald Myers concerning effect of sample size are acknowledged. This study is partly supported by the International Collaboration Program from Enterprise Ireland (No. IC/2002/026).

References

Ahrens, L.H., 1954. The lognormal distribution of the elements (a fundamental law of geochemistry and its subsidiary). Geochim. Cosmochim. Acta 5, 49–73.

Ahrens, L.H., 1965. Distribution of the Elements in Our Planet. McGraw-Hill, New York.

Aubrey, K.V., 1956. Frequency distributions of elements in igneous rocks. Geochim. Cosmochim. Acta 9, 83–89.

Blom, G., 1958. Statistical Estimates and Transformed Beta Variables. Wiley, New York.

Bölviken, B., Stokke, P.R., Feder, J., Jøssang, T., 1992. The fractal nature of geochemical landscapes (a fundamental law of geochemistry and its subsidiary). J. Geochem. Explor. 43, 91–109.

Chakravarti, I.M., Laha, R.G., Roy, J., 1967. Handbook of Methods of Applied Statistics, vol. I. Wiley, New York, pp. 392–394.

Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A., 1983. Graphical Methods for Data Analysis. Chapman & Hall, New York.

Cheng, Q., Agterberg, F.P., Ballantyne, S.B., 1994. The separation of geochemical anomalies from background by fractal methods. J. Geochem. Explor. 51, 109–130.

Clarke, F.W., 1924. The Data of Geochemistry, fifth ed. U.S. Geol. Surv. Bull. 770.

Cornfield, J., 1966. Sequential trials, sequential analysis and the likelihood principle. Am. Statistic. 20, 18–23.

Darnley, A.G., Björklund, A., Bølviken, B., Gustavsson, N., Koval, P.V., Plant, J.A., Steenfelt, A., Tauchid, M., Xie X.J., Garrett, R.G., Hall, G.E.M., 1995. A global geochemical database for environmental and resource management: recommendations for International Geochemical Mapping. Final Report of IGCP Project 259, second revised edition. UNESCO, Paris.

Darnley, A.G., 1997. A global geochemical reference network: the foundation for geochemical baselines. J. Geochem. Explor. 60, 1–5.

Gingerich, P.D., 1995. Statistical power of EDF tests of normality and the sample size required to distinguish geometric-normal (lognormal) from arithmetic-normal distributions of low variability. J. Theor. Biol. 173, 125–136.

Goldschmidt, V.M., 1954. Geochemistry. Oxford University Press, London.

Helsel, D.R., Hirsch, R.M., 1992. Statistical methods in water resources. Studies in Environmental Science, no. 49. New York, Elsevier Publishers, Inc. Available from: <http://water.usgs.gov/pubs/twri/twri4a3/> (accessed 14.08.03).

Krige, D.G., 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. J. Chem. Metall. Min. Soc. S. Afr. 52, 119–139.

Krige, D.G., 1960. On the departure of ore value distributions from lognormal models in South African gold mines. J. S. Afr. Inst. Mining Metall. 61, 231–244.

Krige, D.G., 1966. A study of gold and uranium distribution patterns in the Klerksdorp gold field. Geoexploration 4 (1), 43–53.

Lilliefors, H.W., 1967. On the Kolmogorov–Smirnov test for normality with mean and variance unknown. J. Am. Statist. Assoc. 64, 399–402.

Lindley, D.V., Scott, W.F., 1984. New Cambridge Elementary Statistical Tables. Cambridge University Press, Cambridge.

Mandelbrot, B., 1982. The Fractal Geometry of Nature. Freeman, New York.

Manheim, F.T., Buchholtz ten Brink, M., Hastings, M.E., Mecray, E., 1999. Contaminated-sediment database development and assessment in Boston harbor. U.S. Geol. Surv. Fact Sheet 78-99. Available from: <http://pubs.usgs.gov/fs/fs78-99/> (accessed 08.12.04).

Manku, G.S., Rajagopalan, S., Lindsay, B.G., 1999. Random sampling techniques for space efficient online computation of order statistics of large datasets. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1999, pp. 251–262.

Miesch, A.T., Riley, L.B., 1961. Basic statistical methods used in geochemical investigation of Colorado Plateau uranium deposits. HIMMP Trans. (Mining) 220, 247–251.

Morrison, D.E., Henkel, R.E. (Eds.), 1970. The Significance Test Controversy – A Reader. Butterworths, London.

Newcomb, W.D., Rimstidt, J.D., 2002. Trace element distribution in US groundwaters: a probabilistic assessment using public domain data. Appl. Geochem. 17, 49–57.

Plant, J., Smith, D., Smith, B., Williams, L., 2001. Environmental geochemistry at the global scale. Appl. Geochem. 16, 1291–1308.

Razumovsky, N.K., 1940. Distribution of metal values in ore deposits (in Russian). Dokl. Akad. Nauk, SSSR, 814–816.

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

Sichel, H.S., 1952. New methods in the statistical evaluation of mine sampling data. London Inst. Mining Metall. Trans. 61, 261–288.

Sinclair, A.J., 1974. Selection of threshold values in geochemical data using probability graphs. J. Geochem. Explor. 3 (2), 129–149.

Sinclair, A.J., 1976. Application of probability plots in mineral exploration. Assoc. Explor. Geochem. Spec. vol. 4, 95.

Stephens, M.A., 1974. EDF statistics for goodness of fit and some comparisons. J. Am. Statist. Assoc. 69, 730–737.

U.S. Environmental Protection Agency, 2002. STORET, U.S. Environmental Protection Agency. Office of Water. Available from: <http://www.epa.gov/storet/> (last accessed 08.12.04).

USGS, 2004. The National Geochemical Survey – Database and documentation. U.S. Geological Survey Open-File Report 2004-1001. Available from: <http://tin.er.usgs.gov/geochem/doc/home.htm> (last accessed 08.12.04).

Vinogradov, A.P., 1959. Geochemistry of Rare and Dispersed Chemical Elements in Soil. Acad. Sciences, USSR, Moscow-Leningrad (Trans by Consultants Bureau, Chapman & Hall, London).

Vistelius, A.B., 1960. The skew frequency distributions and the fundamental law of the geochemical processes. J. Geol. 68, 1–22.

Zhang, C.S., Selinus, O., 1998. Statistics and GIS in environmental geochemistry – some problems and solutions. J. Geochem. Explor. 64, 339–354.