Available online at www.sciencedirect.com
ScienceDirect
Journal of Geochemical Exploration 93 (2007) 135–152
JOURNAL OF GEOCHEMICAL EXPLORATION
www.elsevier.com/locate/jgeoexp

Using GIS and statistics to study influences of geology on probability features of surface soil geochemistry in Northern Ireland

Crawford Jordan [a,1], Chaosheng Zhang [b,*], Alex Higgins [a]

[a] *Agriculture, Food and Environmental Science Division, Agri-Food and Biosciences Institute, Newforge Lane, Belfast BT9 5PX, Northern Ireland, United Kingdom*
[b] *Department of Geography, National University of Ireland, Galway, Ireland*

Received 4 July 2006; accepted 3 March 2007
Available online 13 March 2007

Abstract

The probability features of non-normality and non-lognormality are widely observed in geochemistry due to the influences of multiple factors that are difficult to quantify and model. In Northern Ireland, the pseudo-total concentrations of 14 elements (Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, P, Pb and Zn) from 6138 topsoils were measured, and GIS mapping showed that the spatial distribution of these data were in line with the spatial distribution of geology in the area. Investigations into the influences of geology on the concentration data and their probability features were carried out using GIS and statistics in this study. The whole raw data sets for each element were positively skewed and none of them followed either normal or lognormal distributions. Logarithmic transformation was found to have "over-transformed" most of the data sets, changing their skewness from positive to negative values. When soil samples were classified by rock type using a GIS overlay function, obvious differences were observed in the chemical concentrations of soils derived from different rock types. Soils in basalt areas displayed the highest concentrations for most elements under study (Ca, Co, Cr, Cu, Fe, Mg, Mn, Na, Ni, P and Zn) but the lowest concentrations for K, while the highest levels for Cd and Pb occurred in the shale areas. Classifying soils by rock type produced more normally distributed data sets, especially for the igneous rock areas. To restrain the influence of soil type and land cover, samples from both gleys and pastures were extracted via a GIS and it was found the data sets then showed generally greater tendencies towards normality. However, many of the data sets would still not pass a test for normality unless the sample size was small (e.g. of the order of a couple of hundreds). Geology, soil type, land cover and sample size all played important roles in determining soil chemical concentrations and their probability features. However, the influences from other factors were still evident. Attempts made in this study show that it remains a challenging task in geochemistry to separate all the factors and to model their influence at the regional scale.

Keywords: Soil geochemistry; Geology; Probability; Statistics; GIS; Northern Ireland

1. Introduction

Soil geochemistry is controlled by multiple factors with solid geology regarded as a major one. Due to the complicated effects of these factors, non-normal and non-lognormal distributions are widely observed in geochemical databases (Reimann and Filzmoser, 2000). In the

* Corresponding author. Fax: +353 91 525700.
E-mail addresses: Crawford.Jordan@afbini.gov.uk (C. Jordan), Chaosheng.Zhang@nuigalway.ie (C. Zhang).
[1] Fax: +44 2890 255005.

0375-6742/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gexplo.2007.03.001

1950s, Ahrens (1954) stated the lognormal law of geochemistry which was based on trace element concentrations in igneous minerals and rocks. However, geochemists soon presented objections to the universality of this law (Aubrey, 1956; Vistelius, 1960). Ever since, studies on the probability distributions of geochemical data have been on-going. The distributions of the untransformed data were found to be positively skewed for many elements (Davies, 1980; McBratney et al., 1982). In common with the findings of McGrath and Loveland (1992) in their analysis of over 5000 representative topsoils for England and Wales, a logarithmic (base 10) transformation also failed to produce normally distributed data for any of the elements studied. Zhang and Selinus (1998) regarded the lognormal distribution as a special case of more general, positively skewed distributions. Furthermore, Zhang et al. (2005) found none of the 27 elements in a large geochemical database from the U.S. Geological Survey followed either normal or lognormal distributions. In addition to the widely known factors such as multiple populations (caused by geology, soil type, etc.), detection limits and outliers, large sample size was observed to be an independent factor affecting the results of statistical tests (Zhang et al., 2005).

In "The Soil Geochemical Atlas of Northern Ireland" (Jordan et al., 2000), the pseudo-total and extractable concentrations of sixteen elements (cadmium, calcium, chromium, cobalt, copper, iron, lead, magnesium, manganese, molybdenum, nickel, phosphorus, potassium, sodium, sulphur and zinc) in the agriculturally-important soils of Northern Ireland were measured and mapped. These elements were selected either because they were essential for healthy development of micro-organisms, plants and animals, or because they could be important contaminants. The GIS maps of the data sets showed a consistency between soil chemical composition and the spatial distribution of rock types in the study area (Jordan et al., 2000). A recent study by the authors demonstrated influences of rock type on the spatial variations of Ni concentrations in soils of the study area using neighbourhood statistics (Zhang et al., 2007). It is, therefore, essential to separate soil samples by rock type in order to more fully understand the influence of geology on the chemical concentrations of the different groups of soils. It is also important to evaluate if the probability features show better tendencies towards normality when the geology factor is under control. Even though it is widely known that soil geochemistry is affected by multiple complicating factors, attempts to separate or restrain these factors should be of wide interest. GIS techniques make such analyses possible.

In this study, the influence of geology on regional soil geochemistry in Northern Ireland, as defined by the pseudo-total concentrations of 14 selected elements, was examined. The probability features associated with soil geochemistry were further investigated when the effect of geology, as well as soil type and land cover, was constrained.

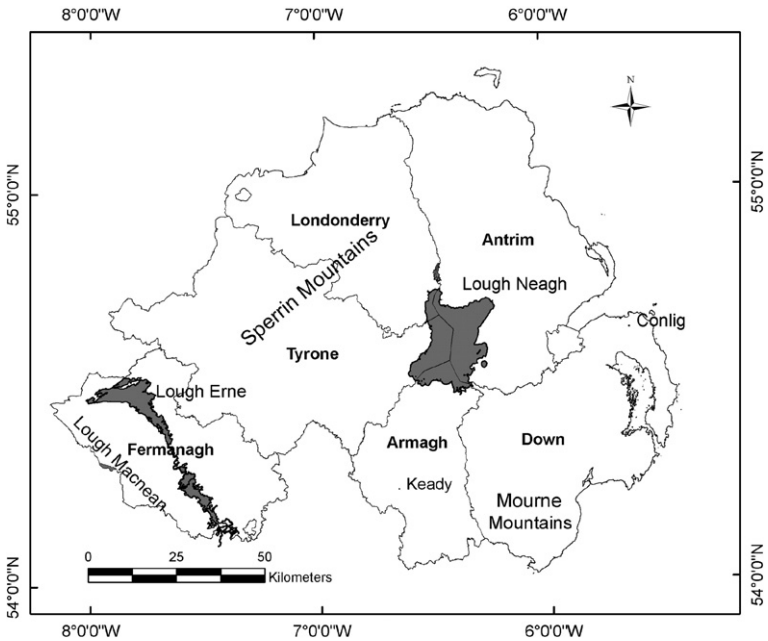

Fig. 1. Study area.

2. Methods

2.1. Study area

Northern Ireland is comprised of six counties viz. (in clockwise order from the north position) Antrim, Down, Armagh, Fermanagh, Tyrone and Londonderry. The Province occupies a total area of 14,120 km^2 of which 13,480 km^2 is land and 640 km^2 is ascribed to inland waters, predominantly lakes Neagh and Erne (Fig. 1). Over 17% of the land area is above 200 m with nearly 6% above 300 m and a maximum altitude of 850 m in the south-east. The Province is a microcosm of the Earth's geology with nearly every period of the Earth's geological history being represented and almost every known rock type found there. The major soil subgroups are climatic peat above 200 m (14%), with acid brown-earths (13%) and gleys (56%) at lower elevations. The dominant soil parent materials are drifts and glacial tills derived from basalts (County Antrim), Silurian shales (counties Down and Armagh), mica-schist (counties Tyrone and Londonderry) and carboniferous-age rocks (counties Fermanagh and Armagh). Significant areas of granite are found in the south-east of the Province (the Mourne mountains). These major solid geological features are shown in a simplified geological map of the Province (Fig. 2) which is based on the Geological Survey of Northern Ireland's 1:250,000 solid geology map (GSNI, 1998).

Historically, iron ore, coal, lead and salt were the dominant minerals mined in Northern Ireland. There are now over 2000 abandoned mine workings across the Province, most of which date from the 18th to early 20th century. However, in recent years, lignite, gold and industrial minerals have dominated commercial exploration activity in Northern Ireland. Full details of the extent and distribution of mineralization in Northern Ireland can be found in Mitchell (2004).

Galena, sphalerite, pyrite and chalcopyrite are the dominant minerals found in vertical veins at a number of locations in the south of County Armagh and in north County Down centred on the towns of Keady and Conlig, respectively. Small amounts of copper, as malachite, are also found in these veins. Significant quantities of zinc and lead are also found associated with gold and silver mineralization in the Dalradian schists of the Sperrin mountains in the north-west of the Province.

Other significant sources of minerals are found throughout Northern Ireland. For example, iron ore was extensively worked in the basalt areas of County Antrim up to the 1920's. Haematite is found at the northern boundary of the shale/granite interface in County Down and a significant magnesium limestone deposit (with Mg > 14%)

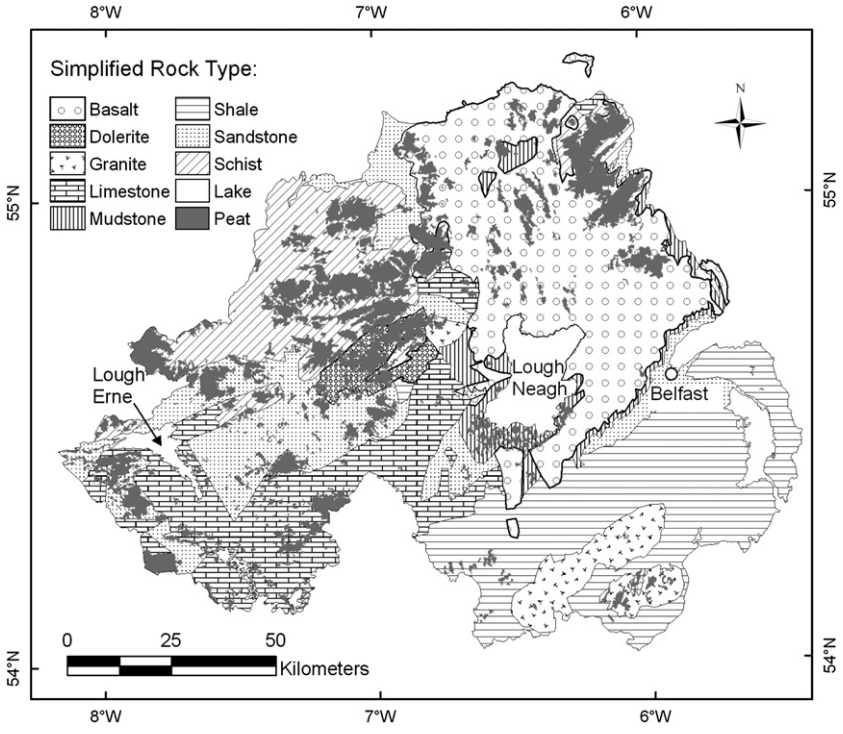

Fig. 2. A simplified geological map of Northern Ireland showing the major solid geologies, lakes and peat (original GIS file from GSNI).

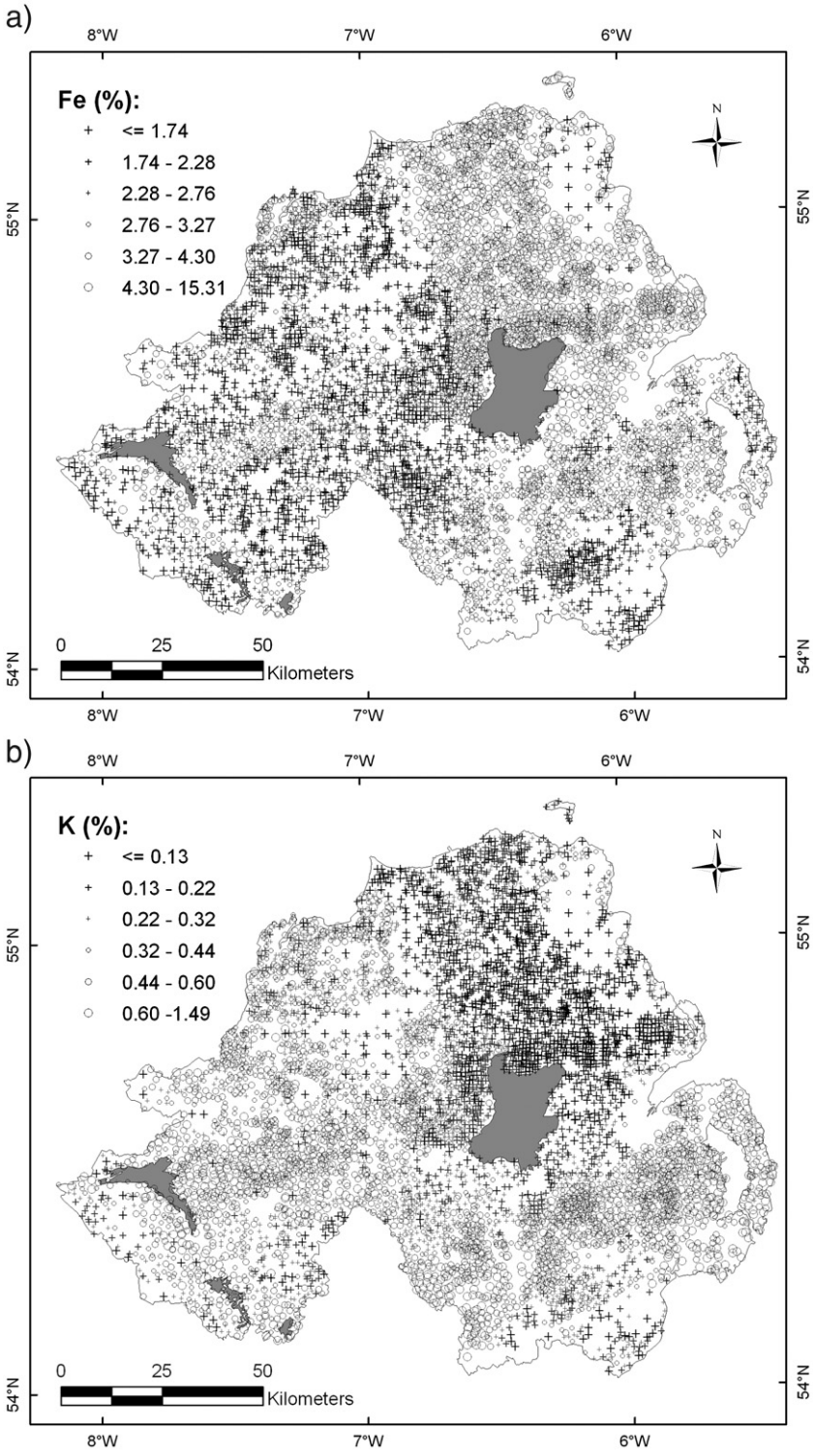

Fig. 3. Symbol maps of Fe and K concentrations in the soils of Northern Ireland: a) Fe and b) K (the symbols are designed to highlight both high and low values).

is quarried just north of Lough Macnean, close to the County Fermanagh border with the Republic of Ireland.

Apart from past mining activity, little evidence of cultural contamination has been proven but may be suspected around industrial centres and particularly in localities where copper sulphate has been used extensively for agricultural purposes (Webb, 1973; Jordan et al., 2000). In modern times, the rapid increase in vehicular transport accounts for higher emission and deposition levels of lead (from fuel additives mainly during the 20th century) and zinc (from tyre wear) in urban areas and along the major road network. Similarly, emission and deposition of cadmium and copper particles from coal and fuel oil combustion are likely to be centred around urban areas (Dore et al., 2005).

The Province has a long-term (1961–90) mean daily air temperature of 8.7 °C, a mean annual rainfall of 1113 mm and a mean annual potential evaporation loss of 384 mm (Betts, 1997). Agricultural land accounts for 80% of the total area of Northern Ireland and this is dominated by grassland (72%) and rough grazing (18%) with only 6% under crops and 1% in broad-leaved woodland (Tomlinson, 1997).

2.2. Soil sampling and preparation

2.2.1. Soil sampling

Soil sampling was carried out in almost every 1 × 1 km square in the lowland area of Northern Ireland (below an elevation of 200 m) and resulted in the collection of 6038 soil samples. A further 100 samples were collected from the upland/semi-natural regions of the Province, at a density of approximately 1 sample per 25 km^2 (see Fig. 3 for sampling locations). Water bodies and urban areas were excluded from this survey. Within each sampled 1 km square, the sample point was located in the dominant soil type indicated by the distribution on the 1:10,000 soil survey field sheet (Cruickshank, 1997). The soil samples were taken in the topsoil within the depth zone 0 to 25 cm (the vast majority of soils were sampled from the ploughed horizon A_p but, for peat soils, the O horizon was sampled). A 1 kg composite soil sample, averaged over the depth of the sampled horizon at each inspection pit, was taken in the field. All the samples were subsequently returned to the laboratory, air-dried, milled and sieved at 2 mm, sub-sampled down to 300 g and stored. Each sample was geo-located to the nearest 100 m using a 6-figure Irish Grid reference (OSI, 1953).

2.2.2. Sample preparation

A 3–4 g homogenised sub-sample was taken from the sieved, sub 2 mm soil sample. This was milled in zirconium oxide grinding equipment to a fine powder and transferred to polyethylene containers prior to digestion. Approximately 2 g of soil (±0.0001 g) were weighed into porcelain crucibles. The samples were oven-dried (105 °C) overnight, cooled in a desiccator and re-weighed. The crucibles and their contents were then placed in a muffle furnace (450 °C) for 12 hours and allowed to cool in a desiccator prior to being quantitatively transferred into digestion tubes.

For the pseudo-total concentrations, samples were digested in block digesters with 15 ml of 50% HCl and 5 ml HNO_3 (*aqua regia*). All acids used were "Aristar" grade. The samples were mixed on a vortex mixer, heated for 3 hours at 60 °C, 1 hour at 105 °C and at 140 °C for 10 hours, until dryness. Twenty-five millilitres of 20% HCl were added to each digestion tube, the contents mixed using a vortex mixer and then heated for 40 min at 80 °C. After cooling, the solutions were again mixed on a vortex mixer and filtered overnight, using Whatman No. 542 filters, into 100 ml volumetric flasks and made up to the mark with ultra-pure water. All filter papers were pre-washed with 0.05 M ethylene diamine tetra-acetic acid (EDTA) and rinsed with ultra-pure water before use.

The samples were digested and analysed in sets of 40. Each set of 40 consisted of 28 samples, 2 samples repeated from the previous set, 2 samples repeated from the current set, 4 internal reference soil samples and 4 blanks. All glassware was cleaned by rinsing with de-ionised water, soaking in a 10% "Decon" solution overnight, rinsing with de-ionised water, soaking in a 10% HNO_3 solution for a minimum of 2 hours and finally rinsing with ultra-pure water and allowing to dry.

2.3. Chemical analysis and quality control

2.3.1. Determination of pseudo-total concentrations

Analysis of pseudo-total Ca, Cd, Co, Cr, Cu, Fe, K, Mg, Mn, Na, Ni, P, Pb and Zn was carried out using an IRIS Inductively Coupled Plasma — Atomic Emission Spectrometer (ICP-AES) from Thermo Jarrell Ash. All concentration values were adjusted for soil moisture content and the results expressed in mg/kg oven-dry soil for trace elements and in % for major elements.

2.3.2. Quality control

All measurements were made in a National Accreditation of Measurement and Sampling (NAMAS, now The United Kingdom Accreditation Service (UKAS)) accredited laboratory. Records of sample weights and ICP-AES analysis data were stored on a central computer system. Microsoft Excel macro programmes were written

Table 1
Percentiles for pseudo-total element concentrations in the soils of Northern Ireland ($n=6138$; for Cu $n=6137$; units in mg/kg except major elements (Ca, Fe, K, Mg, Na, and P) in %); also included are detection limits and mean and coefficient of variation (CV) of the reference soil sample analyses used in the quality control (QC) procedures

Element	Minimum	5%	10%	25% (Q1)	Median (Q2)	75% (Q3)	90%	95%	Maximum	Number of low value outliers [a]	Number of high value outliers [a]	Detection limits (mg/kg)	Mean of QC reference soil sample [b] (mg/kg)	CV of QC reference sample analyses [b] (%)
Ca	0.0002	0.12	0.16	0.24	0.36	0.59	0.85	1.03	15.21	1	54	1.07	3189.8	5.4
Cd	<0.08	<0.08	<0.08	0.2	0.33	0.5	0.72	0.91	5.49	0	87	0.08	0.3	39.2
Co	<0.17	2.72	4.02	6.49	10.53	20.67	35.67	40.31	110.71	6	4	0.17	6.6	5
Cr	<0.17	15.16	19.16	28.71	46.51	68.66	103.33	125.8	654.42	1	34	0.17	22.7	9.5
Cu	<1.07	6.36	10.13	17.26	27.1	44.47	72.37	88.99	278.54	16	43	1.07	15.6	6.1
Fe	<0.001	1.01	1.4	2.01	2.76	3.62	5.49	6.61	15.31	1	28	9.64	22383.8	7.6
K	<0.0003	0.07	0.09	0.17	0.32	0.51	0.71	0.82	1.49	1	0	2.79	3376.2	7.9
Mg	<0.0002	0.15	0.21	0.34	0.59	0.91	1.37	1.66	5.23	1	20	2.05	3388.6	4.7
Mn	<0.03	81	132	249	452	762	1097	1289	22879	4	42	0.03	450.7	4.5
Na	<0.0004	0.013	0.015	0.02	0.027	0.04	0.062	0.081	1.491	1	150	0.37	262.5	10.7
Ni	<0.32	6.43	8.69	15	29.16	60.36	110.94	137.01	545.24	2	32	0.32	12.2	6
P	<0.0004	0.044	0.054	0.071	0.092	0.119	0.154	0.181	0.459	2	40	4.25	988.1	5.3
Pb	<2.62	6.73	9.15	12.83	17.92	26.81	41.65	55.32	1000.61	130	156	2.62	36.3	6.5
Zn	<0.07	21.18	29.09	43.67	65.38	88.81	110.25	125.31	590.56	1	28	0.07	55.7	3.4

[a] Low value outliers were identified by sorting the raw data and by reference to the normal Q-Q plots for the logarithmically transformed data; High value outliers were defined as higher than the third quartile (Q3) plus 3 times the inter-quartile range (IQR = Q3 − Q1).
[b] The number of reference soil samples analysed varied between 705 and 713.

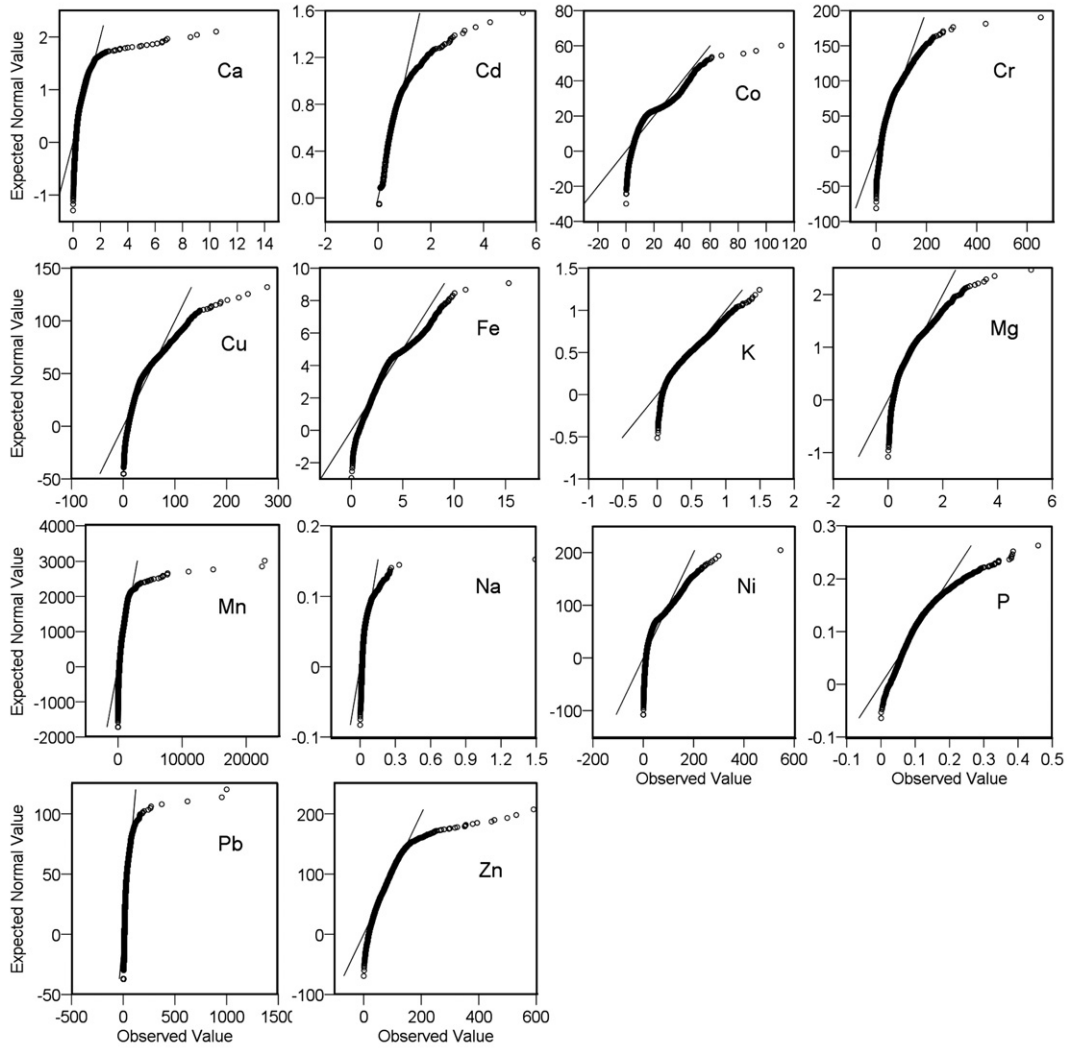

Fig. 4. Normal Q-Q plots for raw concentration data (n=6138; units in mg/kg except major elements (Ca, Fe, K, Mg, Na, P) in %).

to evaluate the final concentration of elements in the samples and highlight any breaches of the quality control (QC) procedures. Each set of digestions had 4 internal reference soils that had been characterised against certified reference materials (e.g. CRM 143R "Sewage sludge amended soil" from the Community Bureau of References, BCR). Each of the elements under analysis were assigned acceptable upper and lower "warning" and upper and lower "critical" concentration levels, based on 2 and 3 standard deviations from the mean ascertained from the data used to characterise the soil. Synthetic QC solutions were included in the analytical sample stream every 20 samples. Batches of samples were re-analysed if the QC soil results met any of the following conditions: were above the critical limits, two or more consecutive QC results were above the warning limits or the standard checks showed the instrument was out of calibration.

As over 700 replicate measurements (between 705 and 713) of the internal reference soil were made for each element, the coefficient of variation (CV) of the measurements about the mean were calculated to give a measure of the precision with which each metal determination was made (see Table 1).

2.4. Database and statistical analyses

The pseudo-total concentrations of the 14 elements in each soil sample were stored in a Microsoft Excel spreadsheet and merged, on the basis of laboratory sample number, with a separate spreadsheet containing

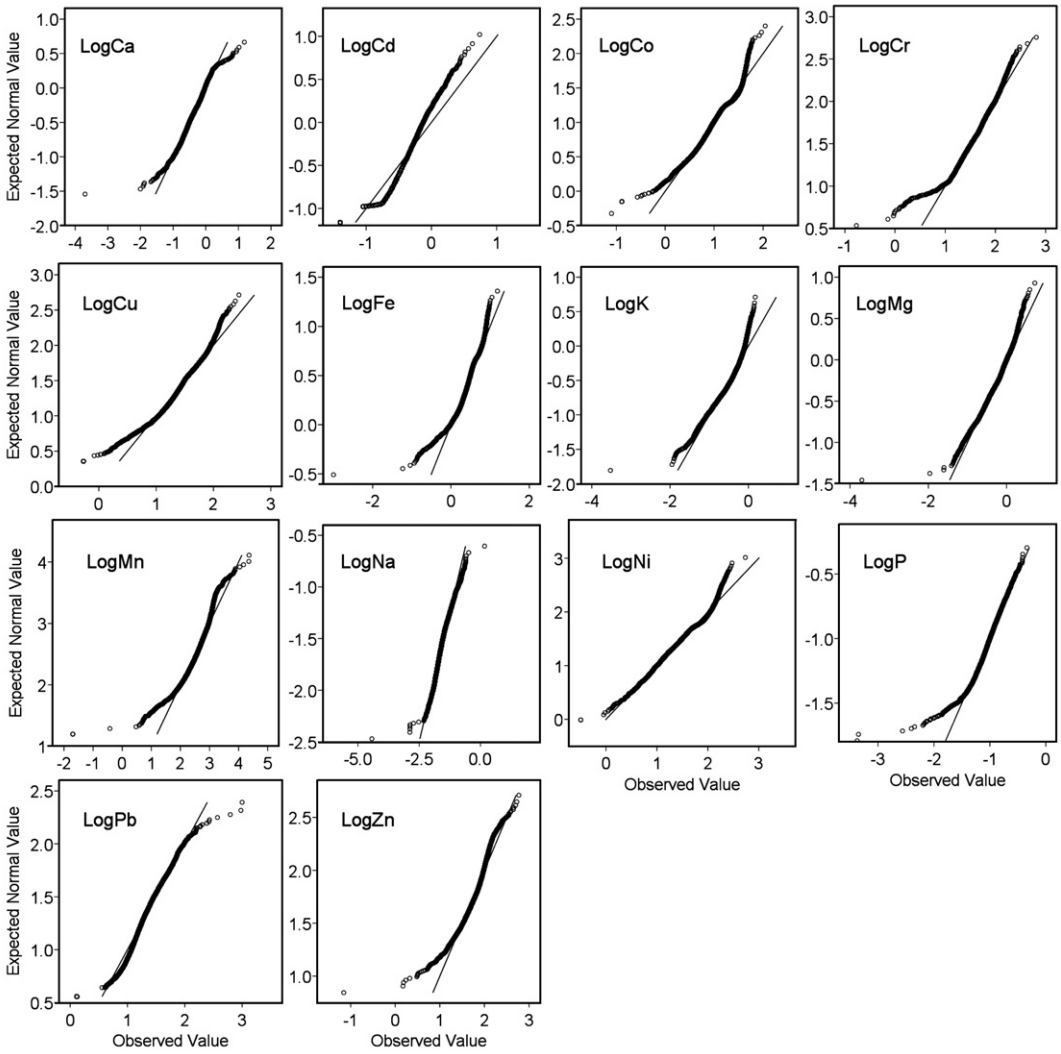

Fig. 5. Normal Q-Q plots for concentration data after logarithmic transformation to the base 10 (n=6138; units for raw data in mg/kg except major elements (Ca, Fe, K, Mg, Na, P) in %).

matching soil attribute data (Cruickshank, 1997). In this way, each sample record acquired locational information (an Irish Grid reference) together with a summary soil description. Moreover, before numerical analysis of the data, those samples with concentrations below the detection limit of the method were assigned a value equal to half the detection limit for the element(s) in question. Only Cd had significant numbers of samples with concentrations at or below the detection limit (1085 or 18% of all samples).

The GIS layers used in this study included geology, soil type and peat. The geology map was based on the digital version of the 1:250,000 "Geology Map of Northern Ireland" (GSNI, 1998) while the peat map was based on a combination of the 1:250,000 soil map of Northern Ireland and the CORINE land cover map of Northern Ireland (Cruickshank and Tomlinson, 1996; Tomlinson, 1997).

In this study, descriptive statistical analyses were carried out to identify the overall features of the data sets, which included percentiles, median, skewness and kurtosis. A statistical test by Kolmogorov–Smirnov (K-S) was applied to test the normality of data sets. Normal quantile–quantile (Q-Q) plots were produced to illustrate the probability features of both the raw and logarithmically transformed data sets. GIS was applied to classify soil samples by rock type using an overlay function. Median values were then

Table 2
Skewness and kurtosis values and significance levels of the Kolmogorov–Smirnov test for normality for the whole data set (n=6138; outliers excluded)

Element	Raw data			Log-transformed data		
	Skewness	Kurtosis	K-S p [a]	Skewness	Kurtosis	K-S p
Ca	1.31	1.80	0	−0.37	0.65	0.01
Cd	1.00	1.36	0	−0.92	−0.22	0
Co	1.18	0.35	0	−0.29	0.09	0
Cr	1.20	1.33	0	−0.79	2.36	0
Cu	1.33	1.46	0	−0.60	0.85	0
Fe	1.08	1.08	0	−1.01	3.23	0
K	0.81	0.24	0	−0.67	0.11	0
Mg	1.15	1.12	0	−0.50	0.21	0
Mn	1.01	0.83	0	−1.16	2.49	0
Na	1.44	1.97	0	0.00	0.72	0
Ni	1.41	1.21	0	−0.14	−0.42	0
P	0.93	1.22	0	−0.92	3.60	0
Pb	1.44	2.00	0	0.06	−0.10	0
Zn	0.70	0.93	0	−1.08	2.49	0

[a] "0" values represent "<0.001".

calculated for each rock type. All of these functions are available in popular statistical software packages, and thus they are not explained in this paper. Statistical analyses were mainly carried out using SPSS® (V.14) and the GIS software used was ArcGIS® (V.9.1).

3. Results and discussions

3.1. GIS visualization of soil geochemistry

Initial GIS mapping showed there was a good association between soil chemical composition and the spatial distribution of rock type. The maps for Fe and K were selected as examples (Fig. 3).

The spatial locations of element concentrations clearly reflected rock type boundaries. High concentrations of Fe (circle symbols in Fig. 3) were found to be predominantly located in the basalt area in the north-eastern part of Northern Ireland, while low concentrations of Fe (+ symbols in Fig. 3) were found in the granite, limestone, and schist areas to the south and west of the Province. Shale and sandstone areas had intermediate Fe concentrations. However, K showed a very different spatial distribution to Fe with high concentrations of K (circle symbols in Fig. 3) located in the shale and sandstone areas in the south and south-west, while the lowest K concentrations (+ symbols in Fig. 3) were found in the basalt area in the north-east of the Province. Other areas have intermediate concentrations of K.

3.2. Normal Q-Q plots for all data sets

To investigate the probability features of the data sets, normal Q-Q plots were produced for both raw data (Fig. 4) and log-transformed data (Fig. 5).

All the whole raw data sets for the elements studied showed significant deviation from a normal distribution, with their samples located away from the diagonal lines in Fig. 4. All the raw data sets displayed a convex shape, with both the high value and low value ends falling below the diagonal lines. Fig. 4 shows that if the raw data were normally distributed, their low values should be lower, and their high values should also be lower. Due to the non-negative constraint of elemental concentrations, it is impossible to find any values below "0". On the other hand, the abnormally high values

Table 3
Median values for soils by dominant rock type and peat (in mg/kg except major elements (Ca, Fe, K, Mg, Na, and P) in %)

Element	Basalt	Dolerite	Granite	Limestone	Mudstone	Shale	Sandstone	Schist	Peat
n [a]	1591	102	199	832	249	1225	855	789	296
Ca	0.68	0.32	0.28	0.30	0.40	0.30	0.27	0.29	0.29
Cd	0.29	0.23	0.22	0.41	0.27	0.41	0.31	0.31	0.27
Co	31.98	7.40	7.45	6.72	12.79	12.01	7.32	7.10	5.38
Cr	86.64	33.42	27.11	32.36	55.22	54.07	37.36	22.68	28.05
Cu	60.25	22.36	21.76	16.89	31.92	27.14	17.89	22.18	15.16
Fe	4.67	2.32	1.88	2.05	2.73	2.93	2.21	2.32	2.02
K	0.12	0.41	0.27	0.36	0.23	0.60	0.41	0.37	0.25
Mg	1.12	0.43	0.42	0.32	0.45	0.74	0.45	0.37	0.28
Mn	906	306	330	285	500	469	289	336	196
Na	0.050	0.025	0.022	0.019	0.027	0.026	0.024	0.022	0.024
Ni	95.72	14.19	15.69	18.38	35.51	35.61	20.57	12.29	12.09
P	0.110	0.079	0.087	0.081	0.081	0.098	0.083	0.088	0.072
Pb	16.09	13.36	18.75	16.74	15.93	27.20	14.93	17.10	16.71
Zn	90.30	42.08	50.59	50.71	58.18	78.07	45.93	46.87	36.72

[a] One value is missing for Cu in peat.

could be explained as the result of rare processes, such as mineralization or human pollution. The high value outliers were easily identified on the normal Q-Q plots, as they were located far away from the majority of the samples. Several elements showed only one or two extremely high values e.g., Fe, Mg, Na and Ni. These are likely to result from errors in the database (e.g. from typing or analytical errors), or the samples may have been contaminated. Another feature of the normal Q-Q plots was that there were multiple kinks (changes in slope) for elements like Co, Cu, Fe, K, Ni and Zn. These kinks are caused by a mixture of multiple populations within the data, showing that there are several groups of samples within the whole data set due, for example, to the presence of groups of different rock type. If the multiple populations are well mixed, the multiple-kink feature may not be significant as was observed for elements other than Co, Cu, Fe, K, Ni and Zn.

For most elements, the normal Q-Q plots for the log-transformed data set (Fig. 5) showed that the data were now more normally distributed than before transformation. However, non-normal behaviour was still found on the low value and high value ends and the multiple-kink feature was still obvious for several elements. One of the major features observed for the log-transformed data set for all elements (Fig. 5) was that the high value outliers observed in Fig. 4 were pushed towards the majority of samples while the reverse occurred with the low value outliers from Fig. 4. The isolated, extremely low values in Fig. 5 confirmed these samples as outliers in the data set.

3.3. Descriptive parameters and test for normality for the whole data sets

Percentiles for the whole raw data set are listed in Table 1. Concentration values for the elements varied over a wide range of several magnitudes, showing the complex nature of soil geochemical features in Northern Ireland, e.g. the minimum and maximum values for Ca were 0.0002% and 15.21%, respectively. When we compared the minimum values and 5th percentiles, some extremely low values (outliers) were observed. Meanwhile, extremely high values (outliers) were also identified by comparing the maximum values and 95th percentiles. Median values are a robust measure of central tendency for data sets with such extreme values. The ranges between 25th and 75th percentiles (inter-quartile ranges) showed that 50% of the data varied over a much narrower range than the whole data set. For example, the inter-quartile range for Cd was 0.20–0.50 mg/kg, compared with a range for the whole data set of <0.08–5.49 mg/kg. One important observation for Cd was that 18% (as explained earlier) of the samples were below the detection limit (0.08 mg/kg). Concentrations of the other elements were generally well above their detection limits. The detection limit for each element studied, together with the CV of measurement of the reference soil samples analysed as part of the QC procedures, is appended to Table 1. Of the elements studied, only Cd had a high CV value which reflects the fact that the internal reference soil used had a mean Cd concentration only three times that of the detection limit for this element.

Since outliers were clearly identified in the normal Q-Q plots of Figs. 4 and 5 (see also Table 1), it is necessary to remove these outliers prior to testing for normality and subsequent analyses (except for the calculation of median values due to their robustness against outliers). Low value outliers were determined by sorting the raw data and by reference to the normal Q-Q plots for the logarithmically transformed data. The high value outliers were identified from the definition of "extreme" values of a Box-and-Whisker plot viz. values higher than the third quartile plus 3 times inter-quartile range. The specific numbers of outliers identified were listed in Table 1. The number of high value outliers were greater than those shown on the normal Q-Q plots of Fig. 4, but they were still only a very small portion of the total number of data values (less than 1%, except for Cd (1.4%), Na (2.4%) and Pb (2.5%)). The high value outliers can be attributed to contamination, natural mineralization, or both. The very high values for Ca were likely due to the presence of limestone debris in soils that have not been well weathered. Most of the low value outliers identified were values below the detection limits for the elements concerned. It should be noted that since 18% of all Cd values were below the detection limit, no low value outlier for Cd was removed.

Non-normality and non-lognormality are widely observed in geochemical databases. Table 2 shows both the skewness and kurtosis values, and the results of the K-S test for normality of the data set, after removal of the outliers.

All elements in the whole raw data set showed positive skewness values with relatively high skewness values found for Na and Pb (both 1.44). It was noted that the skewness values were not extremely high due to the removal of outliers in the data set. The results from the K-S test (Table 2) showed that none of the 14 elements under study passed the test for normality, even after removal of outliers.

The log-transformed data showed, as expected, generally less deviation from normality than the raw data, with relatively smaller skewness values. The significance value (K-S p) for Ca even increased to 0.01. However, with the exception of Na and Pb, most of the skewness values

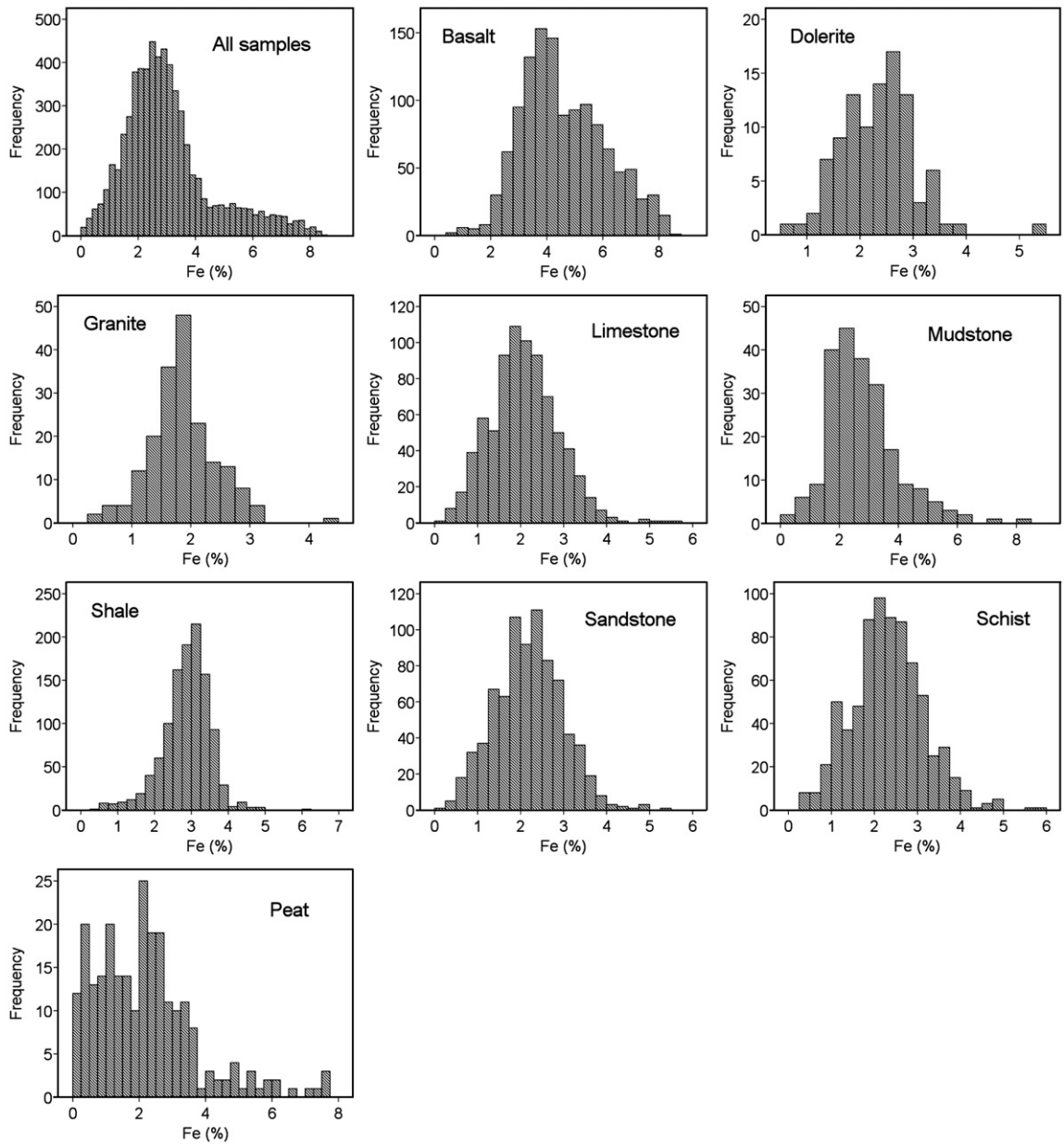

Fig. 6. Comparison between histograms for Fe concentrations in all soil samples ($n=6109$) and samples grouped by rock type or peat (a total of 29 outliers were excluded).

changed from being positive to being negative showing that log-transformation had "over-transformed" most of the data sets. Another important feature for the logarithmic transformation was that the kurtosis values for several elements obviously increased (e.g., Cr, Fe, Mn, P, and Zn). Together with the negative skewness values, it was clear that while logarithmic transformation is effective in reducing the proportion of high values in the data set, it also pushes low values away from the centre of the data set, causing the negatively skewed and sharp distribution of the transformed data. For most elements, the shift towards normality (with generally smaller absolute skewness values) after the log-transformation was related to their positive skewness values before transformation. While a log-transformation is inflexible, a power transformation (e.g. Box–Cox) will provide the flexibility to choose an appropriate power to push the skewness towards "0" (Zhang and Selinus, 1998).

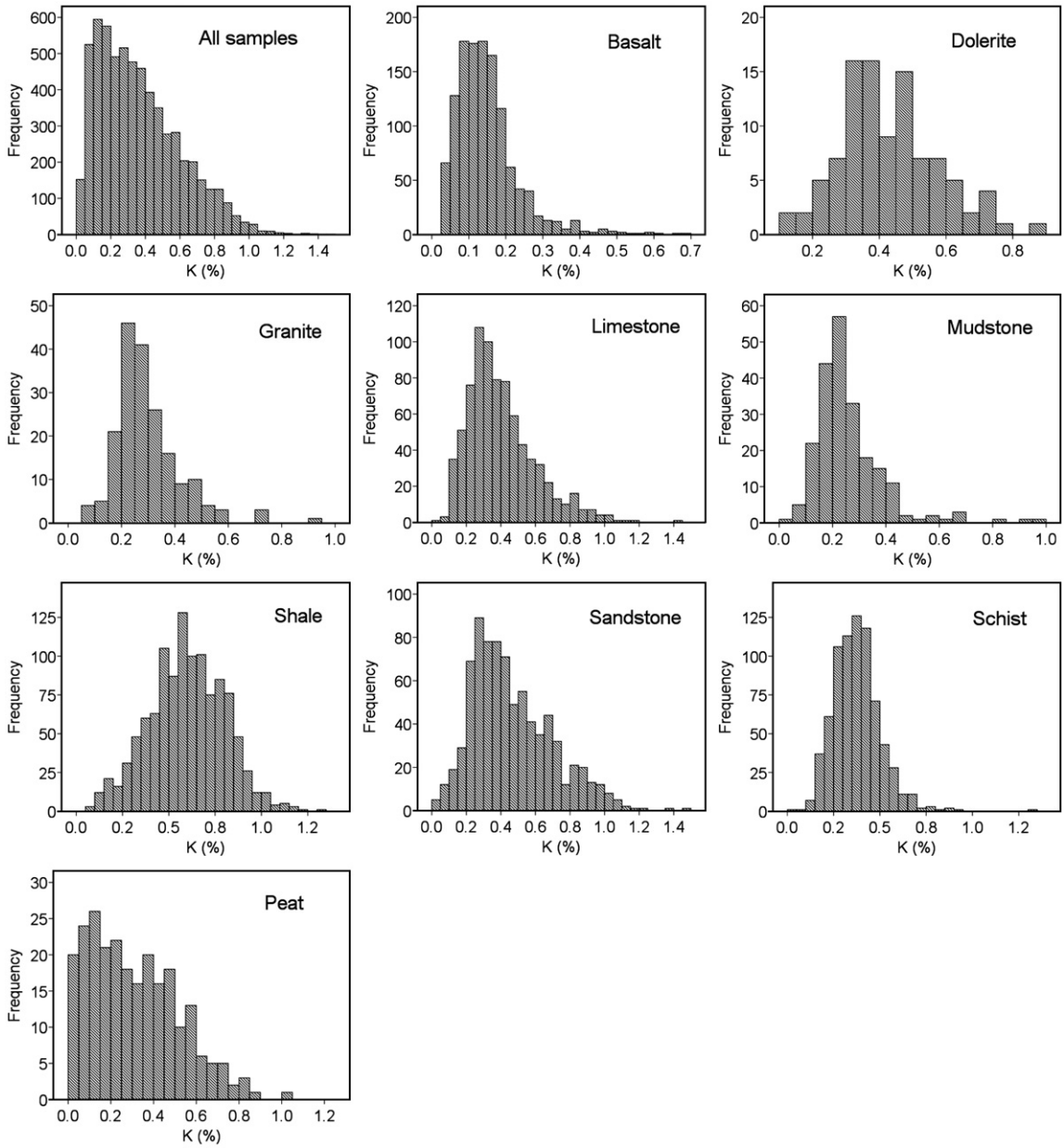

Fig. 7. Comparison between histograms for K concentrations in all soil samples (n=6137) and samples grouped by rock type or peat (1 outlier was excluded).

3.4. Comparisons of data grouped by rock type

Factors such as climate, topography, soil genesis, geology and human activities strongly influence the chemical nature of soils. Based on the initial GIS mapping (Fig. 3), geology (mainly rock type) plays a dominant role in soil geochemistry in the study area. If soil samples are classified by rock type, each group of samples should belong to a more similar population, and thus their probability distribution should display a better tendency towards normality. Our hypothesis is: "IF rock type is a main influencing factor for soil geochemistry, THEN soil samples separated by rock type should have a better tendency towards normality." It should be noted that the logic order of the hypothesis cannot be swapped i.e. we are, specifically, *not* testing "IF there is a better tendency towards normality for soils separated by rock type, THEN rock type is a main influencing factor."

To test our hypothesis, the raw data for 6138 samples were classified into 8 major rock types (viz. basalt, dolerite,

Table 4
Significance levels of the Kolmogorov–Smirnov test for normality (after Lilliefors correction) for raw soils data grouped by dominant rock type and peat [a]

Element	Basalt	Dolerite	Granite	Limestone	Mudstone	Shale	Sandstone	Schist	Peat
n [b]	1591	102	199	832	249	1225	855	789	296
Ca	0	0.006	0	0	0	0	0	0	0
Cd	0	0	0	0	0	0	0	0	0
Co	0	0.001	0	0	0	0	0	0	0
Cr	0	0	0	0	0	0	0	0	0
Cu	0	0	0.013	0	0	0	0	0	0
Fe	0	>0.200	0	0.034	0	0	0.062	0.004	0
K	0	>0.200	0	0	0	>0.200	0	0	0
Mg	0	0.111	0	0	0	0	0	0	0
Mn	0	0	0	0	0	0	0	0	0
Na	0	>0.200	0.003	0	0	0	0	0	0
Ni	0.002	0.003	0	0	0	0	0	0	0
P	0	0.024	0.005	0	0	0	0	0	0
Pb	0	0	0	0	0	0	0	0	0
Zn	0.002	0.011	0	0	0.005	0	0	0	0

[a] "0" values represent "<0.001".
[b] Actual sample number varies among different rock types and elements due to different number of outliers removed.

granite, limestone, mudstone, shale, sandstone, schist) and peat using the simplified geology map of Northern Ireland (Fig. 1). The peat class includes some organic mineral soils such as humic rankers and organic mineral gleys which raise the elemental concentrations for some samples above that expected for "pure" peat. The median values, calculated with all the raw data including the previously identified outliers, by rock type are listed in Table 3.

The differences between rock types are clearly illustrated in Table 3. Soils in basalt areas have significantly higher median concentrations for most elements under study (Ca, Co, Cr, Cu, Fe, Mg, Mn, Na, Ni, P, and Zn), while the highest values for Cd and Pb occurred in the shale areas. Soils in mudstone/shale areas generally displayed the second or third highest median concentrations for these elements among the 8 rock types and peat.

The high concentrations of most of the elements in soils are inherited from the parent rock through weathering. Basalt rock is very fine grained and contains almost no quartz. Dolerite and granite contain large amounts of quartz and feldspar, but generally contain low concentrations of trace elements. Soils in areas of sedimentary rock generally contain intermediate concentrations of most elements but mudstone and shale contain higher concentrations of many elements due to their relatively fine grain sizes. Trace elements in sandstones are diluted by quartz. Limestone contains very low concentrations of most elements except Ca (and Sr which is not included in this study). However, limestone is easily weathered and soils developed on limestones can become enriched with trace elements. In this study, Cd had the highest median values in the limestone and shale areas (0.41 mg/kg). Schist contains intermediate amounts of most elements, similar to the average level for the whole data set. Peat generally contains very low concentrations of most elements under study (except for Pb which may be due to historical atmospheric deposition) as it is comprised almost entirely from organic matter but, as our sample also includes organic mineral soils under this heading, the elemental concentrations are higher than for pure peats. Based on these results, geology is the overriding influence on soil geochemistry in the study area.

3.5. Probability features of data grouped by rock type

To further investigate the effects of rock type on soil geochemistry, histograms were produced for Fe (Fig. 6) and K (Fig. 7) in soils grouped by rock type and peat (after removal of outliers).

When the soil samples were grouped by rock type and peat, the histograms showed significant improvement towards normality compared with those for all samples. The histogram for Fe in all samples showed a clear multi-modal and positive skewness feature, with the high values mainly contributed by soils from the basalt areas. It was clear that the histograms for Fe in granite, limestone, sandstone and schist areas were quite symmetric. The histogram for K in all samples was also clearly positively skewed and became more symmetric when the samples were grouped by rock type.

Meanwhile, deviations from normality can still be observed for most of the histograms. Fe in basalt was positively skewed, with a few very low values. These

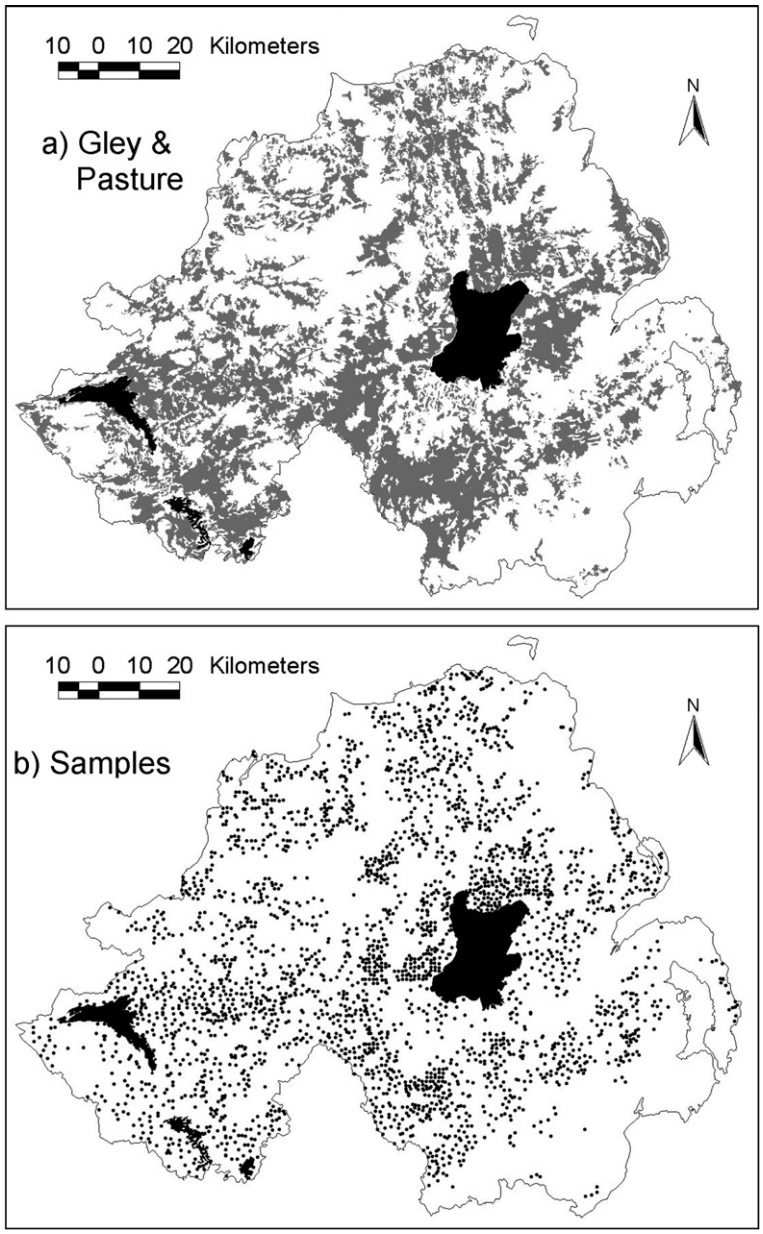

Fig. 8. Spatial distribution of gley and pasture and locations of soil samples extracted from both areas: a) Gley and pasture, b) soil samples.

low values in the basalt area may be due to the presence of peaty soils. On the other hand, there were some high values of Fe in almost all the other rock type areas, which can be related to the spatial heterogeneity of different rocks (e.g. with different degrees of mineralization or metamorphic processes). Since many of the high value outliers were already excluded from these histograms, the impact of other factors must be taken into account. In particular, the effect of the glacial movement which caused the significant mixing of soil parent materials needs to be considered. Thus, some soils have developed from tills that contain materials transported from other places. This, together with the fact that simplified rock type classes based on the dominant rock type present was used to stratify the original data set, explains why outliers were observed in almost every rock type area. For example, quartzite and limestone were both found in the schist area.

Histograms for both Fe and K in peat appeared much more skewed than in the rock type areas, with obviously high value outliers. The outliers observed in the "peat" data set were expected due to the inclusion of other

Table 5
Significance levels of the Kolmogorov–Smirnov test for normality (after Lilliefors correction) for raw data from both gley and pasture areas grouped by dominant rock type [a]

	Basalt	Dolerite	Granite	Limestone	Mudstone	Shale	Sandstone	Schist
n [b]	835	43	25	552	121	466	495	345
Ca	0	>0.200	>0.200	0	0	0	0	0
Cd	0	0	>0.200	0	0.002	0	0	0
Co	0.046	0.076	0.098	0	0	0.105	0	0
Cr	0	0.004	>0.200	0	0	0	0	0
Cu	0	0.002	0.030	0	0.052	0	0	0
Fe	0	>0.200	>0.200	0.006	0	0.145	>0.200	>0.200
K	0	>0.200	0.018	0	0	0.150	0	0.002
Mg	0	0.180	0.114	0	0	0.001	0	0
Mn	0.001	0.005	>0.200	0	0.004	0	0	0
Na	0	0.025	>0.200	0	0	0	0	0
Ni	0	0.028	>0.200	0	0	0.001	0	0
P	0	0.198	>0.200	0	0.001	0	0	0.017
Pb	0	0.013	0.006	0	0	0	0	0
Zn	0.063	0.012	0.048	0	0.193	0	0.006	0

[a] "0" values represent "<0.001".
[b] Actual sample number varies among different rock types and elements due to different number of outliers removed.

organic soils such as humic rankers and organic mineral gleys in this data set.

The normality of the data sets by rock type and peat was tested using the K-S test after Lilliefors correction (Table 4).

The elements Fe, K, Mg, and Na in dolerite, K in shale and Fe in sandstone areas passed the K-S test at a significance level of 0.05. Most elements in most rock type groups had a very low significance level $p<0.001$. This showed that even though the soil samples were separated by rock type, it was still difficult for most elements to pass a test for normality. Elements in the dolerite area had the best chance of passing a test for normality as shown in Table 4.

One of the reasons more elements in dolerite areas passed the test for normality was that dolerite is not very much mixed with other rock types. However, this is also the case for basalt and granite, but very few elements in these two areas passed the test. This may be explained by sample size. Zhang et al. (2005) suggested that when the sample size is large, hypothesis tests gain power, resulting in rejection of the null hypothesis for most real data sets. Zhang et al. (2005) also suggested a sample number of 1000 as indicative of a large sample size. In this study, the sample size for soils in the basalt area was 1591 and was one of the main reasons why it was hard for elements in soils of this rock type to pass the test for normality. On the other hand, there were only 102 samples in the dolerite area, and such a relatively small sample size would favour normality. An exception was the raw data for K in shale, which passed the K-S test with a sample size of 1225. This element may have been quite homogenous in the shale area with little mineralization. It is expected that, if the sample size became smaller, more elements would pass the tests for normality (Zhang et al., 2005). This is demonstrated by the histograms in Figs. 6 and 7 where a greater tendency towards normality was achieved by classifying samples by rock type and peat. Meanwhile, the statistical constraints of sample size, together with the non-normality or non-lognormality for most elements, imply that factors other than rock type play an important role at small spatial scales within the hierarchy of geochemical landscapes with spatial variations at different scales.

3.6. Probability features of data from both gley and pasture areas grouped by rock type

To further investigate the probability features of soil geochemistry, factors of soil type and land cover were considered. In Northern Ireland, gley is the main soil type (56%), and pasture is the main land cover (90% including rough grazing). Using the overlay function of the GIS, areas with both gley and pasture were combined and used to extract a total of 2882 soil samples from a total of 6138 (Fig. 8).

In the 2882 samples selected, the effects of soil type and land cover can be treated as being relatively constrained. These soil samples were classified by rock type and the probability distributions for the raw data tested using the K-S test after Lilliefors correction (Table 5).

If we compare the results in Tables 4 and 5, we can see that more opportunities of passing the test for normality

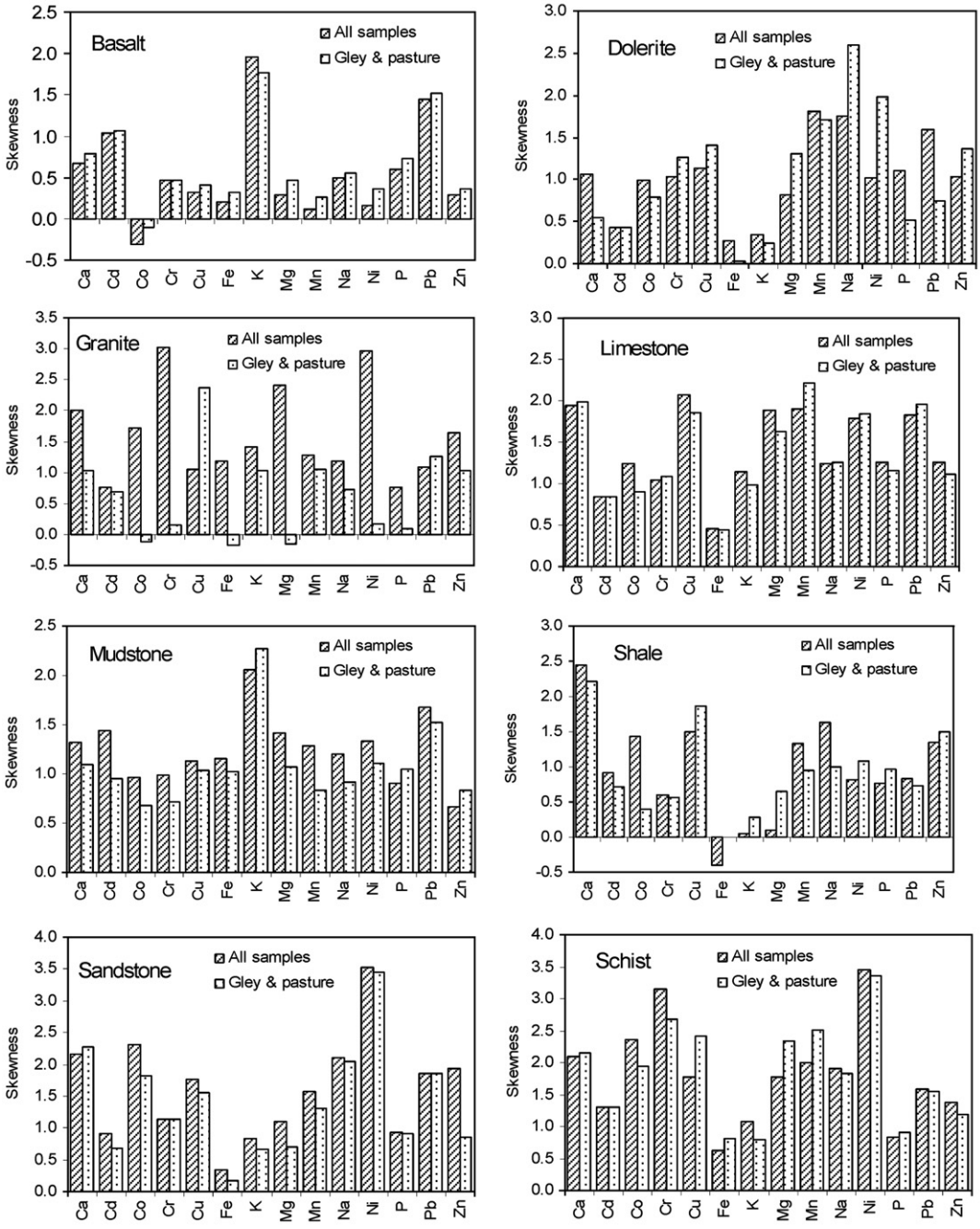

Fig. 9. Comparison between skewness values for pseudo-total element concentrations by rock type from all soils and from soils in gley and pasture areas only.

were obtained for the raw (untransformed) data when the effects of soil type and land cover were constrained. Using a p value of 0.05 as the threshold, 24 tests passed the K-S test (Table 5) as compared to a total of 6 tests (Table 4) when soil type and land cover were not considered. Samples in the dolerite and granite areas showed the greatest tendency to pass the tests for normality, which can be partly attributed to their relatively small sample size of 43 and 25, respectively.

Since sample size has been shown to play a role in the K-S test for normality when soil type and land use were constrained, skewness values were calculated for the elements by rock type from all soils and from soils in gley and pasture areas only (Fig. 9). The results appeared

fairly mixed for dolerite, limestone, shale and schist areas with both increased and decreased skewness values when the soil type and land use were constrained. This demonstrated the complexity of soil geochemical probability features. However, soils from granite (except for Cu and Pb), mudstone (except for K, P and Zn), and sandstone (except for Ca, Cr and Pb) showed obviously decreased skewness values, which was in line with the results of the K-S test. An exception is soils from basalt areas which displayed increased skewness values for most elements. Taking results from both the K-S test and skewness values into consideration, improvement towards normality was evident when the sample was constrained by soil type and land use, even though such improvement was not as significant as when the samples were classified by rock type.

In this study, the effects of rock type, as well as soil type, land cover and sample size on probability features of soil geochemistry were investigated. It is expected that other factors may also have influenced soil geochemistry, such as different degrees of mineralization and metamorphic processes of bedrocks, non-independence of sampling and the non-zero feature of geochemical concentrations. Meanwhile, it should be mentioned that the GIS overlay operations were made based on generalized maps, which may have also omitted the variations of soil geochemistry at small scales. Attempts made in this study show that it is still a challenging task in environmental geochemistry to separate all the factors controlling soil geochemistry and to investigate their influences at the regional scale.

4. Conclusions

Outliers in soil geochemistry were observed using normal Q-Q plots. High value outliers were identified using the normal Q-Q plots for raw data while low value outliers were better detected using the normal Q-Q plots for the log-transformed data. These outliers were probably associated with rare processes such as mineralization and human pollution.

None of the whole raw data sets for the total concentrations of 14 elements followed either normal or lognormal distributions and all displayed positive skewness values. Logarithmic transformation "over-transformed" most of the raw data sets changing their skewness from positive to negative values.

A GIS overlay function was useful to classify soil samples by rock type which enabled further statistical analyses to be made. Obvious differences in chemical concentrations between rock types were observed: soils in basalt area had the highest levels of most metals in the study (Ca, Co, Cr, Cu, Fe, Mg, Mn, Na, Ni, P and Zn) but the lowest levels of K, while the highest levels for Cd and Pb occurred in the shale areas. After classification by rock type and peat, soil geochemistry values showed better tendencies towards normality. Such a trend was strengthened when the influences of soil type and land cover were restrained. However, the data sets still had difficulty satisfying a test for normality unless the sample size was small.

The factors of rock type, soil type, land cover and sample size were found to influence the probability features of soil geochemistry and the results of the statistical tests. However, the influences of other factors cannot be omitted and it remains a challenge to properly quantify these influences.

Acknowledgments

The authors acknowledge the support provided from the EU Structural Fund, made available through the Environment and Heritage Service of the Department of the Environment of Northern Ireland, in funding a significant portion of the work associated with the creation of the Soil Geochemical Atlas for Northern Ireland (Jordan et al., 2000) which was the source of the data used in this study. The support of field and analytical staff in the Agri-Environment Branch, Agriculture, Food and Environmental Science Division of the Agri-Food and Biosciences Institute is also gratefully acknowledged for assistance in the sampling and preparation of soils and their analysis. The authors also gratefully acknowledge the comments and suggestions made by the two reviewers which have improved the quality of this paper.

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.

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

Betts, N.L., 1997. Climate. In: Cruickshank, J.G. (Ed.), Soil and Environment: Northern Ireland. Agricultural and Environmental Science Department, The Queen's University, Newforge Lane, Belfast, pp. 63–84. Chapter 4.

Cruickshank, J.G. (Ed.), 1997. Soil and Environment: Northern Ireland. Agricultural and Environmental Science Department, The Queen's University of Belfast, Newforge Lane, Belfast.

Cruickshank, M.M., Tomlinson, R.W., 1996. Application of CORINE land cover methodology to the UK — some issues raised from Northern Ireland. Glob. Ecol. Biogeogr. 5, 235–248.

Davies, B.E., 1980. Trace element pollution. In: Davies, B.E. (Ed.), Applied Soil Trace Elements. John Wiley & Sons, Chichester, pp. 287–351.

Dore, C.J., Watterson, J.D., Murrells, T.P., Passant, N.R., Hobson, M.M., Baggott, S.L., Thistlethwaite, G., Goodwin, J.W.L., King, K.R.,

Adams, M., Walker, C., Downes, M.K., Coleman, P.J., Stewart, R.A., Wagner, A., Sturman, J., Conolly, C., Lawrence, H., Cumine, P.R., 2005. UK emissions of air pollutants 1970 to 2003. AEA Technology, 551 Harwell, Didcot, Oxon. OX12 0QJ.

GSNI, 1998. The solid geology of Northern Ireland: a vector map at 1:250,000 scale. Geological Survey of Northern Ireland, Belfast.

Jordan, C., Higgins, A., Hamill, K., Cruickshank, J.G., 2000. The Soil Geochemical Atlas of Northern Ireland. Agriculture, Food and Environmental Science Division, Department of Agriculture and Rural Development, Newforge Lane, Belfast.

McBratney, A.B., Webster, R., McClaren, R.G., Spiers, R.B., 1982. Regional variation of extractable copper and cobalt in the topsoil of south-east Scotland. Agronomie 2, 969–982.

McGrath, S.P., Loveland, O.J., 1992. The Soil Geochemical Atlas of England and Wales, (SGAEW). Blackie Academic & Professional, London.

Mitchell, W.I. (Ed.), 2004. The Geology of Northern Ireland: Our Natural Foundation, 2nd Edition. Geological Survey of Northern Ireland, Belfast.

Ordinance Survey of Ireland (OSI), 1953. Tables for the transverse Mercator projection of Ireland. Ordnance Survey. Phoenix Park, Dublin.

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.

Tomlinson, R.W., 1997. Land Cover — based on the CORINE land cover programme. In: Cruickshank, J.G. (Ed.), Soil and Environment: Northern Ireland. Agricultural and Environmental Science Department, The Queen's University, Newforge Lane, Belfast, pp. 99–117.

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

Webb, J.S., Applied Geochemistry Research Group, 1973. Provisional Geochemical Atlas of Northern Ireland. Imperial College of Science and Technology, London.

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

Zhang, C.S., Manheim, F.T., Hinde, J., Grossman, J.N., 2005. Statistical characterization of a large geochemical database and effect of sample size. Appl. Geochem. 20, 1857–1874.

Zhang, C.S., Jordan, C., Higgins, A., 2007. Using neighbourhood statistics and GIS to quantify and visualize spatial variation in geochemical variables: an example using Ni concentrations in the topsoils of Northern Ireland. Geoderma 137 (3–4), 466–476.