



PERGAMON

Applied Geochemistry 17 (2002) 185–206

www.elsevier.com/locate/apgeochem

**Applied
Geochemistry**

Factor analysis applied to regional geochemical data: problems and possibilities

Clemens Reimann^{a,1}, Peter Filzmoser^{b,*}, Robert G. Garrett^c

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

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

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

Received 5 July 2000; accepted 1 March 2001

Editorial handling by R. Fuge

Abstract

A large regional geochemical data set of C-horizon podzol samples from a 188,000 km² area in the European Arctic, analysed for more than 50 elements, was used to test the influence of different variants of factor analysis on the results extracted. Due to the nature of regional geochemical data (neither normal nor log-normal, strongly skewed, often multimodal data distributions), the simplest methods of factor analysis with the least statistical assumptions perform best. As a result of this test it can generally be suggested to use principal factor analysis with an orthogonal rotation for such data. Selecting the number of factors to extract is difficult, however, the scree plot provides some useful help. For the test data, a low number of extracted factors gave the most informative results. Deleting or adding just 1 element in the input matrix can drastically change the results of factor analysis. Given that selection of elements is often rather based on availability of analytical packages (or detection limits) than on geochemical reasoning this is a disturbing result. Factor analysis revealed the most interesting data structures when a low number of variables were entered. A graphical presentation of the loadings and a simple, automated mapping technique allows extraction of the most interesting results of different factor analyses in one glance. Results presented here underline the importance of careful univariate data analysis prior to entering factor analysis. Outliers should be removed from the dataset and different populations present in the data should be treated separately. Factor analysis can be used to explore a large data set for hidden multivariate data structures. © 2002 Elsevier Science Ltd. All rights reserved.

1. Introduction

The principal aim of factor analysis, which was developed initially by psychologists, is to explain the variation in a multivariate data set by as few “factors” as possible and to detect hidden multivariate data structures. The term “factor” used by psychologists is equivalent to “controlling processes” in geochemistry. Thus, theoretically, factor analysis should be ideally suited for an easy presentation of the “essential” information

inherent in a geochemical data set with many analysed elements.

In regional geochemistry an advantage would be that instead of presenting maps for 40–50 (or more) elements only maps of 4–6 factors may have to be presented, containing a high percentage of the information of the single element maps. It is even more informative if factor analysis can be used to reveal unrecognised multivariate structures in the data that may be indicative of certain geochemical processes, or, in exploration geochemistry, of hidden mineral deposits. Factor analysis has been successfully used for this purpose (e.g. Garrett and Nichol, 1969; Chork and Govett, 1985; Chork, 1990; Chork and Salminen, 1993), but is still a controversial method. A focal point of critique is that too many different techniques are available, all giving

* Corresponding author. Tel.: +43-1-58801-10733.

E-mail addresses: p.filzmoser@tuwien.ac.at (P. Filzmoser), garrett@gsc.nrcan.gc.ca (R.G. Garrett), clemensreimann@yahoo.co.uk (C. Reimann).

¹ Present address: Margaretenstr. 110/6, A-150 Wien, Austria.

slightly different results (Rock, 1988). It is argued that the statistically untrained users will thus always be tempted to experiment until finding a solution that fits their preconceptions. Another problem is that users are not always aware of some of the basic requirements for carrying out a successful factor analysis. As a result, factor analysis is very often merely applied as an exploratory tool, and results could often have been predicted using much simpler methods.

Using a large regional scale geochemical data set containing 605 samples and more than 50 variables, the objective of this study is to answer some fundamental questions with regard to the use of factor analysis in geochemistry:

- Can (and should) factor analysis be applied to such a high-dimensional data set? What are the prerequisites for applying factor analysis to such a data set?
- What are the results of factor analysis when such a large data set is investigated?
- Can the information contained in more than 50 single element maps be presented in just a few (e.g. 3–6) factor maps?

Furthermore, an attempt is made to answer the question as to which parameters have the largest influence on the results of factor analysis. The influence of the:

- actual method used (principal factor analysis (PFA) versus maximum likelihood (ML), method of factor rotation),
- number of factors extracted, and
- number of elements entered into the factor analysis will be discussed.

2. Materials and methods

2.1. The Kola project

From 1992 to 1998, the Geological Surveys of Finland (GTK) and Norway (NGU) and Central Kola Expedition (CKE), Russia, carried out a large, international multi-media, multi-element geochemical mapping project, covering 188,000 km² north of the Arctic Circle. The entire area between 24 and 35.5°E up to the Barents Sea coast (Fig. 1) was sampled during the summer of 1995. Results of the “Kola Ecogeochemistry” project are documented on a web site (<http://www.ngu.no/Kola>) and in a geochemical atlas (Reimann et al., 1998). One of the sample media for this project was the C-horizon of podzol soil profiles, developed on glacial drift. The average sample density was 1 site per 300 km². C-horizon samples were taken at 605 sites and subsequently analysed by a number of different techniques for more than 50 elements, resulting in 89 variables. The

project was primarily designed to reveal the environmental conditions in the area, as reflected by the very low sample density, and an aqua regia extraction method applied to the <2 mm grain size fraction of the soils for chemical analysis.

For a geochemical mapping or exploration project, a much higher sample density is usually considered necessary. In the case of stream sediments, about 1 sample per 1–3 km² is collected; in the case of soils, up to several hundred samples per km² are often required to intersect the anomalous patterns related to a target of the size of an average ore body. On the other hand, it has been proposed that mineral deposits occur in “geochemical provinces” — large areas with enhanced concentrations of certain elements (Hawkes and Webb, 1962), which should be easily detectable with low-density geochemistry (Bølviken et al., 1990, 1992). For regional geochemical mapping using glacial till, the fine fraction (<0.063 mm) has mostly been used (Koljonen, 1992). In Europe geochemists and geologists commonly study total element concentrations in their samples. With an aqua regia extraction of the <2 mm fraction, mineral weathering properties and secondary processes will play an important role in determining the element concentrations found in the samples. A geological interpretation of the C-horizon results is presented in Reimann and Melezhik (2001).

2.2. Geology

The bedrock of the study area includes rocks forming part of the basement of the Fennoscandian Shield. They are mostly Archean or Early Proterozoic in age, autochthonous and parautochthonous cover rocks of Late Proterozoic to Early Cambrian age, allochthonous basement and cover rocks of Late Proterozoic to Early Cambrian age. Autochthonous rocks of similar ages occur in windows within the Caledonian Orogen. The rocks in large areas of the Kola Peninsula and in Finland are Archean. In the Kola Peninsula, the supra-crustal rocks include felsic gneisses, iron quartzites (BIF) and amphibolites (Fig. 2). In Finland, the rocks belonging to this group are similar in the northeastern and eastern areas. In a large area in the western part of the study area, the bedrock also includes felsic gneisses and tonalites. In the Granulite Complex, 2000–1900 Ma in age, extending from the western part of the Kola Peninsula through northern Finland to Norway, the rocks are mainly mafic and felsic granulites, or high-grade gneisses. On the coast of the Barents Sea, in the northern part of the Kola Peninsula, there are Archean felsic gneisses and tonalites as well as arkosic and quartzitic sandstones of late Proterozoic age.

A greenstone belt extends from Russia, through Finland, to Norway. It consists predominantly of basaltic and komatiitic volcanites of Proterozoic age (Bølviken

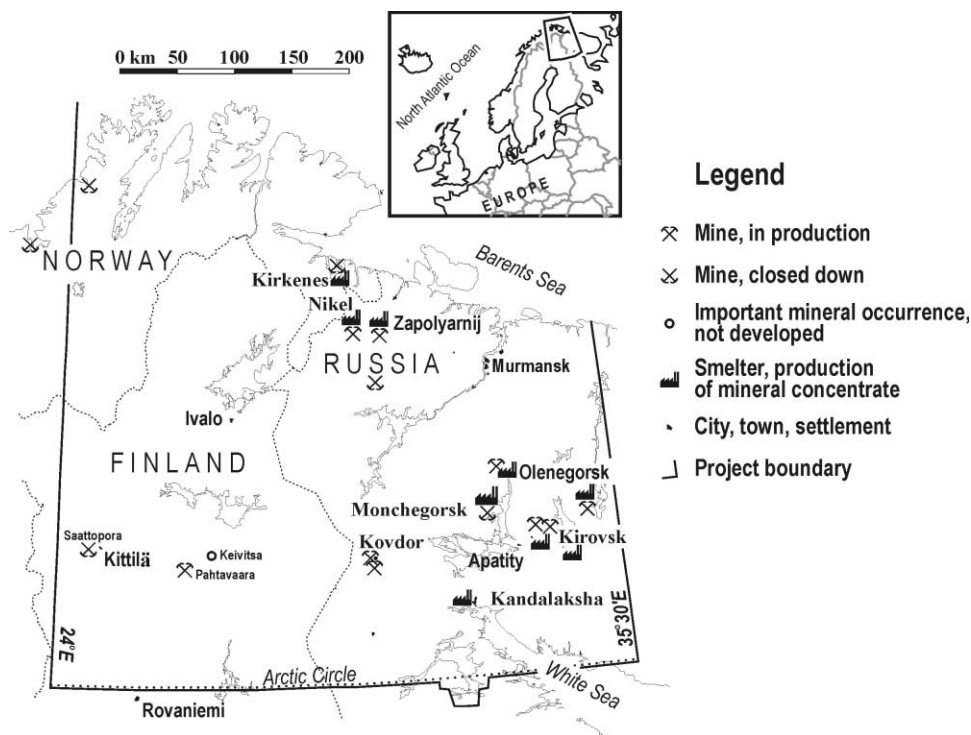


Fig. 1. General location map of the study area for the Kola Project (Reimann et al., 1998). Locations named in the text are given.

et al., 1986; Koljonen, 1992). This type of bedrock (in a general sense) exists also north of the White Sea in Russia and in the Pechenga Belt in the northern part of the Kola Peninsula. On the margins of this belt, there are also andesitic and dacitic rocks. The greenstone belts contain several ore bodies, e.g. Fe ores in western Finnish Lapland, a Cu deposit (Kittilä), Au-ores (Pahtavaara (in production), Saattopora and Bidjovagge), Ni-ores (Nikel and Zapolyarnij) and others (e.g. Keivitsa Ni–Cu–PGE) (see Table 1 in Reimann and Melezhik, 2001).

The study area includes intrusive complexes of various ages, from Archaean to Palaeozoic. North of the Granulite Complex there are granite, granodiorite and alkaline granite intrusions (Vainospää and Litza-Ura-Guba). The Koitelainen layered gabbro intrusive occurs south of the artificial lake Lokka. Intrusive complexes of alkaline rocks are found at Sokli (phosphorite-bearing and a possible Nb-occurrence) in Finland, and at Kovdor (apatite-magnetite) in Russia. The alkaline intrusions in the Khibiny/Lovozero Mountains are among the most intensely studied of their kind in the world. The related apatite and loparite deposits have been mined for decades. Ultramafic alkaline rocks occur east of Kirovsk. In Central Lapland, there are granitoids, a group constituting silica-rich, felsic rocks.

In Norway, along the northern coast, there are supracrustal rocks of Late Proterozoic–Cambrian age. In the western part, these are metamorphosed and

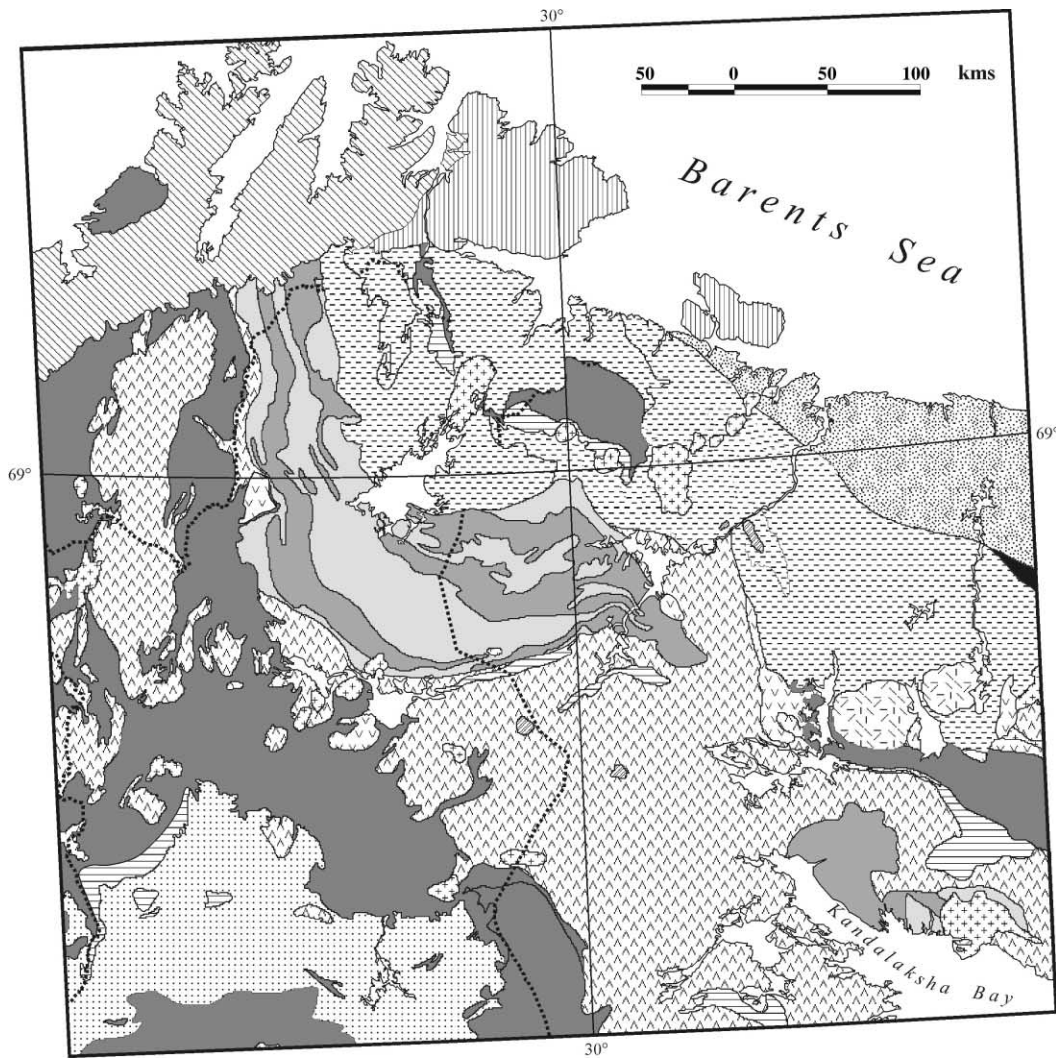
include quartzites, meta-arkoses, phyllites and gray-wackes. In the eastern part, the rocks are sedimentary and Proterozoic in age: conglomerates, sandstones, siltstones and mudstones, locally enriched in P.

This extremely heterogeneous geological environment is well reflected in the simplified geological map of the survey area (Fig. 2). For a colour version of this map, see Reimann et al. (1998) or Reimann and Melezhik (2001).


2.3. Quaternary geology

The study area is part of the glaciated terrain of Northern Europe. The area was entirely covered by ice during the Pleistocene, which began 2–3 Ma ago (Eriksson, 1992). During this period, Northern Europe was glaciated and deglaciated at least 3 times; warmer climatic periods intervened, giving rise to ice-free interglacials. For a shorter period during the last, Weichsel, glaciation, the ice in Fennoscandia melted almost totally.


The main Quaternary deposits are till and peat; there are also large areas without any drift, covered by outcrops and boulder fields (see map of Quaternary deposits of Finland and north-western part of Russian Federation and their resources; Niemelä et al., 1993). Till consists of an unsorted mixture of rock and mineral fragments from boulders to clay size. The rock material




Caledonian rocks


 Quartzite, meta-arkose, dolostone, phyllite, greywacke, gabbro


Neoproterozoic


 Conglomerate, gritstone, sandstone, siltstone, mudstone, in places enriched in phosphorous


Palaeoproterozoic

 Basaltic volcanites, subordinate 'black schist', conglomerate, quartzite, dolostone, gabbro

 Andesite, picrite, basalt, greywacke, 'black schist', chert, limestone


 Granite, granodiorite, rappakivi granite, gneiss, greywacke, marble


 Gabbro

 Granite, granodiorite, alkaline granite

Archaean

 Felsic/mafic granulite

 Gneiss, granite, tonalite, iron quartzite, amphibolite

 Tonalite, granite/gneiss, tonalite, amphibolite

 Basalt, komatiite

Palaeoproterozoic to Palaeozoic

 Alkaline/ultramafic alkaline igneous rock

Fig. 2. Simplified geological (lithological) map of the study area.

Table 1
Statistical summary of the test data set^{a,b,c}

Element	Unit	DL	<DL	Min	Max	Mean	Median	S.D.	MAD	<i>p</i>	<i>p</i> _{ln}	<i>p</i> _{Box-Cox}
Ag	mg/kg	0.001	0.2	<0.001	0.119	0.011	0.008	0.011	0.004	<0.001	<0.001	<0.001
Al	mg/kg	10	0	1840	85,900	12,665	9910	9814	5834	<0.001	0.002	0.5
Al_XRF	wt. %	0.03	0	2.92	12.08	7.34	7.38	0.969	0.667	<0.001	<0.001	<0.001
As	mg/kg	0.1	1.7	<0.1	30.7	1.25	0.5	2.349	0.445	<0.001	<0.001	<0.001
Ba	mg/kg	0.5	0	4.7	1300	60.15	43.5	74.33	28.91	<0.001	0.059	0.5
Ba_INAA	mg/kg	50	0	210	3000	600	575	224	170	<0.001	0.046	0.002
Be	mg/kg	0.05	0	0.06	14	0.442	0.235	1.06	0.141	<0.001	<0.001	0.5
Bi	mg/kg	0.005	2.3	<0.005	3.89	0.049	0.026	0.164	0.021	<0.001	<0.001	0.001
Ca	mg/kg	3	0	110	41,700	2279	1905	2383	1075	<0.001	<0.001	<0.001
Ca_XRF	wt. %	0.005	0	0.03	6.76	2.133	2.17	0.899	0.801	0.007	<0.001	0.003
Cd	mg/kg	0.001	0	0.007	0.221	0.029	0.024	0.02	0.01	<0.001	<0.001	<0.001
Ce_INAA	mg/kg	3	0	12	500	58.84	45	53.23	23.72	<0.001	0.001	0.023
Co	mg/kg	0.2	0	1.2	44.3	8.22	7	5.029	3.706	<0.001	0.5	0.5
Co_INAA	mg/kg	1	0.2	<1	57	14.27	13	6.718	5.93	<0.001	<0.001	<0.001
Cr	mg/kg	0.5	0	2.2	471	36.16	28.35	35.09	16.23	<0.001	0.033	0.5
Cr_INAA	mg/kg	5	0	11	910	116.15	99	87.51	45.96	<0.001	<0.001	<0.001
Cu	mg/kg	0.5	0	2	149	21.96	16.2	18.44	10.82	<0.001	0.5	0.5
Eu_INAA	mg/kg	0.2	0	0.3	14.3	1.239	1.05	1.006	0.371	<0.001	<0.001	<0.001
Fe	mg/kg	10	0	3310	79,200	17,236	14,700	10,189	7154	<0.001	0.5	0.5
Fe_XRF	wt. %	0.02	0	0.59	12.35	3.605	3.43	1.4	1.342	<0.001	0.5	0.5
Hf_INAA	mg/kg	1	0	2	120	6.47	6	6.588	1.483	<0.001	<0.001	<0.001
K	mg/kg	200	0.5	<200	11,000	1478	1100	1295	741	<0.001	0.004	0.004
K_XRF	wt. %	0.004	0	0.36	5.24	1.558	1.41	0.593	0.482	<0.001	<0.001	0.064
La	mg/kg	0.5	0	3.5	203	17.94	12.8	20.96	6.449	<0.001	<0.001	0.5
La_INAA	mg/kg	0.5	0	6.1	310	30.64	24	29.10	13.34	<0.001	<0.001	0.015
Li	mg/kg	0.5	0	1.7	70.9	9.129	7.2	6.883	4.3	<0.001	0.048	0.5
Lu_INAA	mg/kg	0.05	0	0.05	2.67	0.368	0.3	0.263	0.163	<0.001	0.038	0.5
Mg	mg/kg	5	0	370	70,500	4741	3720	4815	2002	<0.001	0.002	0.028
Mg_XRF	wt. %	0.02	0	0.12	7.32	1.271	1.15	0.677	0.526	<0.001	0.001	0.076
Mn	mg/kg	0.5	0	33.8	2140	185	128.5	180	65.83	<0.001	<0.001	0.5
Mn_XRF	wt. %	0.008	0	0.015	0.356	0.059	0.054	0.031	0.022	<0.001	<0.001	<0.001
Na	mg/kg	15	0	20	19,400	338	140	1368	88.96	<0.001	<0.001	<0.001
Na_XRF	wt. %	0.01	0	0.08	4.87	2.26	2.45	0.678	0.504	<0.001	<0.001	<0.001
Nd_INAA	mg/kg	5	1.3	<5	220	22.35	18	19.44	8.896	<0.001	<0.001	<0.001
Ni	mg/kg	1	0	1.2	228	23.41	18.65	21.09	11.56	<0.001	0.5	0.5
P	mg/kg	7	0	59	7170	446	393	368	185	<0.001	0.012	0.009
P_XRF	wt. %	0.004	0	0.004	0.589	0.045	0.039	0.032	0.019	<0.001	<0.001	<0.001
Pb	mg/kg	0.2	0	0.3	45.3	2.748	1.6	3.326	0.741	<0.001	<0.001	<0.001
S	mg/kg	5	0.5	<5	531	41.00	30	42.63	12	<0.001	<0.001	0.004
Sc	mg/kg	0.1	0.2	<0.1	15.4	2.816	2.3	1.809	1.186	<0.001	0.001	<0.001
Sc_INAA	mg/kg	0.1	0	1.7	36	13.60	13	5.695	5.93	<0.001	<0.001	<0.001
Si	mg/kg	10	0	50	590	154	140	64.94	44.48	<0.001	<0.001	<0.001
Si_XRF	wt. %	0.23	0	17.05	40.27	31.461	31.74	2.579	2.216	<0.001	<0.001	0.024
Sm_INAA	mg/kg	0.1	0	0.9	37	3.964	3.4	3.113	1.631	<0.001	0.042	0.061
Sr	mg/kg	0.5	0	1.6	1040	25.34	7.7	98.23	3.781	<0.001	<0.001	<0.001
Th_INAA	mg/kg	0.2	0	1	54	7.164	5.8	4.953	3.41	<0.001	0.016	0.022
Ti	mg/kg	0.5	0	48.8	5730	895	807	515	405	<0.001	<0.001	0.5
Ti_XRF	wt. %	0.003	0	0.053	1.9	0.362	0.347	0.16	0.151	<0.001	0.001	0.5
V	mg/kg	0.5	0	4.5	183	34.99	30.9	19.65	15.72	<0.001	0.5	0.5
Y	mg/kg	0.5	0	0.9	169	6.366	4.4	10.97	2.372	<0.001	<0.001	0.004
Yb_INAA	mg/kg	0.2	0	0.3	19.9	2.377	1.9	1.835	1.038	<0.001	0.003	0.038
Zn	mg/kg	0.5	0	3.7	348	27.40	20.9	24.17	12.45	<0.001	0.1	0.5
LOI	wt. %	0.1	0	0.34	16.4	2.34	1.8	1.854	0.72	<0.001	0.027	0.077
pH		0.1	0	3.7	7.6	5.83	5.8	0.358	0.2	<0.001	<0.001	<0.001

^a Minimum, maximum, mean, median, standard deviation, median absolute deviation (MAD), and *p*-values of a Kolmogorov–Smirnov test for normality of the original, the log-transformed (*p*_{ln}), and the Box–Cox transformed (*p*_{Box-Cox}) data.

^b C-horizon podzol samples, fraction <2 mm. No extension: aqua regia extraction; _INAA: analysed by instrumental neutron activation analysis; _XRF: analysed by X-ray fluorescence; DL: detection limit, <DL: % of samples below the detection limit.

^c The data are taken from Reimann et al. (1998).

in the till is mainly of local origin, although some cobbles and boulders may have been transported over several kilometres. The moraine formations in the study area are mostly gravelly and sandy tills, locally hummocky moraines occur.

In-situ occurrences of weathered bedrock are distinctive features of the bedrock in Finnish Lapland (Hirvas, 1991). The weathering products are mixed with till to a varying degree, thus demonstrating that weathering took place before the last glaciation. The chemical composition of the gravelly, weathered bedrock differs little from that of the fresh bedrock. The thickness of the weathered bedrock varies, being only a few metres in most places, but tens of metres in others.

In valleys, between moraine formations and outcrops, there are areas covered by peat formations, which are generally not very thick. In the river valleys, fluvial deposits dominate. In many places, there are eskers and sorted, ice-marginal formations. In the northwestern part of northern Finland some aeolian deposits can be found.

Climatic conditions varied during Holocene (post-glacial) times. In the climatic maximum, during the latter part of the Atlantic Period (6–4.6 ka BP), pine forests extended over most of the Kola Peninsula (Khotinskiy, 1984), leaving a narrow zone of coniferous middle taiga (spruce) and a strip of tundra along the NE coast of the peninsula.

2.4. Sampling, sample preparation and analyses

A detailed description of sample site selection criteria and sample methods is given in Äyräs and Reimann (1995) and in Reimann et al. (1998). In short, complete podzol profiles were dug at carefully selected sites and about 2 kg of the C-horizon material was sampled. A field duplicate was taken at every 15th site. The randomised samples were air dried and subsequently sieved through 2 mm nylon screening, and the <2 mm fraction was retained for analysis.

Analytical procedures and all analytical results are detailed in Reimann et al. (1998). Quality control procedures followed the methods suggested in Reimann and Wurzer (1986). In short, after insertion of a project standard at a rate of 1 in 20 and sample duplicates at a rate of 1 in 15, a 2 g subsample of the <2 mm fraction of the C-horizon samples was digested in aqua regia (3:1 HNO₃/HCl) at 90°C at Geological Survey of Finland's (GTK) laboratory. The solutions were analysed by ICP-AES for 32 elements (Niskavaara, 1995) and by graphite furnace atomic absorption spectrometry (GFAAS) for Ag, As, Cd and Pb. A second aliquot was analysed after pre-concentration using reductive co-precipitation (Niskavaara and Kontas, 1990) for Bi, Sb, Se and Te by GFAAS. Aqua regia extraction will not yield total element concentrations. The determined amounts will be strongly influenced by differences in mineralogy between

the samples, and mostly reflect secondary geochemical processes like weathering, scavenging of elements by Fe-oxides/hydroxides and/or the amount of sulphides and clay minerals in the individual samples.

Additionally, the samples were analysed for major elements (Al, Ca, Fe, K, Mg, Mn, Na, P, Si, Ti) by XRF in the Norwegian Geological Survey's (NGU) laboratory, and for more than 30 other elements by instrumental neutron activation analysis (INAA) at Activation Laboratories in Canada. These 2 techniques result in total element concentrations.

2.5. Data analysis

For rapid factor analysis and mapping of the results a combination of S-PLUS [Mathsoft — <http://www.splus.mathsoft.com/>; Venables and Ripley (1997) for factor analysis] and DAS[®] [Dutter et al. (1992) for mapping] was used. Class selection for the black and white symbol maps was undertaken by box plot analysis as proposed by Kürzl (1988).

3. Results

Table 1 summarises the variables and analytical results (minimum, maximum, mean, median, standard deviation, median absolute deviation (MAD) of the data, and *p* values of a Kolmogorov–Smirnov test for normality (see Afifi and Azen, 1979) of the original, the log-transformed, and the Box–Cox (Box and Cox, 1964) transformed data.

Factor analysis is a very data-sensitive technique, a fact that is often neglected. A careful univariate analysis should be carried out for any data set prior to its being used for factor analysis. Some users may be surprised — is not factor analysis chosen to “simplify” data analysis? This paper will start with a discussion of the prerequisites, and determine whether or not the data set fulfils the statistical requirements for undertaking a factor analysis.

3.1. Factor analysis versus principal component analysis

In geochemical textbooks and publications there is much confusion as to what is “factor analysis” and what is “principal component analysis” (PCA). In addition, both methods can be carried out based on either the correlation matrix or the covariance matrix. The data entered can be transformed and/or standardised. To further confuse the issue, the data could be first standardised and then transformed or first transformed and then standardised. The choice and sequence of these steps will have a major influence on the results. This will be discussed below.

The major difference between factor analysis (FA) and PCA is that PCA performs a transformation of the

data with no statistical assumptions whereas factor analysis assumes a statistical model with certain prerequisites. PCA accounts for maximum variance of all variables, while FA is based on the correlation structure of the variables. The model of factor analysis allows that the common factors do not explain the total variation of the data. This implies that factor analysis allows for the existence of some unique factors that have a completely different behaviour than the majority of all other factors. Thus unusual variables will not enter the common factors. PCA in contrast will always show the total structure in the data (all variables are “forced” into the result). In practice, this means that factor analysis is better suited to detect common structures in the data. In geochemistry chances to detect common processes determining element behaviour are thus better when using factor analysis.

In factor analysis, 2 main methods exist for extracting the common factors: principal factor analysis (PFA) and the maximum likelihood method (ML). PFA works in principle like PCA but with a reduced correlation or covariance matrix. Only the common structure of all variables but not any special behaviour (uniqueness) of each single variable is thus used. ML, in contrast, uses a complicated statistical optimisation procedure to extract the factors.

3.2. *Mixing major, minor and trace elements and overcoming differences in amount of variation*

In multi-element analysis of geological materials one usually deals with elements occurring in very different concentrations. In rock geochemistry, the chemical elements are divided into “major”, “minor” and “trace” elements. Major elements are measured in % or tens of %, minor elements are measured in about 1% amounts, and trace elements are measured in ppm, or even ppb. This may become a problem in multivariate techniques considering all variables simultaneously because the variable with the greatest variance will have the greatest influence on the outcome. Variance is obviously related to absolute magnitude. As one consequence, one should not mix variables quoted in different units in one and the same multivariate analysis (Rock, 1988). Transferring all elements to just one unit is not an easy solution to this problem, as the major elements occur in much larger amounts than the trace elements. Possible solutions to the problem include transformation and/or standardisation.

Standardisation to zero mean and unit variance of the raw data makes little sense in geochemistry because it is known that the distributions are strongly skewed. Thus the data should be transformed first before a decision on standardisation is taken. Log-transformation is most widely used to approach considerably more homogenous ranges. Previously, Bartlett (1947) and Bartlett and

Kendall (1946) noted that log-transformation aids in obtaining homogeneity of variance. It will, however, emphasise the influence of variables with a large variation. For example, in the Kola data set, Ag has a range (defined as maximum–minimum) of 0.119 in the raw data, while Si has a range of more than 200,000. After a log transform Ag has a range of 7.1 while that for Si is now 0.86. If this effect of emphasising the influence of variables with a large variation is wanted, a log-transform of the data will suffice. If this effect is not wanted the data have in addition (and after the log-transformation!) to be standardised to zero mean and unit variance. Standardisation is equivalent to using the correlation matrix and not the covariance matrix in factor analysis. This approach would be the standard choice of most statisticians and is implemented automatically in many software packages (e.g. S-PLUS).

3.3. *Closed data (complete subcompositional independence)*

The problem of entering statistical analysis with “closed number systems” has been much discussed in the literature (e.g. Butler, 1976; Le Maitre, 1982; Woronow and Butler, 1985; Aitchison, 1986). All compositional data expressed as % (or ppm), which sum up to a constant value (e.g. 100 wt.% — major elements analysed by XRF) are closed data. Given values for $N-1$ variables the N th value is automatically known. This has quite serious consequences in correlation analysis (on which factor analysis is based) that are often neglected. For example, a negative correlation is less significant and a positive correlation is more significant in a closed array than in an open array (Rock, 1988). High “artificial” internal correlations between variables can lead to ill-conditioned matrices and give unstable and even erroneous results (Rock, 1988). Another important point is that for closed number systems the correlation matrix has not full rank. The estimation of the factor scores by the usual regression method is thus not possible since the inverse of the correlation matrix is needed. Data closure cannot be overcome by any conventional data-transformation method.

3.4. *Normal distribution*

Before carrying out a classical factor analysis, it should generally be tested whether or not all variables have a normal distribution. Just as for many other statistical techniques, factor analysis is very sensitive to non-normally distributed data (Pison et al., 1999). It is now well known amongst geochemists that regional geochemical data practically never show a normal distribution (Reimann and Filzmoser, 2000). ML requires not only a normal distribution for all the variables entered but also a multivariate normal distribution. When using PFA a

normal distribution is not a must, but this method is based on the correlation or covariance matrix and these are strongly affected by non-normally distributed data and the presence of outliers (see below). In many published examples of the use of factor analysis, it is neglected that regional geochemical (and environmental) data almost never follow a normal distribution. Continuing with factor analysis in such a case must lead to biased results. A proper normalisation procedure must first be applied to the data to approach a normal distribution. In many cases a logarithmic transformation is used, and many geochemical textbooks state that geochemical data usually follow a log-normal distribution. It remains somewhat unclear from where the authors take this observation without carrying out the proper statistical tests. Vistelius (1960) pointed out that many of the apparently log-normal distributions observed in geochemical data sets come about from combining data from several parent normal distributions. Garrett et al. (1980) demonstrate this in the context of regional geochemical reconnaissance data, where, a priori, geochemists accept that the data are drawn from different lithological populations, and hopefully mineralisation, and have been influenced by a variety of secondary processes. The question whether to enter factor analysis with the original or somehow transformed data can easily be answered. All entered variables should come as close to a normal distribution as possible (Reimann and Filzmoser, 2000).

For the test data, Table 1 demonstrates that not a single variable follows a normal distribution. After log-transformation most variables are still not normally distributed (Table 1). Many different wide spread transformations were investigated (e.g. square root, log10, logit, double logarithmic including scale transformation) and none resulted in a normal distribution for more than 20% of all variables (Reimann and Filzmoser, 2000). If it is essential to approach normal distributions a Box–Cox transformation (Box and Cox, 1964; Howarth and Earle, 1979) is one of the few remaining solutions. Table 2 shows that after a Box–Cox transformation more than 2/3 of all variables approach normality. This would thus be a good data transformation when working with geochemical data and statistical methods that require a normal distribution. Unfortunately, there are few statistical software packages that allow for a Box–Cox transformation, and the second choice is then to log-transform all data prior to entering factor analysis. Thus the much more wide spread log-transformation was employed. Factor analysis was, however, also tested with Box–Cox transformed and standardised data. In general, the first 3 factors stayed comparable, major differences could be found in higher order factors. Although results were different, they were not better or more easily interpretable than the results obtained with the same data following a log-transformation.

Table 2
Summary statistics for different methods of factor analysis^a

Method	F1	F2	F3	F4	Total	Residuals	
						m	s
XRNA PFA Q4	42	15	12	4	73	0.213	0.362
XRNA PFA O4	37	17	12	5	71	0.21	0.33
XRNA PFA P4	40	21	9	6	76	0.007	0.083
XRNA PFA V4	40	21	9	6	76	0.005	0.084
XRNA ML Q4	42	15	12	4	73	0.15	0.309
XRNA ML O4	36	16	14	5	71	0.144	0.282
XRNA ML P4	40	21	8	6	75	0.009	0.09
XRNA ML V4	40	21	8	6	75	0.005	0.091

^a Columns F1–F4 represent the explained variance for each factor. “Total” gives the sum of the explained variance in %. This should be as high as possible. “Residuals” represent the unexplained part of the factor analysis, m: mean, s: standard deviation over all elements of the residual matrix. The mean should be close to 0. The standard deviation should be as small as possible. XRNA: only results of total analyses either with XRF or INAA used. PFA: principal factor analysis, ML: maximum likelihood method, Q: quartimax, O: oblimin, P: promax and V: varimax rotation.

3.5. Data outliers

Regional geochemical data sets practically always contain outliers. The outliers should not simply be ignored but they have to be analysed because they contain important information about data quality and unexpected behaviour in the region of interest. In fact, finding data outliers that maybe indicative of mineralisation (in exploration geochemistry) or of pollution (in environmental geochemistry) is one of the major aims of any geochemical survey. Outliers can have a severe influence on factor analysis, since the parameter estimates are based on the correlation or covariance matrix (Pison et al., 1999). Outliers should thus be removed prior to entering a factor analysis (or statistical methods able to handle outliers should be used). This is rarely done. One reason maybe ignorance, another that with the outliers removed the expected or wanted “obvious” results will no longer be revealed in the factors. Finding data outliers is not a trivial task, especially in high dimensions. One way of identifying outliers is to compute Mahalanobis distances (Garrett, 1989) or better Mahalanobis distances based on robust estimates of location and scatter (Rousseeuw and Van Zomeren, 1990). A more elegant way to reduce the impact of outliers is to apply robust factor analysis (Filzmoser, 1999; Pison et al., 1999). The aim is to estimate the parameters in the factor analysis model to fit the *majority* of data points, contrary to classical (least squares) estimation

where *all* data values, including the outliers, are fitted. When using a robust factor analysis the outliers can be identified and interpreted by looking at the residuals of the robust fit.

There is a further problem that often occurs when working with geochemical data: the detection limit problem. For some determinations a proportion of all results are below the lower limit of detection of the analytical method. For statistical analysis these are often set to a value of half the detection limit. However, a sizeable proportion of all data with an identical value can seriously influence an estimate of correlation. For the study dataset several variables had more than 25% of the data below detection. It is very questionable whether or not such elements should be included at all in a factor analysis. Unfortunately it is often the elements of greatest interest that contain the highest number of missing data (e.g. Au) — the temptation to include these in a factor analysis is thus very high. Note that in addition to the elements/parameters given in Table 1 the elements Au, B, Br, Cs, Hg, Ir, Mo, Rb, Se, Ta, Tb, Te and U were also analysed. These were not used for this study because for all these elements a substantial part (> 3–100%) of all data were below the respective limits of detection.

3.6. Inhomogenous data set

Geochemical data sets are often extremely inhomogeneous. Usually they consist of a mixture of different populations. In different subsets antagonistic data structures may exist. In the case of the test data set there exists some knowledge about the processes governing the data structure. The geological map could be used to construct more homogenous subsets of the samples that were collected. If this is done and the correlation matrix for each of these subsets is estimated it becomes at once clear that each of these lithologically based data subsets has a quite different correlation matrix (Fig. 3). Entering factor analysis with such inhomogeneous data will almost invariably lead to unstable results. The existence of such problems can easily be detected in a simple CDF-diagram. Fig. 4 shows 4 examples displaying breaks in the data structure, caused by the presence of different lithologies. This must influence the results of factor analyses. The main result being that factors will be governed by the elements that show high (or low) values in one of the sub-populations. It is predictable that factor analysis will have as one of its main outcomes an elucidation of some of the known lithological units. Such data behaviour can be expected in the vast majority of cases when working with regional geochemical or environmental data. It is doubtful that factor analysis is the correct method of multivariate data analysis in such cases.

The correct data analysis procedure would be to first disaggregate the total data set into more structurally

homogenous subsets, e.g. via a clustering procedure or by the use of already existing knowledge (e.g. the geological map). However, in most cases, and definitely in the case of the test data set, there would not remain sufficient samples in any of these subsets to carry out a meaningful factor analysis.

3.7. Spatial independence

Factor analysis assumes that the data represent random, independent samples from a multivariate distribution. However, geographic variables usually have a spatial dependence, i.e. the observations are correlated. Factor analysis of correlated observations will reflect such correlations, which cannot be removed by a random shuffling of the data points. For a correct statistical investigation, the data should be viewed as a realisation of a stochastic process, which will result in complicated and expensive statistical procedures (see, e.g. Basilevsky, 1994). The data set considered here comes much closer to spatial independence than most data from regional or environmental geochemical investigations because of the low sample density employed, i.e. 1 site per 310 km².

3.8. Dimensionality

One of the first requirements for stable results from a factor analysis is that there are a sufficient number of samples for the number of variables. Different rules have been suggested (Le Maitre, 1982), e.g. $n > p^2 + 3p + 1$ (where n is the number of samples and p the number of variables) — for the 54 variables from Table 1 this gives 3079 samples — but the data set consists of “only” 605 samples. Even if more tolerant rules are used (e.g. $n > p^2$ or $n > 9p$, or just $n > 8p$) the number of samples in this study is rather small in relation to the number of variables. Factor analysis should thus preferably not be entered with the full set of elements. The above cited rules suggest that for a data set consisting of 605 samples factor analysis should not be entered with many more than about 23 variables. Some possibilities of reducing the number of variables are discussed below.

3.9. Factor analysis

Judged by the statistical criteria presented one should probably come to the conclusion that the Kola data set (and most other geochemical data sets) are not suited for factor analysis. There are often too many variables for the number of samples. There is a closed number problem. There are many data outliers, and the data are not normally distributed. Even after a log-transformation they do not show a normal distribution. The data are extremely inhomogeneous. Other techniques, e.g. cluster analysis, should probably be used prior to factor analysis to gain more stable data subsets. This, however,

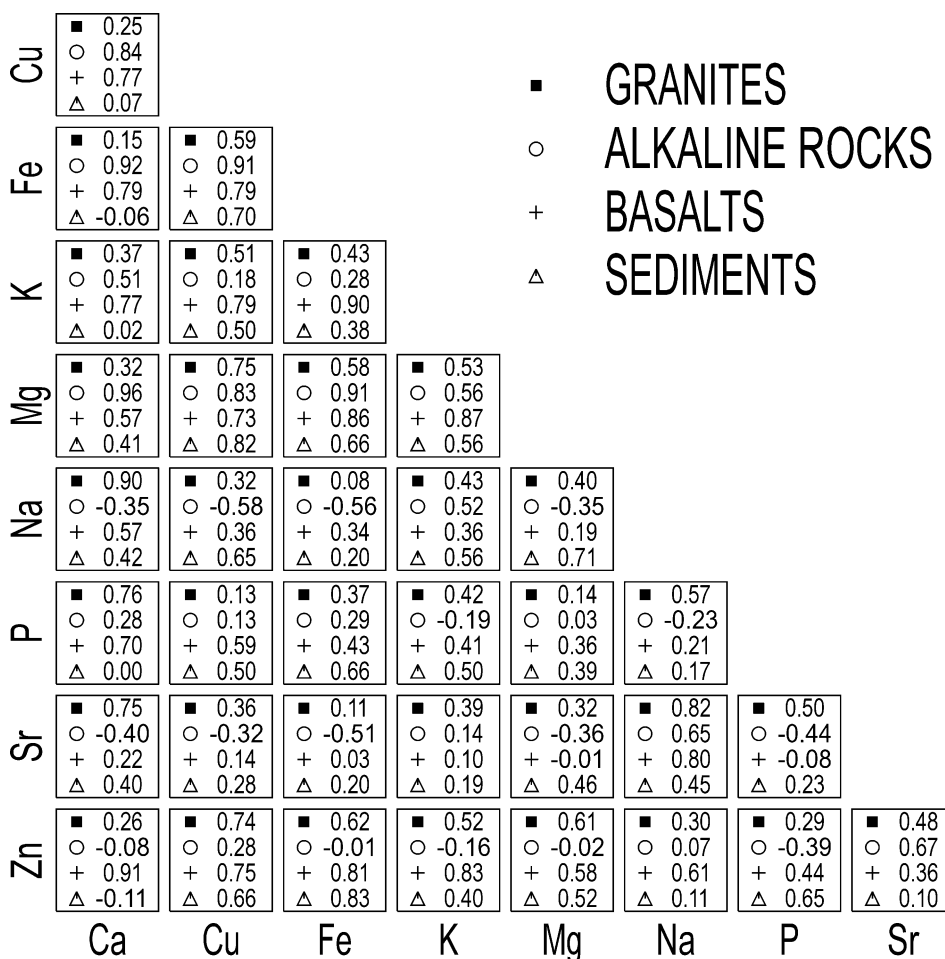


Fig. 3. Pearson correlation coefficients within 4 selected rock units underlying the C-horizon in the survey area for 9 elements (Granites: $n=94$, Alkaline rocks: $n=12$, Basalts: $n=42$, Sediments: $n=68$). Note that the same pair of elements can show positive as well as negative correlation depending on rock type.

adds considerable work, and with the small number of samples in the original data set it would most likely result in too few cases in the resulting subsets to carry out stable factor analyses.

Having clearly stated that the study data set is hardly suited for classical factor analysis, factor analysis is still entered (as is so often done) and the results of different approaches are discussed in view of the issues presented above. To avoid as many as possible of above discussed problems all data were log-transformed and standardised prior to entering factor analysis. This approach approximates normality and gives the same weight to all variables.

3.10. Factor analysis method and rotation

As the next step, the actual method for carrying out factor analysis has to be chosen. The options are the classical PFA or more complicated techniques like maximum likelihood (ML). Is this choice likely to have

a fundamental influence on the results of the factor analysis? Furthermore, the method of factor rotation must be selected: Varimax (Kaiser, 1958), Promax (Hendrickson and White, 1964), Oblimin (Harman, 1976) or Quartimin (Carroll, 1953) are just some examples. Varimax and Promax are orthogonal rotations, i.e. the rotated factors are not correlated, Oblimin and Quartimin are oblique rotation methods, i.e. the rotated factors can be correlated. Most geochemists will likely accept the standard choice offered by the software package being used. The mathematical background is well documented in a number of specialised textbooks (e.g. Harman, 1976; Seber, 1984; Basilevsky, 1994; Mardia et al., 1979; Johnson and Wichern, 1998). In the light of above discussions on the statistical assumptions that need to be met, PFA may be the "safer" method for geochemical applications (Pison et al., 1999).

To ease comparison of different versions of factor analysis, which usually result in long tabulations, one

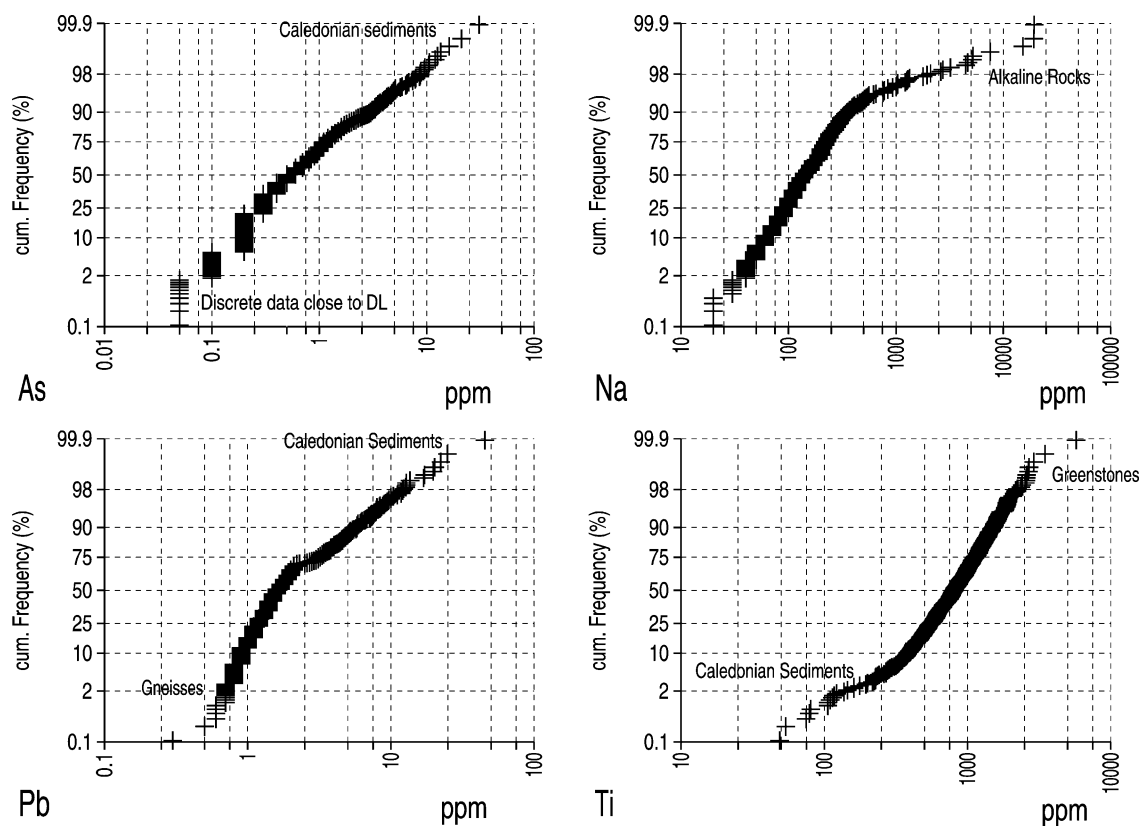


Fig. 4. CDF-diagrams scaled for linearity of lognormal distributions for 4 selected variables demonstrating some inherent problems of the data-set with regards to the suitability for factor analysis. Some of the main reasons (mostly certain lithologies) causing breaks in the diagrams are marked (DL: detection limit).

can plot the resulting factor loadings into a simple XY -graph, where the X -axis is scaled according to the explained variance for the whole data set for each factor. The Y -axis is scaled from +1 through 0 to -1 and shows the factor loadings of the different variables entering each factor. Names of variables with an absolute value of the loadings < 0.3 are not plotted. Fig. 5 shows 8 results of factor analyses using a selection of all the elements where total concentrations were determined either by XRF or INAA techniques. In the first example, 4 factors result in a total explained variance of 76%, F1 contributes 40%, F2 21%, F3 9% and F4 only 6%.

At a first cursory glance, the factors look similar in all 8 cases (Fig. 5). This is not surprising because the factor analyses are based on the same correlation matrix, and, independent of the method chosen, the same elements always will show high correlations. In all 8 tested cases, between 71 and 76% of the total variability is explained by just 4 factors, which were chosen for this comparison because in all cases the 5th (and the following) factors explained less than 5% of the total variability.

When studying the resulting factors more carefully, it is apparent that there are slight differences between the

results. Some elements are exchanging positions between different factors, depending on method and/or rotation. Such slight shifts can result in dramatic changes in the regional distribution displayed in factor maps. Results obtained by the ML-method appear to be very sensitive to factor rotation — F1 and F2 exchange position here between Varimax/Promax and Oblimin/Quartimin rotations. For geochemical reasons none of the 8 results can be selected as the best. There are a number of statistical parameters that can be used to judge the quality of a factor analysis (Basilevsky, 1994). Unfortunately in practice it is often difficult to decide which of these parameters is really indicative of a “useful” result. Criteria for the quality of the results can be the explained variance of each single factor (should be high, e.g. $> 5\%$) and the total explained variance (should reach a certain cut-off, e.g. $> 70\%$). Another measure is the simple structure of the rotated factors: more separated factors should have an easier interpretation. According to Table 2, PFA with an orthogonal rotation performs best.

In a further step, all resulting factor maps were plotted ($8 \times 4 = 24$ maps). For several of the principally different

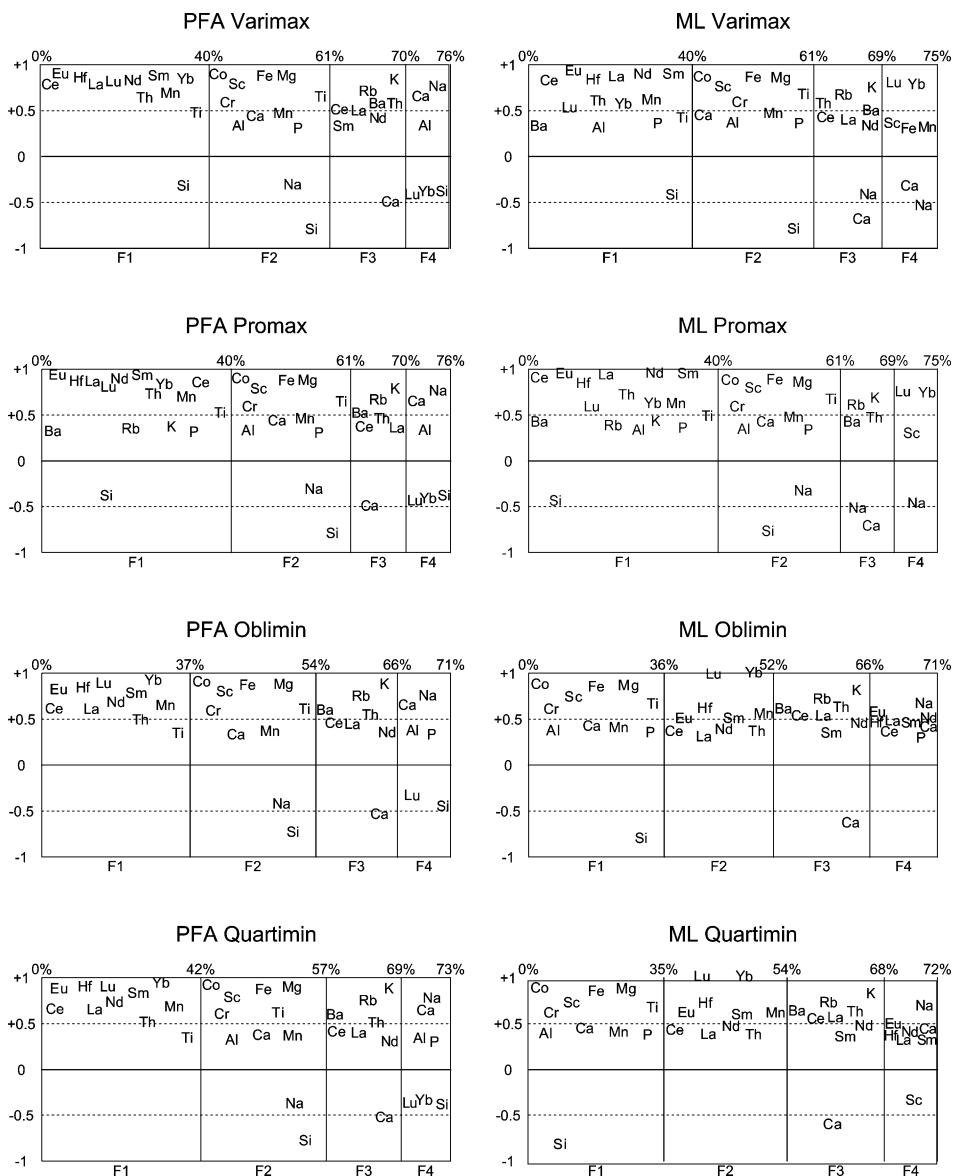


Fig. 5. Factor loadings for the first 4 factors (F1–F4) extracted with 8 different methods of factor analysis from the same data set. Only elements analysed with XRF or INAA were used here. Method of factor analysis (PFA or ML) and method of factor rotation is given on top of each graphic.

methods, the same procedure was carried out for 3, 5 and 6 factors giving in total over 100 maps (not shown). Careful investigation of all these showed that PFA with Varimax rotation gave the most “stable” maps. When using the ML method, one factor more or less could result in completely different maps. The different rotation methods resulted in completely different regional distributions of the factors. In general, the ML-maps were rather noisy. For geochemical data, PFA, as the method making the least statistical assumptions, is thus

the best choice. Orthogonal rotations (e.g. Varimax) are preferable to the more complicated oblique rotation methods. Fortunately this will be the default selection in most statistical packages.

With the many choices at hand, geochemists using factor analysis will invariably be tempted to choose the one result that is most interpretable on the basis of their knowledge in general, and the study area in particular. This may result in some really important but unexpected results being neglected.

3.11. Number of factors extracted

There are different procedures to determine the “optimum” number of factors to be extracted from factor analysis, including, a number of statistical tests (see Basilevsky, 1994). However, the procedures most often used are more of the “rule of thumb” type, e.g.:

- To select as many factors as there are eigenvalues larger than the average;
- To select enough factors to reach a certain pre-selected explained variance (e.g. > 70%); and
- To use the scree plot (Cattell, 1966) where the number of factors is plotted against explained variance (see Fig. 6 for an example) and the cut-off is chosen at the point where the function flattens out.

It is also possible to enter factor analysis with a pre-selected number of factors (although, in most cases, one may argue that if one has enough knowledge to pre-select the number of factors one probably does not need to use factor analysis). Results can drastically change with the number of factors extracted. This effect is especially pronounced when factor analysis is entered with only a few variables. Experience from studying many factor maps suggests that it might in general be better to extract a low number of factors. The scree plot is probably the best of the “simple” techniques for determining the optimum number of factors, although the answer is again not a clear one.

3.12. Selection of elements

Following decisions on data transformation (log), standardisation, selection of method (PFA) and factor rotation (Varimax) the influence of the number of elements entered into factor analysis was studied. Because factor analysis is mostly used as a method to reduce

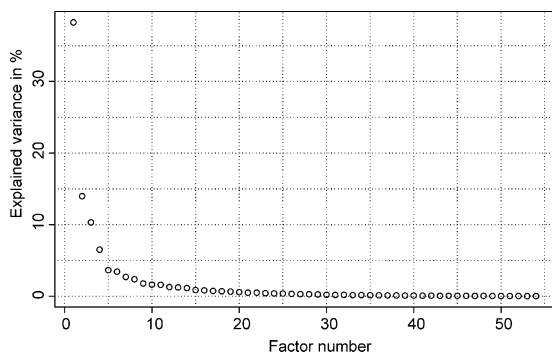


Fig. 6. Scree plot for a factor analysis (PFA, Varimax rotation) with all elements/parameters (54), demonstrating the very minor increase in the explained variance with an increasing number of factors.

dimensionality in a data set it will normally be entered with all available variables. Thus, here this test was started with all 54 elements/parameters named in Table 1 ($605/54 = 11.4$ — see discussion on dimensionality). The scree plot shown in Fig. 6 is constructed with the results of factor analysis using all these variables. The resulting curve suggests that 4, 5, 6 or 7 factors could be extracted. All these possibilities were plotted (factor loadings and maps) and finally the results for 6 factors, which explain 76% of the total variation, were selected. Only these results are shown in Fig. 7 and Table 3. Fig. 8 gives the 6 factor maps, which all show clear regional structures.

Geochemically F1 is dominated by the rare earth elements (REEs), high concentrations of these occur in the soils overlying the alkaline intrusions, the granulite complex and parts of the Caledonian sediments (Fig. 2). In the F1 map, these are reflected with the highest values. Most interesting is the linear continuation of this zone from the northeastern end of the granulite belt towards the coast of the Barents Sea (Fig. 8). F2 shows a “mafic” association of elements. High values in the map reflect the greenstone belts (Figs. 2 and 8) and parts of the granulite complex containing mafic layers. The third factor associates some surprising elements. LOI (loss on ignition — content of water/organic material in the samples) plays an important role and suggests that secondary geochemical processes dominate this factor. High factor scores are visible along the coast of the Barents Sea in Northern Norway and parts of Russia, where very thin soil profiles rich in organic material are observed. A second area with high scores occurs on top of the alkaline intrusions, from where a trend extends in a westerly direction towards the alkaline intrusions near Kovdor and the Sokli Carbonatite in Finland. It is very likely that two different processes are reflected here in the same factor: admixtures of organic material even in the C-horizon along the coast (different mineral weathering conditions in coastal areas could be an alternative explanation) and a high proportion of hydrous minerals occurring in the alkaline intrusions. Factor 4, dominated by Na and Sr, clearly separates the sedimentary sequences in Northern Norway from the alkaline intrusions in Russia. The very large area with high values surrounding these intrusions is most likely indicative of extensive hydrothermal alteration accompanying the emplacement of the intrusions. The heavy REEs and several “mafic” elements enter Factor 5. Although the map shows a very clear regional structure it cannot be explained on the basis of current geological knowledge. Phosphorus and Ca are the most important elements in F6, and this element association points to Ca-phosphate, apatite, determining the values in this factor. The map is rather noisy, however, the most interesting structure is a half-circular area with high values running from Murmansk to the Norwegian Russian border and

Table 3
Factor loadings of a factor analysis (PFA, Varimax rotation) carried out with all 54 available variables^a

	F1	F2	F3	F4	F5	F6	Comm.
Ag	0.22	0.11	0.44	0.13	0.08	0.14	0.3
Al	0.35	0.43	0.7	0.28	-0.12	-0.07	0.89
Al_XRF	0.3	0.22	0.04	0.49	0.14	-0.14	0.42
As	0.16	0.17	0.72	-0.18	-0.05	0.05	0.6
Ba	0.47	0.4	0.38	0.39	0.07	-0.06	0.69
Ba_INAA	0.64	-0.07	-0.05	0.06	0.06	-0.29	0.51
Be	0.54	0.07	0.71	0.23	-0.11	-0.06	0.86
Bi	-0.02	0.06	0.7	-0.26	-0.01	-0.03	0.57
Ca	-0.06	0.17	0.01	0.53	0.65	-0.06	0.73
Ca_XRF	-0.28	0.26	-0.39	0.65	0.32	-0.16	0.86
Cd	0.28	0.22	0.56	0.31	0.06	0.2	0.59
Ce_INAA	0.91	0.09	0.28	0.09	0.09	0.06	0.94
Co	0.1	0.77	0.55	-0.07	0.11	0.02	0.92
Co_INAA	-0.03	0.87	0.16	0.17	0.12	0.24	0.89
Cr	0.01	0.8	0.26	0.12	-0.1	-0.23	0.78
Cr_INAA	-0.15	0.76	-0.07	0.28	-0.18	0.04	0.73
Cu	0.07	0.69	0.44	0.09	0.15	0.1	0.72
Eu_INAA	0.82	0.17	0.15	0.34	0.13	0.22	0.91
Fe	0.27	0.66	0.56	-0.25	0.14	-0.05	0.9
Fe_XRF	0.33	0.75	0.14	0.21	0.08	0.4	0.89
Hf_INAA	0.78	-0.05	0.16	0.17	-0.04	0.18	0.7
K	0.33	0.27	0.42	0.14	0.17	-0.14	0.43
K_XRF	0.59	-0.31	0.37	-0.24	-0.12	-0.27	0.72
La	0.8	0.04	0.37	0.18	0.16	-0.2	0.88
La_INAA	0.93	0.08	0.21	0.12	0.06	0	0.94
Li	-0.01	0.2	0.79	-0.12	0.13	-0.23	0.76
LOI	0.23	0.26	0.75	0.1	-0.12	0	0.71
Lu_INAA	0.71	0.28	0.09	-0.03	0	0.55	0.89
Mg	0.04	0.7	0.57	-0.03	0.19	-0.08	0.86
Mg_XRF	-0.05	0.84	0.02	0.28	0.05	0.27	0.86
Mn	0.25	0.27	0.77	0.06	0.16	0.15	0.77
Mn_XRF	0.33	0.39	0.24	0.48	0.18	0.48	0.81
Na	0.24	0.04	0.07	0.84	0.19	-0.02	0.8
Na_XRF	-0.22	-0.19	-0.28	0.51	0.25	-0.42	0.66
Nd_INAA	0.87	0.14	0.19	0.12	0.08	0.08	0.84
Ni	-0.01	0.77	0.42	0.14	-0.09	-0.04	0.81
P	0.1	0.09	0.03	0.08	0.86	0.01	0.77
P_XRF	0.35	0.12	0.12	0.3	0.78	0.09	0.86
Pb	0.42	-0.05	0.67	-0.24	0	0.06	0.69
pH	0.02	0.08	-0.03	0.13	0.01	0.06	0.03
S	0.23	0.18	0.53	0.12	0.01	0.06	0.38
Sc	0.18	0.76	0.35	-0.15	0.12	-0.09	0.77
Sc_INAA	0.14	0.83	-0.14	0.01	0.12	0.35	0.87
Si	0.02	-0.02	-0.05	0.21	-0.13	0.07	0.07
Si_XRF	-0.29	-0.47	-0.31	-0.57	-0.17	-0.09	0.76
Sm_INAA	0.89	0.21	0.2	0.14	0.12	0.2	0.95
Sr	0.4	-0.06	0.37	0.7	0.18	-0.02	0.83
Th_INAA	0.83	0	0.27	-0.22	-0.08	-0.11	0.82
Ti	0.28	0.5	0.2	0.32	0.2	-0.42	0.69
Ti_XRF	0.52	0.55	0.25	0.2	0.1	0.3	0.77
V	0.22	0.82	0.24	0.03	0.16	-0.19	0.83
Y	0.73	0.09	0.45	0.1	0.15	0.1	0.78
Yb_INAA	0.71	0.26	0.11	-0.01	0	0.56	0.9
Zn	0.38	0.27	0.79	0.09	0.09	0.08	0.86

^a Comm.: Communality, or that part of the variance explained by the common factors. A high value (e.g. >0.5) indicates that this variable is well explained by the factor model.

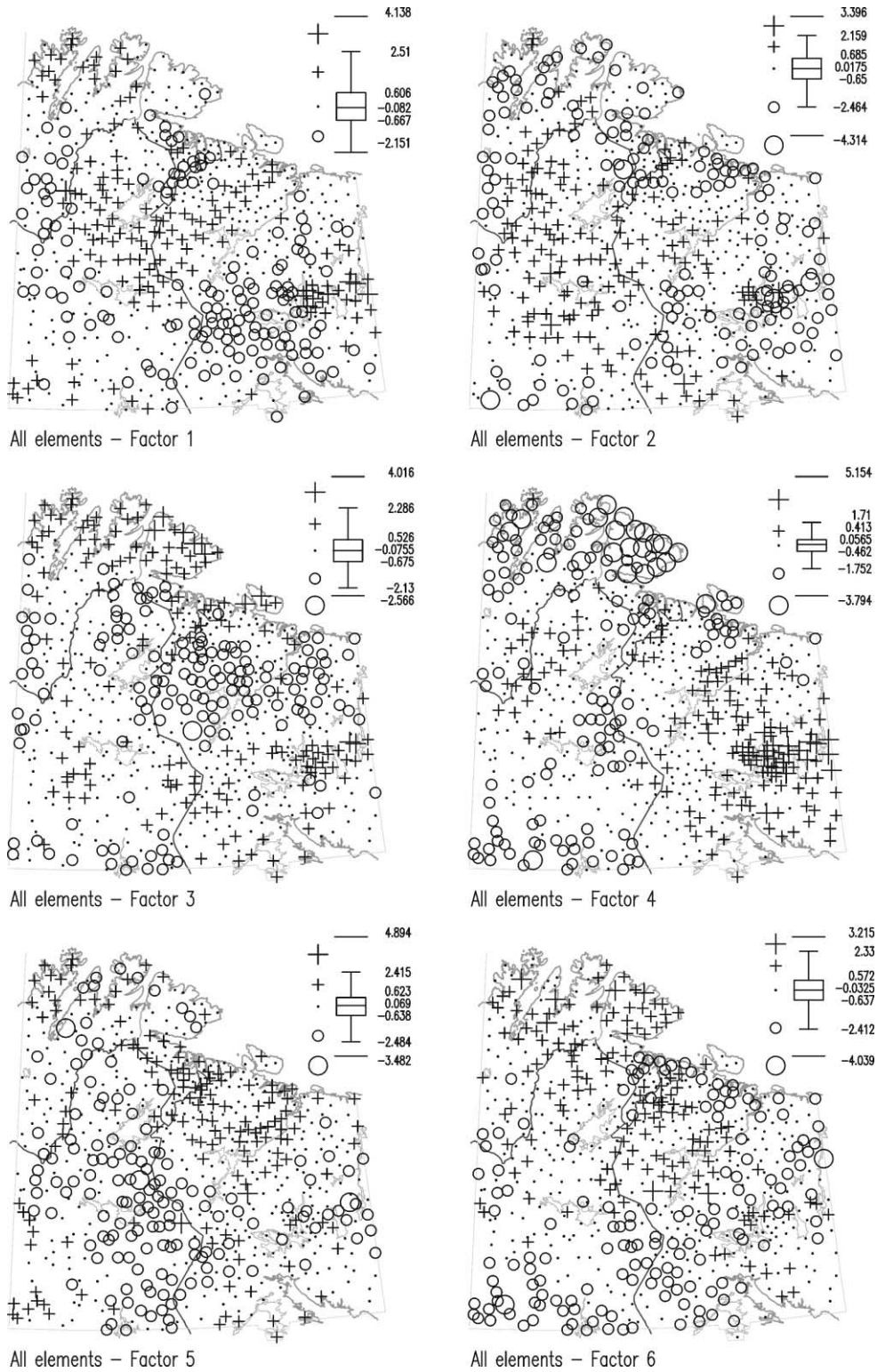


Fig. 8. Factor score maps for the 6 factors shown in Fig. 7: for locations, see Fig. 1; for general geology of the area Fig. 2.

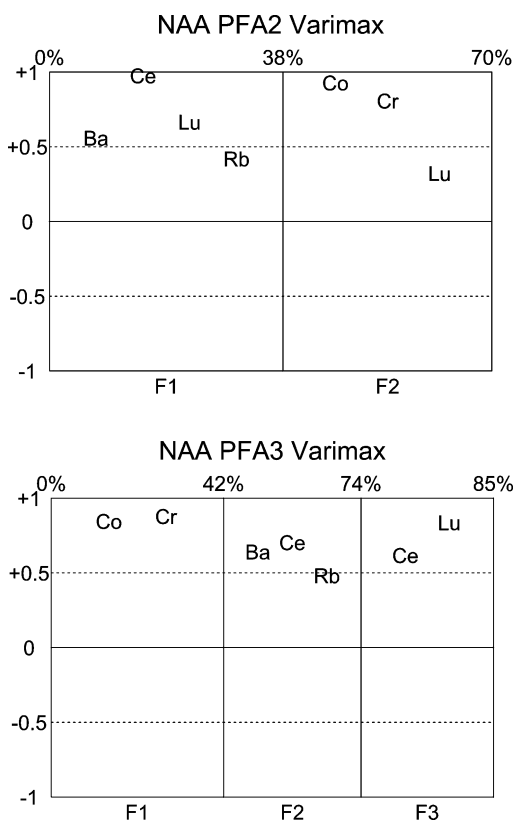


Fig. 9. Difference in factor loadings depending on whether 2 or 3 factors are extracted. NA: analyses with INAA.

addition, geochemical reasoning (e.g. geochemical associations and/or pathfinder elements for different types of ore deposits) was used to select further sub-sets of variables. In geochemistry, the selection of elements entered will in practice most often arbitrarily depend on which elements have been analysed.

As the main result of this investigation, again leading to hundreds of factor maps, it can be stated that the selection of elements with which factor analysis is entered has a most important effect on the results. Just one element more or less can give very different factor solutions and maps. This is rather disturbing when considering that the choice of elements analysed often depends quite arbitrarily on the multi-element package a laboratory is offering, detection limits reached and price rather than on good geochemical reasoning. In general, it was easier to explain results obtained with a factor analysis when the number of variables entered is small. Structures in the maps are often much “sharper” than in the maps where very many elements were employed (compare also the maps in Fig. 8 with those in Fig. 10). Surprisingly, the most general criteria for selection of variable sub-sets (i.e. “total” and “aqua regia extracted” elements separated, “major”, “minor” and “trace” elements analysed separately) gave the most convincing results.

None of the variable selections according to geochemical criteria resulted in better factors or new interesting features in factor maps. The most likely reason for this is the data inhomogeneity issue documented above.

A last word of caution is necessary with regard to the number of factors extracted. Fig. 9 shows an example where factor analysis was entered with a very limited number of elements (Ba, Ce, Co, Cr, Lu and Rb) all analysed by INAA (total concentrations). It was possible to extract 2 or 3 factors, the first version explaining 70% of the total variance, while 3 factors explain 85%. Judged by statistical reasoning the second result is by far superior. In both cases the elements entering the different factors are fully acceptable for the geochemist. However, when these factors are mapped, 2 factors result in two very informative maps (not shown but comparable to those shown in Fig. 10A and B) revealing important and new structures in the data. Mapping 3 factors results in one informative map and 2 maps that could not be interpreted (not shown).

3.14. A collection of the most interesting results

Fig. 10 presents a collection of factor maps. These are the most interesting results of all the different approaches tested. The first two maps come from an extended version of the above example of factor analyses with 10 elements (instead of 6 as above), all analysed by INAA. In both maps very clear and interesting regional structures with sharp boundaries emerge. F1 (Ce, Lu, Ba) marks the alkaline intrusions, the northern half of the sediments along the Norwegian coast and parts of the granulite belt in Finland (compare also with Fig. 8, F1). The most unusual feature is a clear linear zone (marked in grey) extending from the granulite belt towards the coast of the Barents Sea, probably marking a lineament with numerous alkaline intrusions (Fig. 10A). The second map (Fig. 10B) shows a linear anomaly along the Ura Guba zone (grey) — a very exciting feature for Cu–Ni deposit exploration in the study area. In addition, the well known greenstone belts are marked by high values. This is also one of the very few maps where both, the Cu–Ni deposits near Monchegorsk and the Pechenga deposits near the Norwegian Russian border are indicated by high values. The features shown in Fig. 10A and B were detected with slight differences in resolution in a number of different factor analyses with varying element combinations of REEs (Fig. 10A) and elements indicative of mafic rocks (Co, Cr, Fe, Mg, Mn, Sc, Ti — F2 — Fig. 10B).

Another large-scale linear feature was observed in 3 maps resulting from 3 different factor analyses (only two shown: Fig. 10C and D). It appears in F4 (Al) of a factor analysis carried out with the major elements as determined by XRF (total concentrations of Al, Fe, Ca, K, Mg, Mn, Na, P, Si, Ti). The area marked in grey in

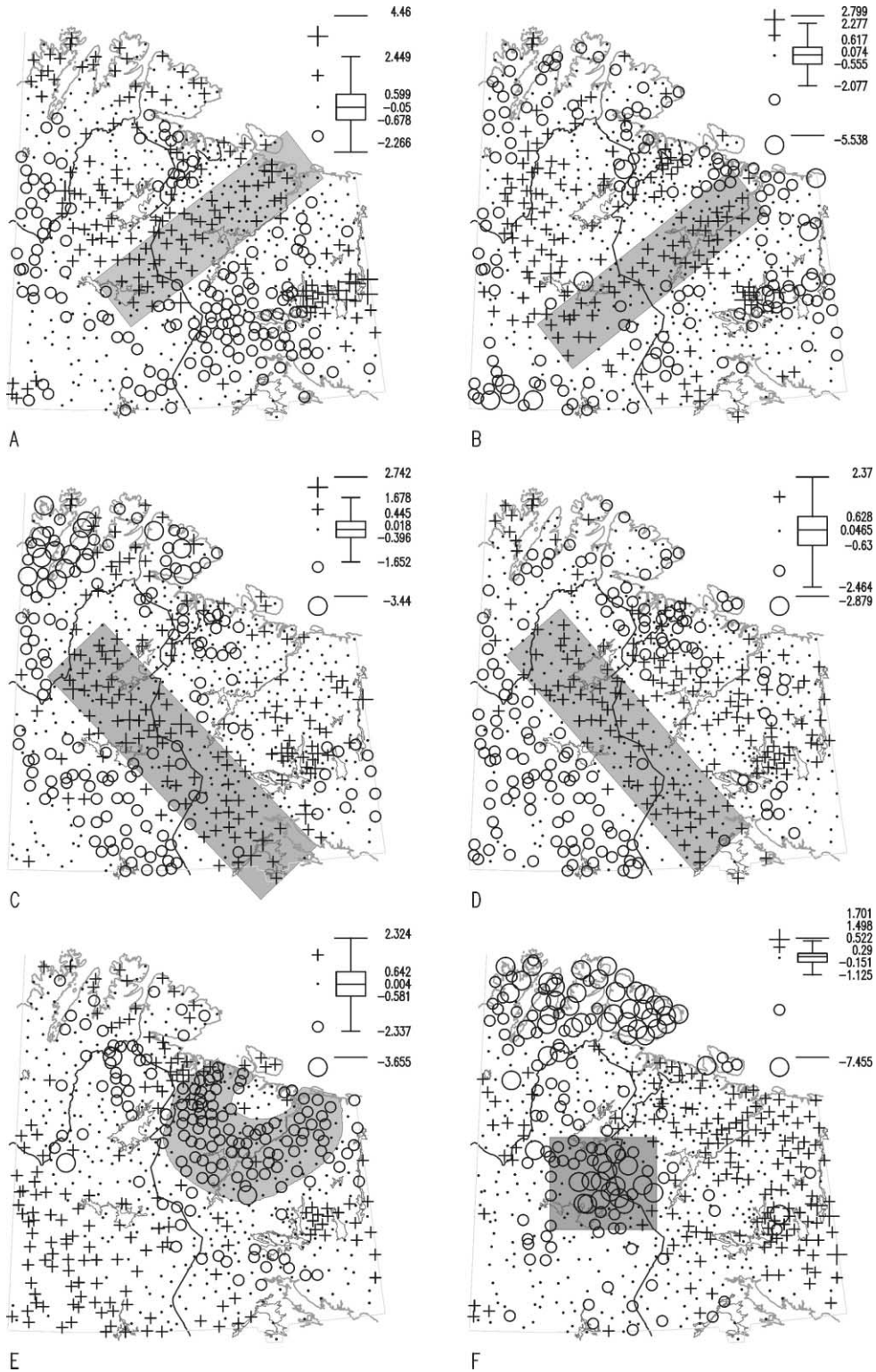


Fig. 10. Maps of factor scores showing the most interesting data structures detected in different factor analyses (see text for explanations).

Fig. 10C cuts several geological units, and is most likely indicative of alteration processes related to a deep-seated fault. It was revealed again in a factor analysis carried out with all those elements extracted by aqua regia as F6, dominated by K and Ba, showing the same structure (Fig. 10D).

A circular structure running from Murmansk to Kirkenes became visible in F5 when all elements analysed were used for factor analysis (Fig. 8). A similar structure appears in two more factor maps. In Fig. 10E, the distribution shows F5 (of 6) of a factor analysis entered with all elements analysed by XRF and INAA. Here, the circular structure reflects low values of Rb and K. When entering the same elements but extracting only 4 factors, the same feature becomes visible via high values of Ca, Na and Al (not shown). The same factor analysis identified another interesting area characterised by exceptionally low values in the centre of the map (Fig. 10F). This feature was visible in quite a number of different factor analyses whenever Ca and Na dominated a factor. One could assume that it might in some way be related to feldspar weathering or to alteration processes related to the emplacement of the Sokli Carbonatite. In contrast, a rather large area in Russia is marked by positive loadings in these maps. A regional scale low marks also the sedimentary rocks on the Varanger Peninsula in Fig. 10F (compare with the geological map, Fig. 2).

4. Conclusions

The worked example using the C-horizon soil data of the Kola project demonstrated that even when neglecting all prerequisites for a sensible factor analysis, interesting results may emerge that can be interpreted based on geochemical reasoning. However, most of the results are governed by regional geology and could have been predicted using pre-existing information. Even the more informative results presented in Fig. 10 can all be extracted from a small selection of single element maps as presented in Reimann et al. (1998). It is disturbing that a reader (or reviewer) not being unusually experienced with factor analysis or the regional geology of the survey area has practically no chance of judging whether or not these results really make sense and introduce new knowledge.

Should factor analysis then be used when studying regional geochemical data? The answer to this question depends strongly on how, and for what purpose, it is used. Complex polypopulation regional geochemical (or environmental) data are not really suited for factor analysis. In nature the chemical elements show neither a normal nor a log-normal distribution. Different geochemical processes can govern the regional distribution of one and the same factor. It is not sufficient to use just

one method of factor analysis with one set of elements to explain the inherent information of a whole data set. Different choices of parameters result in a multitude of results — some may be useful, others cannot be interpreted. The fact that single elements do not enter any of the factors does not mean that the regional distribution of these elements is uninteresting — for an interpretation these may be the most interesting elements of all. Although, in general, when entering factor analysis with all available elements the resulting factors can be explained by geochemical reasoning, the resulting factor maps will not reveal the same information as is revealed in single element maps or combinations thereof. It must be realised that the majority of regional geochemical data sets will be plagued by the above problems, independent of their provenance (exploration or environmental geochemistry). They are thus not well suited for direct entry into factor analysis. Other techniques (like, for example, cluster analysis) should be used to disaggregate the whole data set into more homogenous data subsets prior to factor analysis. Single element maps will always have to be produced first. A solid univariate data analysis is a pre-requisite for using advanced statistical techniques. Exploratory multivariate analysis, e.g. Chi²-plots (Garrett, 1989) for detecting multivariate outliers and cluster analysis should all be used prior to entering a factor analysis. A careful univariate data analysis of the test data set can be found in Reimann et al. (1998). Tests for normality are presented in Reimann and Filzmoser (2000). A comparison of robust and non-robust techniques of factor analysis is given by Filzmoser (1999).

Factor analysis cannot be used as a proof for the existence of certain processes — it can indicate certain relations and help stimulate ideas, they have to be proven in a different way. Much justified critique of factor analysis as applied to geochemical data is caused by the misuse of the technique. Factor analysis can be used to explore the data for hidden multivariate structures, which then have to be explained by different means. If used correctly, it will not result in a reduction of geochemical maps but rather in additional maps. To objectively judge the quality of these maps is very difficult. In regional geochemical mapping, a good result is probably best indicated by stable geochemical maps, displaying large-scale patterns. The best success is if new patterns, that do not necessarily fit the established geological maps and concepts and that may be difficult to see in single element maps, are revealed. However, even such large-scale patterns may often be difficult to interpret based on current knowledge. They can be used to develop new ideas about processes influencing element distribution on a regional scale or to re-interpret the geological map. To really be able to interpret the results beyond pure speculation may require considerable new fieldwork. Factor analysis may be useful in guiding the field activities into especially interesting areas.

Researchers critical of the use of factor analysis are well justified in stating that it is not very scientific to play with the selection of elements and number of factors extracted until one “finds” an “interesting” result. On the other hand, even all the different results presented here are based on identical data, the data are not being changed to generate these results. Patterns reflected in the maps are thus not “artefacts” of the selection process, but are based on the multivariate behaviour of real data. However, the patterns may reflect the influence of individual samples (outliers), or small groups of samples that perturb the correlations of the variables so that they do not reflect the underlying correlations in the main mass of the data. In such cases, the patterns are artefacts of the constraints of orthogonality, which force lower order factor axes into data space not occupied by actual samples. Variable loadings on these factors, and scores for samples computed on them, will commonly be hard or impossible to interpret.

Geochemists are not likely to stop using factor analysis as a tool to study their data as a result of this paper. Discussions on the advantages and disadvantages of factor analysis applied to geochemical data were published as long as 30 a ago (e.g. Miesch, 1969; Temple, 1978; Kufs, 1979). The results of this paper indicate that there are a number of general rules for the application of factor analysis to regional geochemical data (compare also with Howarth and Sinding-Larsen, 1983):

1. Before entering factor analysis the distribution of each of the variables must be carefully studied. A Box–Cox or a log-transformation may be required to approach normality and to ensure homogeneity of variance to meet the requirements of a least squares based procedure.

2. To enter geochemical raw data, including major, minor and trace elements into factor analysis does not make sense because it can be predicted that the minor and trace elements would have almost no influence on the result. If major, minor, and trace elements are mixed in one the same factor analysis log-transformation may not be sufficient to reach homogeneity of variance. In this case, standardisation to zero mean and unit variance guarantees an equal influence of all variables. Note that many standard software packages automatically carry out this standardisation, which does not allow carrying out a factor analysis based on the covariance matrix.

3. Furthermore any large regional geochemical data set should be subdivided into the most obvious data subsets showing different geochemical behaviour, either based on pre-existing knowledge of the regional geology or based on a cluster analysis prior to entering factor analysis.

4. The safest method for factor analysis with geochemical data is principal factor analysis (PFA). More advanced techniques like maximum likelihood (ML) depend even more on the normal distribution of the

data entered and, given the nature of regional geochemical data, will thus produce misleading results. Note that robust methods of factor analysis were also tested on this data set (e.g. Filzmoser, 1999). Statistically these perform clearly better, however, the geochemical interpretation is not necessarily easier. One reason may be that also robust techniques cannot overcome the problem related to a multimodal dataset. In addition they are still not available in most standard software packages, used by geochemists.

5. Note that data outliers are a characteristic of regional geochemical data — when using non-robust methods of factor analysis these should be recognised and removed prior to entering factor analysis.

6. For factor rotation an orthogonal method should be chosen, e.g. Varimax.

7. The number of factors to extract is very difficult to determine. Most of the existing rules do not really help in practise. The most practical approach is to use a scree plot for guidance, and then try some different numbers of factors and study the results (loadings and maps) in detail, although this introduces a lot of subjectivity. In most cases a low number of extracted factors gave the best results in terms of interpretability.

8. The selection of elements entered will govern the results of factor analysis. It may be worthwhile to first enter all elements, and afterwards test different combinations of elements. Just one element more or less can substantially alter the results of factor analysis. It is disturbing, that in geochemistry the selection of elements will very often arbitrarily depend on what has been analysed. This is often governed by price, methods available, and/or detection limits, and not by science.

9. For factor analysis that is carried out with geochemical data it is important to have a computer system that allows fast and easy graphical presentation of results, e.g. loading plots, scree plot and factor score maps. This allows the investigator to “play” with the data in a true exploratory data analysis approach.

Acknowledgements

The authors wish to thank the sampling and laboratory teams from Finland, Norway and Russia. Galina Kashulina (INEP, Apatity) and Viktor Melezhik (NGU, Trondheim) gave helpful comments on an earlier draft of this paper. Viktor Melezhik supplied us with the black and white version of the geological map used in this paper. Discussions on the results of factor analyses of the Kola C-horizon data with V. Chekushin, R. Dutter, F. Koller, H. Niskavaara and R. Salminen were appreciated. F. Koller provided support for C.R. in Vienna without this help over several months this work would not have been possible. The Norwegian Ministry of the Environment financed the Norwegian participation

in the Kola Project with special project funds from the Norwegian Ministry of Foreign Affairs. R. Howarth and an anonymous second reviewer provided us with very helpful comments.

References

- Afifi, A.A., Azen, S.P., 1979. *Statistical Analysis: A Computer Oriented Approach*. Academic Press, New York.
- Aitchison, J., 1986. *The Statistical Analysis of Compositional Data*. Wiley, New York.
- Åyräs, M., Reimann, C., 1995. Joint ecogeochemical mapping and monitoring in the scale of 1:1 mill. in the West Murmansk Region and contiguous areas of Finland and Norway — 1994–1996. *Field Manual. Nor. Geol. Unders. Rep.* 95.111.
- Bartlett, M.S., 1947. The use of transformations. *Biometrics* 3, 39–52.
- Bartlett, M.S., Kendall, D.G., 1946. The statistical analysis of variance — heterogeneity and the logarithmic transformation. *J. Roy. Stat. Soc.* 8/1 (Suppl.), 128–138.
- Basilevsky, A., 1994. *Statistical Factor Analysis and Related Methods. Theory and applications*. John Wiley & Sons, New York, USA.
- Böhm, P., Wolterbeek, H., Verburg, T., Musilek, L., 1998. The use of tree bark for environmental pollution monitoring in the Czech Republic. *Environ. Pollut.* 102, 243–250.
- Bølviken, B., Bergström, J., Björklund, A., Kontio, M., Lehmuspelto, P., Lindholm, T., Magnusson, J., Ottesen, R.T., Steenfelt, A., Volden, T., 1986. *Geochemical Atlas of Northern Fennoscandia. Scale 1:4,000,000. Geological Surveys of Finland, Norway and Sweden, Helsinki, Trondheim and Stockholm*.
- Bølviken, B., Kullerud, G., Loucks, R.R., 1990. Geochemical and metallogenic provinces: a discussion initiated by results from geochemical mapping across northern Fennoscandia. *J. Geochem. Explor.* 39, 49–90.
- Bølviken, B., Stokke, P.R., Feder, J., Jössang, T., 1992. The fractal nature of geochemical landscapes. *J. Geochem. Explor.* 43, 91–109.
- Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. *J. Roy. Stat. Soc. (B)* 26, 211–252.
- Butler, J.C., 1976. Principal components analysis using the hypothetical closed array. *Math. Geol.* 8, 25–36.
- Carroll, J.B., 1953. An analytic solution for approximating simple structure in factor analysis. *Psychometrika* 18, 23–38.
- Cattell, R.B., 1966. The scree test for the number of factors. *Mult. Behav. Res.* 1, 245–276.
- Chork, C.Y., 1990. Unmasking multivariate anomalous observations in exploration geochemical data from sheeted-vein tin mineralisation near Emmaville, N.W.S., Australia. *J. Geochem. Explor.* 37, 205–223.
- Chork, C.Y., Govett, G.J.S., 1985. Comparison of interpretations of geochemical soil data by some multivariate statistical methods, Key Anacon, N.B., Canada. *J. Geochem. Explor.* 23, 213–242.
- Chork, C.Y., Salminen, R., 1993. Interpreting geochemical data from Outokumpu, Finland: an MVE-robust factor analysis. *J. Geochem. Explor.* 48, 1–20.
- Dutter, R., Leitner, T., Reimann, C., Wurzer, F., 1992. *Grafische und geostatistische Analyse am PC. Beiträge zur Umweltstatistik. Schriftenreihe der Technischen Universität Wien, Bd. 29*, pp. 78–88.
- Eriksson, K., 1992. Glacigenic deposits. In: Govett, G.J.S. (Ed.), *Regolith Exploration Geochemistry in Arctic and Temperate Terraines. Handbook of Exploration Geochemistry, Vol. 5*. Elsevier, Amsterdam, pp. 13–40.
- Filzmoser, P., 1999. Robust principal component and factor analysis in the geostatistical treatment of environmental data. *Environmetrics* 10, 363–375.
- Garrett, R.G., 1989. The chi-square plot: a tool for multivariate outlier recognition. *J. Geochem. Explor.* 32, 319–341.
- Garrett, R.G., Nichol, I., 1969. Factor analysis as an aid in the interpretation of regional geochemical stream sediment data. In: Canney, F.C. (Ed.), *Internat. Geochem. Explor. Symp. Quart. Colorado School of Mines*; pp. 245–264.
- Garrett, R.G., Kane, V.E., Zeigler, R.K., 1980. The management, analysis and display of regional geochemical data. *J. Geochem. Explor.* 13, 115–152.
- Harman, H.H., 1976. *Modern Factor Analysis, 3rd Edition*. University of Chicago Press, Chicago.
- Hawkes, H.E., Webb, J.S., 1962. *Geochemistry in Mineral Exploration*. Harper & Row, New York.
- Hendrickson, A.E., White, P.O., 1964. PROMAX: a quick method for rotation to oblique simple structure. *Brit. J. Stat. Psychology* 17, 65–70.
- Hirvas, H., 1991. Pleistocene stratigraphy of Finnish Lapland. *Geol. Surv. Finland, Bull.* 354.
- Howarth, R.J., Earle, S.A.M., 1979. Application of a generalised power transformation to geochemical data. *J. Math. Geol.* 11, 45–62.
- Howarth, R.J., Sinding-Larsen, R., 1983. Multivariate analysis. Howarth, R.J. *Handbook of Geochemical Exploration, Vol. 2: Statistics and Data Analysis in Geochemical Prospecting*, Elsevier, Amsterdam, pp. 207–289.
- Johnson, R.A., Wichern, A.D., 1998. *Applied Multivariate Statistical Analysis*. Prentice Hall, Upper Saddle River, NJ, USA.
- Kaiser, H.F., 1958. The Varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23, 187–200.
- Khotinskiy, N.A., 1984. Holocene vegetation history. In: Velichko, A.A., Wright, H.E., Barnovsky, C.W. (Eds.), *Late Quaternary Environments of the Soviet Union*. University of Minnesota Press, Minneapolis, pp. 179–200.
- Koljonen, T. (Ed.), 1992. *Geochemical Atlas of Finland, Part 2: Till*. Geological Survey of Finland, Espoo, Finland.
- Kufs, C.T., 1979. Another view of the use of factor analysis in geology. *Math. Geol.* 11, 717–720.
- Kürzl, H., 1988. Exploratory data analysis: recent advances for the interpretation of geochemical data. *J. Geochem. Explor.* 30, 309–322.
- Le Maitre, R.W., 1982. *Numerical Petrology*. Elsevier, Amsterdam.
- Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. *Multivariate Analysis, 6th Print*. Academic Press, London, UK.
- Miesch, A.T., 1969. Critical review of some multivariate procedures in the analysis of geochemical data. *Math. Geol.* 1, 171–184.

- Niemelä, J., Ekman, I., Lukashov, A. (Eds.), 1993. Quaternary Deposits of Finland and Northwestern Part of Russian Federation and Their Resources 1:1,000,000. Geological Survey of Finland, Espoo, Finland.
- Niskavaara, H., 1995. A comprehensive scheme of analysis for soils, sediments, humus and plant samples using inductively coupled plasma atomic emission spectrometry (ICP-AES). In: Autio, S. (Ed.), Geological Survey of Finland, Current Research 1993–1994. Geol. Surv. Finland, Espoo, Spec. Pap. 20: pp. 167–175.
- Niskavaara, H., Kontas, E., 1990. Reductive coprecipitation as a separation method for the determination of gold, palladium, platinum, rhodium, silver, selenium and tellurium in geological samples by graphite furnace atomic adsorption spectrometry. *Anal. Chim. Acta* 231, 273–282.
- Pison, G., Rousseeuw, P.J., Filzmoser, P., Croux, C., 1999. Factor Analysis in a Robust Way. Univ. Antwerp, Belgium: [http://win-www.uia.ac.be/u/statis/\(preprint\)](http://win-www.uia.ac.be/u/statis/(preprint)).
- 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, 1001–1014.
- Reimann, C., Melezhik, V., 2001. Metallogenic provinces, geochemical provinces and regional geology — what causes large-scale patterns in low-density geochemical maps of the C-horizon of podzols in Arctic Europe? *Appl. Geochem.*
- Reimann, C., Wurzer, F., 1986. Monitoring accuracy and precision — improvements by introducing robust and resistant statistics. *Mikrochim. Acta* II, 31–42.
- Reimann, C., Äyräs, M., Chekushin, V.A., Bogatyrev, I., Boyd, R., Caritat, P. de, Dutter, R., Finne, T.E., Halleraker, J.H., Jæger, Ø., Kashulina, G., Niskavaara, H., Lehto, O., Pavlov, V., Räsänen, M.L., Strand, T., Volden, T., 1998. Environmental Geochemical Atlas of the Central Barents Region. NGU-GTK-CKE special publication. Geological Survey of Norway, Trondheim, Norway.
- Rock, N.M.S., 1988. Numerical Geology. Lecture Notes in Earth Sciences 18. Springer Verlag, New York/Berlin/Heidelberg.
- Rousseeuw, P.J., Van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. *J. Am. Stat. Assoc.* 85, 633–651.
- Seber, G.A.F., 1984. *Multivariate Observations*. Wiley, New York, USA.
- Temple, J.T., 1978. The use of factor analysis in geology. *Math. Geol.* 10, 379–387.
- Venables, W.N., Ripley, B.D., 1997. *Modern Applied Statistics with S-PLUS*, 2nd Edition. Springer, New York.
- Vistelius, A.B., 1960. The skew frequency distributions and the fundamental law of the geochemical processes. *J. Geol.* 68, 1–22.
- Woronow, A., Butler, J.C., 1985. Complete subcomposition independence testing of closed arrays. *Computers Geosci.* 12, 267–279.