Application of the Method of Statistical Comparison of XRD- and XRF-Data for Identification of the Most Representative Rock Samples: A Case Study of an Extensive Collection of Carbonatites and Aluminosilicate Rocks of the Kontozero Alkaline Complex (Kola Peninsula, NW Russia)

Ekaterina Fomina 1, Evgeniy Kozlov 1
1 Geological Institute of Kola Science Centre of Russian Academy of Sciences (GI KSC RAS), 14 Fersmana Street, Apatity, 184209 Russia
e.kozlov@ksc.ru

Abstract. We investigated carbonatites and aluminosilicate rocks from the Kontozero Devonian carbonatite paleovolcano complex (198 samples). Some specific features complicate the geological exploration of this object: a) the rocks of the Kontozero complex are predominantly volcanic and therefore exhibit small dimension of mineral grains and the diversity of their structural relationships; b) because breccias are common in many parts of the complex, the rocks are mostly inhomogeneous; c) Kontozero belongs to the alkaline-carbonatite formation, which is typically characterised by mineral diversity and the presence of rare minerals. The purpose of this study was to develop an algorithm for selecting from an extensive collection of rock material the most informative for mineralogical and geochemical studies samples. As a tool for this selection, we chose an original method of statistical comparison of XRD and XRF data using factor analysis (FA). This methodological approach enables mathematical identification of all major, minor, and several accessory minerals and a rough estimation of their contents (Fomina et al., 2019). We carried out the mineralogical interpretation of the factors according to the peak positions on the graphs of factor loadings and qualitative analysis of diffraction data of rock samples with maximum factor scores. For the studied rock collection, this approach allowed us to identify more than 20 rock-forming minerals based only on XRD data. Also, we found about ten mineral phases, the lines of which are low-intensity and/or overlap by more intense peaks of other minerals in the diffraction patterns. The mineralogical interpretation of the factors of these hidden minerals requires verification by an electron microscope investigation of the samples selected with FA. Based on the results of this study, we developed an algorithm that facilitates choosing rock samples that are most contrasted in mineral and chemical composition and contain the entire set of mineral phases characteristic of the rocks of the study object with minimal effort. In the case of our Kontozero rock collection, 30 representative samples were selected, which is about 15% of the initial sample set.

1. Introduction
When studying extensive collections of geological samples (e.g., core samples), geologists face the problem of selecting representative samples that can provide the most comprehensive information (mineralogical, geochemical, isotopic, etc. data) with minimal labour. The standard solution to this
problem is mostly intuitive. Earlier, we proposed the original technique of statistical comparison of XRD and XRF data using FA [1]. This methodological approach enables mathematical identification of all major (> 10 vol %), minor (1-10 vol %), and several accessory minerals and a rough estimation of their contents. As shown in [1], this technique also clarifies in which sample the concentration of the mineral phase of interest is highest, which significantly simplifies the mineralogical study. FA makes the study blind, which substantially reduces the influence of the researcher on the result. We present an example of an FA-based investigation of X-ray diffraction patterns and complementary XRF data on an extensive rock sample collection of the Kontozero carbonatite complex. Based on the results of this study, we developed an algorithm that remarkably facilitates choosing samples that are most contrasted in mineral and chemical composition and contain the entire set of mineral phases constituting the rocks of the study collection.

2. Materials and methods

2.1. Samples description

This study involved 198 core samples from the Kontozero volcano-plutonic alkaline-carbonatite complex. Kontozero belongs to the Kola Alkaline Province [2], [3], which intrusions formed in the Devonian in the interval of 360-380 Ma [4]. The predominance of volcanic rocks distinguishes the Kontozero complex among other complexes of the Kola Province. Still, it complicates petrographic studies due to the small dimension of minerals and the diversity of their structural relationships. Other features of the complex are the ubiquity of breccias and, like in any alkaline-carbonatite formation, mineral diversity and the presence of rare minerals. All these factors, along with the insufficient geological exploration ([5], [6]), much puzzles studying the rocks of Kontozero. Thus, by the beginning of this study, we had only minimum information about the mineral composition of the studied samples. Based on the chemical composition, we knew that the sample collection includes 1) sensu-stricto carbonatites (calcic, magnesian, and ferroan) containing < 20 wt% SiO2, 2) silicocarbonatites (essentially carbonate rocks of endogenous origin containing > 20 wt% SiO2), and 3) a variety of aluminosilicate rocks (from normal to alkaline; with both Na and K alkalinity type).

2.2. Analytical techniques

The primary source of information on the mineral composition was X-ray powder diffraction (XRD) from bulk rock samples. The chemical compositions of each sample were determined by X-ray fluorescence analysis (XRF). Both analytical methods are express and allow to obtain the results for extensive sample collections at the earliest stage of the research.

2.2.1. XRD. The X-ray powder diffraction data (XRD) of the bulk rock samples were collected at room temperature by the Shimadzu XRD-6000 diffractometer (Shimadzu Corp., Kyoto, Japan), using a Cu target X-ray generator with graphite monochromator. The scan range of the Bragg angle (2θ) was from 4.00° to 70.00° in the continuous regime with a scan speed of 2.00 °/min; sampling pitch was 0.02°. The work was performed on the analytical equipment of IM UB RAS, Miass, Russia (http://www.mineralogy.ru).

2.2.2. XRF. The X-ray fluorescence (XRF) data of the bulk rock samples were collected by the wavelength dispersive X-ray fluorescence spectrometer S4 Pioneer (Bruker AXS, Germany). Instrumental operation conditions for the main rock-forming (Na, Mg, Al, Si, P, K, Ca, Ti, Mn, Fe) and some minor elements (Ba, Sr, Zr) were the following: 30 kV at 80 mA for Na Kα, Mg Kα, Al Kα, Si Kα, P Kα, K Kα, and Ca Kα analytical lines and 50 kV at 40 mA for Ti Kα, Mn Kα, Fe Kα, Sr Kα, Zr Kα, and Rh Kα lines. For minimisation mineral and particle size effects, the samples were homogenised using the fusion sample preparation technique. Samples were preliminarily dried and calcined to determine loss of ignition values. Then 0.5 g of calcined sample was mixed with 7.5 g of flux (a mixture of lithium metaborate and lithium tetraborate) and fused in electric furnace TheOX (Claise, Canada) to obtain appropriate for the further analysis glasses. Certified reference materials
RM of igneous and sedimentary rocks, as well as apatite concentrates, were applied to build calibration curves. The lower detection limits were 0.05 wt % for all measured elements. Analytical procedure accounting of spectral overlaps and matrix effects correction using the fundamental parameters method, as well as the calculation of Sr and Zr contents utilising the intensity of the incoherent (Compton) anode emission (Rh) scatter peak. The research was performed using equipment of the Joint use Center «Isotope-geochemical Research» (IG SB RAS, Irkutsk, http://www.igc.irk.ru).

2.3. Data processing
Before data processing, we applied some spectral manipulations such as baseline correction and smoothing of the diffraction patterns. Smoothing was performed by using PeakFit v. 4.12 (Systat Software Inc., San Jose, USA) with Loess regression (level 0.5%). Baseline correction of diffractograms was accomplished in QualX v. 2.24 program (Institute of Crystallography (IC)-CNR, Bari, Italy) [7], using “Bezier Spline” (points selected by the program were interpolated via Bézier curve). The data processed in this way were then compiled into a single database in Microsoft Excel. Other preparatory manipulations detailed in [1] performed by using this program: (1) the removal of variables, the values of which after the baseline fitting dropped to zero in all diffractograms and (2) the addition of a small constant (for example, 0.001) to each intensity value. We also considered that after processing the data by using QualX, the maximum peak of the diffractogram automatically sets to 1000. The diffractograms were scaled by multiplying intensities at each 2θ by the coefficient $k = (I_{\text{max}} - I_{\text{min}})$, where $I_{\text{max}}$ and $I_{\text{min}}$ represent the maximum and minimum values in the corresponding “raw” diffractogram. Because most diffractograms showed subhorizontal baselines, this simplified approach to estimate the intensity of the principal peak satisfies the correctness. The set of diffractograms thus transformed was then supplemented with the contents of chemical elements in the corresponding samples. Factor analysis in the modification of the principal component method was performed using IBM SPSS Statistics v. 23 (IBM Corp., Armonk, NY, USA) [8]. VARIMAX rotation [9], which is the most commonly used in FA orthogonal rotation, was also applied. Factors were identified by using the online XRD AMCS database [10], the QualX v. 2.24 program with the indexed XRD database of open-access POW_COD [11], and the commercial PDF2 [12].

3. Results and discussions
FA aims at ascribing a large and unwieldy set of original (measured) variables to a smaller set of latent variables called factors. In this study, FA was applied to minimise the number of variables (factors) to the extent necessary to represent all non-random differences in the combined dataset, containing both diffractograms and chemical data, while retaining all relevant information. FA is performed in two steps, which are the determination of the factors and interpretation of the factor meanings. The processing of the Kontozero collection data yielded 107 factors describing a 100% variance of the raw data. Interpretation of the results involves the decryption of the information hidden in the parameters of scores and loadings. The score value describes the magnitude of a factor. The loading is the correlation coefficient between the factor and the original variable. Analysis of the graphs of factor loadings ($r_{x}$) showed that more than a third of the factors describe the noise component (figure 1). The graphs of noise factors do not exhibit distinct peaks but contain many “outliers” (single points with high $r_{x}$ values surrounded by low $r_{x}$ values). We excluded all these factors from consideration. About a third of $r_{x}$ graphs have either one intense peak or a series of low-intensity peaks. Factors with these $r_{x}$ graphs are interpretable only in exceptional cases (when analysing them, we used the techniques described below but in most cases to no avail).
Figure 1. The $r^F_A$ graphs of noise factors (42 out of 107 factors).

Only the remaining third of the factors, which showed many distinct peaks on $r^F_A$ graphs (figure 2), turned out to be informative. This group includes mainly the first 30 factors that together account for 90% of the total data variance.

Figure 2. Examples of $r^F_A$ graphs of the most informative factors (red line – factor No. 6 “analcime”; blue line – factor No. 16 “natrolite”; green line – factor No. 18 “quartz”).

The peak positions on the $r^F_A$ plots of the selected factors coincide with the corresponding peak positions on the diffractograms of certain minerals from the databases. For several factors, we discovered a functional relationship between the score values and the content of chemical components in the corresponding mineral phases. An indicative example is the apatite factor (from now on, the names of factors denote minerals to which the factors correspond). Its $r^F_A$ graph of factor loadings has intense peaks at the same $2\theta$ as the fluorapatite diffractogram from the RRUFF database [13] (figure 3A). Noteworthy, the score values of the apatite factor are linearly dependent on the $P_2O_5$ content (figure 3B). Hence, the score value of the factor allows us to make a rough estimate of the mineral content.

Figure 3. (A) Comparison of the fluorapatite diffractogram from the RRUFF database (ID R050122, blue line) and the $r^F_A$ graph of the fluorapatite factor (red line); (B) The functional relationship between the value of the fluorapatite factor and the content of $P_2O_5$. 
We compared the $r^{F4}$ graphs and the diffraction patterns of those samples in which the score values of the corresponding factors were the highest (i.e., the highest expected content of the mineral associated with the factor). Most of these diffraction patterns have intense peaks, the positions of which coincide with the positions of the peaks in the $r^{F4}$ graphs (figure 4a). For mineralogical identification of these factors (with intense peaks in the diffraction patterns of the samples), we used two techniques: 1) mineral search in the online XRD AMCSD database [10] using the “Diffraction Search” tool for the most intense peaks in the $r^{F4}$ with the “Tolerance” parameter equal to 0.1 (figure 4b); 2) qualitative identification of the phase of interest in the diffraction pattern with the maximum score values using QualX v. 2.24 program [7] (figure 4c). As shown in [1], in cases where the contents of some minerals in the rocks are close to proportional, it is possible to combine several minerals into one factor. During this study, we also observed peaks of several minerals (for example, magnetite and diopside) in some $r^{F4}$ graphs of factor loadings. Given the possibility of “mixing” several minerals into one factor, analysis of diffraction patterns with a maximum score value is preferable. However, it is technically more complex, and we achieved the best results by combining both interpretation techniques. The analysis of factor loadings on geochemical variables yielded additional evidence on the mineralogical nature of factors. There is a clear pattern in the attribution of the maximum factor loadings: Ca (0.83) and L.O.I. (0.75) for calcite factor; Na (0.56) and Al (0.53) for albite factor; Fe (0.41) and Mg (0.35) for the factor combining magnetite and diopside; P (0.95) for apatite factor; K (0.66) for orthoclase factor; S (0.67) for pyrite factor, etc. (the analysis of geochemical loadings is detailed in [1]).

**Figure 4.** (A) Comparison of the $r^{F4}$ graph of factor № 14 (“fluorite”, red line) with the diffractogram of a sample characterised by the maximum score value of this factor (blue line); (B) Comparison of $r^{F4}$ graph of fluorite factor with fluorite peaks from the AMCSD online XRD database; (C) QualX v. 2.24 Dialog box with a diffractogram of the sample with the maximum score value of fluorite factor (red peaks – PDF2 fluorite card)
The described methods identified about 20 factors associated with 21 minerals: Ca-Mg-Fe carbonates (calcite, ferruginous dolomite, and siderite), strontianite, feldspars (orthoclase and albite), garnet (andradite), monticellite, diopside, biotite, chlorite, serpentine, zeolites (natrolite and analcime), quartz, magnetite, ilmenite, fluorite, fluorapatite, pyrite and anatase. Subsequently, we found all these phases by electron microscope studies in thin sections of samples identified by the FA as a priority for their high score values (figure 5).

![Figure 5](image)

**Figure 5.** (A)–(F) The minerals of the studied rocks of the Kontozero complex: Ab – albite; Adr – andradite; Anc – Analcime; Ant – anatase; Ap – apatite; Brt – baryte; Bt – biotite; Cal – calcite; Chl – chlorite; Di – diopside; Dol – dolomite; Fl – fluorite; IIm – ilmenite; Mag – magnetite; Mtc – Monticellite; Ntr – natrolite; Or – orthoclase; Py – pyrite; Qz – quartz; Sd – siderite; Srp – serpentine; and Str – strontianite. All images are BSE photos.

We identified several factors that also have intense peaks in the $r^{FA}$ graphs. However, even in the diffraction patterns of samples with a maximum score value of these factors, only small lines present in the regions of $r^{FA}$ peaks often obscured by lines of other minerals (e.g., figure 6A). For the interpretation of such factors, the qualitative identification of the phase of interest by the diffractogram with the maximum score value is not practical. Thus, we diagnosed these factors by the position of the most intense peaks in the AMCSD database. The subsequent mineralological study of samples with maximum score values confirmed our assumptions. Ultimately, a comparison with the identified factors revealed several minerals in the studied rocks, which, although not abundant, are of significant petrological interest. Thus we found, for example, burbankite $(Na, Ca)_3(Sr, Ba, Ce)_3(CO_3)_{5}$ (figure 6B). Similarly, baryte, talc, the second generation of chlorite, and several varieties of biotite were found.
**Figure 6.** (A) Comparison of the $r^{FA}$ graph of factor No. 17 ("burbankite", the red line) with the diffraction pattern of the sample showing the maximum score value of this factor (blue line); (B) Burbankite (Bur) in the sample with the maximum score value of the burbankite factor

**4. Conclusions**

Summarizing the results of the study, we propose the following algorithm for the search for the most representative samples in an extensive collection of rocks of a geological object:

1) XRD and XRF analysis and primary data processing, including baseline corrections and removal of zero values. The output of this step is a database suitable for FA;

2) Factor analysis of the obtained results. The output is tables of $r^{FA}$ values and graphs, values of factor loadings on geochemical variables, and factor values for each sample;

3) Compilation and examination of all $r^{FA}$ graphs on a single chart. The output is the rejection of all non-interpreted noise factors;

4) Comparison of each factor graph with the diffraction pattern of the sample, which showed the maximum score value of this factor. The output is a division into easily and difficultly interpreted factors;

5) Interpretation of the easily interpretable factors by combining the two proposed techniques. The output is a highly confident mineralogical explanation of these factors;

6) Interpretation of difficultly interpreted factors by searching for the mineral(s) according to the position of the most intense peaks in the AMCS D database. The output is an assumption about the nature of these factors;

7) Mineralogical (optical investigation, SEM + EDX, Raman) study of the samples with the highest factor score values. The output is a verification of the FA results, a collection of the most representative samples, an idea of the mineral composition of the study object rocks at the level of main, minor, and some accessory minerals in a reasonably short time.

**Acknowledgments**

This research was funded by the Russian Science Foundation, grant number № 19-77-10039. Core sampling was carried out in TFGI NWFD (Apatility, Russia) under the GI KSC RAS research topic № 0226-2019-0053. The authors express their deepest gratitude to the TFGI NWFD employees for their assistance and Sidorov M.Yu (GI KSC RAS, Apatility, Russia) for his help with sample preparation. The authors thank Chuparina E.V. (IGC SB RAS, Irkutsk, Russia) for XRF analysis, Khvorov P.V. (IM UB RAS, Miass, Russia) for XRD analysis, and Bazai A.V. (GI KSC RAS, Apatility, Russia) for electron microprobe analysis and BSE images.

**References**

[1] E. Fomina, E. Kozlov, and S. Ivashevskaia, “Study of diffraction data sets using factor analysis: a new technique for comparing mineralogical and geochemical data and rapid diagnostics of the mineral composition of large collections of rock samples,” Powder Diffr., vol. 34, no. S1,
[2] H. Downes, E. Balaganskaya, A. Beard, R. Lifervich, and D. Demaiffè, “Petrogenetic processes in the ultramafic, alkaline and carbonatitic magmatism in the Kola Alkaline Province: A review,” Lithos, vol. 85, no. 1–4, pp. 48–75, 2005, doi: 10.1016/j.lithos.2005.03.020.

[3] A. G. Bulakh, V. V. Ivanikov, and M. P. Orlova, “Overview of carbonatite-phoscorite complexes of the Kola Alkaline Province in the context of a Scandinavian North Atlantic Alkaline Province,” in Phoscorites and carbonatites from mantle to mine, F. Wall and A. N. Zaitsev, Eds. London: Mineralogical Society of Great Britain and Ireland, 2004, pp. 1–43.

[4] U. Kramm, L. N. Kogarko, V. A. Kononova, and H. Vartiainen, “The Kola Alkaline Province of the CIS and Finland: Precise Rb-Sr ages define 380–360 Ma age range for all magmatism,” Lithos, vol. 30, no. 1, pp. 33–44, 1993, doi: 10.1016/0024-4937(93)90004-V.

[5] A. A. Arzamastsev and M. N. Petrovsky, “Alkaline volcanism in the Kola Peninsula, Russia: Paleozoic Khibiny, Lovozero and Kontozero calderas,” Proc. MSTU, vol. 15, no. 2, pp. 277–299, 2012.

[6] M. N. Petrovsky, E. A. Savchenko, and V. Y. Kalachev, “Formation of eudialyte-bearing phonolite from Kontozero carbonatite paleovolcano, Kola Peninsula,” Geol. Ore Depos., vol. 54, no. 7, pp. 540–556, 2012, doi: 10.1134/S1075701512070057.

[7] A. Altomare, N. Corriero, C. Cuocci, A. Falcicchio, A. Moliterni, and R. Rizzi, “QUALX2.0: a qualitative phase analysis software using the freely available database POW COD,” J. Appl. Crystallogr., vol. 48, no. 2, pp. 598–603, 2015, doi: 10.1107/S1600576715002319.

[8] D. George and P. Mallery, IBM SPSS Statistics 23 Step by Step. New York: Routledge, 2016.

[9] H. F. Kaiser, “The varimax criterion for analytic rotation in factor analysis,” Psychometrika, vol. 23, no. 3, pp. 187–200, 1958, doi: 10.1007/BF02289233.

[10] R. T. Downs and M. Hall-Wallace, “The American Mineralogist crystal structure database,” Am. Mineral., vol. 88, no. 1, pp. 247–250, 2003.

[11] S. Gražulis, A. Daškevič, A. Merkys, D. Chatiegnier, L. Lutterotti, M. Quiros, N. R. Serebryanaya, P. Moeck, R. T. Downs, and A. Le Bail, “Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration,” Nucleic Acids Res., vol. 40, no. D1, pp. D420–D427, 2012, doi: 10.1093/nar/gkr900.

[12] S. Gates-Rector and T. Blanton, “The Powder Diffraction File: a quality materials characterisation database,” Powder Diffr., vol. 34, no. 4, pp. 352–360, 2019, doi: 10.1017/S0885715619000812.

[13] B. Lafuente, R. T. Downs, H. Yang, and N. Stone, “The power of databases: The RRUFF project,” in Highlights in Mineralogical Crystallography, T. Armbruster and R. M. Danisi, Eds. Berlin: DE GRUYTER, pp. 1–30.