EXPLORE
NUMBER 176 SEPTEMBER 2017
Newsletter for the Association of Applied Geochemists
www.appliedgeochemists.org

Finally, a correlation coefficient that tells the geochemical truth

Robert G. Garrett[1], Clemens Reimann[2], Karel Hron[3], Petra Kynčlová[4] and Peter Filzmoser[4]; *[1]Emeritus Scientist, Geological Survey of Canada, Natural Resources Canada, 601 Booth St., Ottawa, Ontario, K1A 0E8, Canada; [2]P.O. Box 6315, Torgard, NO-7491, Trondheim, Norway; [3]Department of Mathematical Analysis and Applications of Mathematics, Palacký University, 17.listopadu 12, 77146 Olomouc, Czech Republic; [4]Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, WiednerHauptstr. 8-10, 1040 Vienna, Austria.*

https://doi.org/10.70499/YKCX6512

Introduction

Geochemists have long been aware of the problems surrounding estimating correlation coefficients for their analytical data sets. Very often they just don't make sense on the basis of the mineralogy of the sample material and our knowledge of mineral stoichiometry. The problem lies in the nature of geochemical analyses, they are relative measures reported in such units as weight %, parts per million (mg/kg), μg/L, etc., the sum of the parts, individual measures, add to a constant. Because of the relative units it does not matter whether all the parts have been determined in the analysis, the problem remains whatever the number of parts determined, even just two. The problems related to correlations were recognized by Pearson as long ago as 1897. The first geoscientist to study the problem systematically was Felix Chayes (1960) a research petrologist who worked for the Carnegie Institution's Geophysical Laboratory and for the Smithsonian Institution. The true information in a geochemical data set lies in the ratios between the parts, and Tom Pearce (1970) was the first geoscientist to promote the use of ratios in petrology, leading to a number of diagrams that are effective in classification and genetic studies. The mathematical groundwork for properly handling compositional data was laid out by John Aitchison (1984, 1986) with his exposition on the use of log-ratios. Since then numerous papers and books have been published on compositional data analysis, see for example Pawlowsky-Glahn *et al.* (2015) and the references in Reimann *et al.* (2017). Today a common approach in multivariate analysis, e.g., Principal Components or Factor Analysis, is to use a centred log-ratio (clr) of the data set prior to carrying out the analysis (e.g., Fig. 1). It might seem apparent then to also calculate the correlation coefficients on the clr-transformed data. However, this does not lead to consistent results, because clr variables are driven by their zero sum constraint. As a consequence, a negative bias occurs when correlation analysis in clr variables is performed. It is quite natural that different sub-compositions, i.e. subsets of the parts, for a data set do not yield the same correlation coefficients for the two parts of interest. The reason for this is the computation of the clr-transform involves dividing the value for each part (variable) by the geometric mean of all the parts in the subset for an individual sample; and different subsets for a sample will have different geometric means. One can also express each clr variable as a (scaled) sum of all pairwise log-ratios with the respective compositional part – a kind of intuitive result, when all information in compositional data is contained in log-ratios. A careful choice of parts, involved in the analysis, is thus always necessary.

A solution to the problem of negative bias of correlation analysis in clr variables has been proposed by Kynčlová *et al.* (2017) and involves the computation of symmetric coordinates, an extension of isometric log-ratios (Egozcue *et al.*, 2003). The symmetric coordinates are computed as weighted log-ratios that take the total number of parts into consideration. This procedure has been demonstrated with two large sets of geochemical (environmental) soil data by Reimann *et al.* (2017). The purpose of this article is to demonstrate the procedure and discuss the results for a small set of petrochemical data whose mineralogy will be familiar to readers. As such, this article is a tutorial rather than a contribution of original science. The data set of 16 'averages' for common plutonic rocks was published by Nockolds

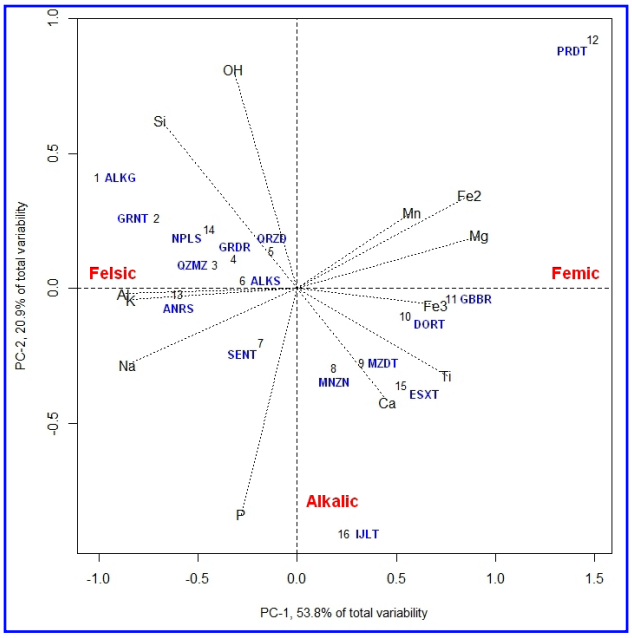

Figure 1. Principal Components Analysis for the clr-transformed Nockolds data set. Lithologies: 1- Alkali Granite; 2 - Granite; 3 - Quartz Monzonite; 4 – Granodiorite; 5 – Quartz Diorite; 6 – Alkali Syenite; 7 – Syenite; 8 – Monzonite; 9 – Monzodiorite; 10 - Diorite; 11 - Gabbro; 12 - Peridotite; 13 - Anorthosite; 14 - Nepheline Syenite; 15 - Essexite; 16 - Ijolite

continued on page 5

Finally, a correlation coefficient that tells the geochemical truth...

continued from page 1

(1954) as oxide percentages. More recent compilations have been made, but the Nockolds data suffices for the demonstration. The original oxides have been converted to cation percentages and H_2O+ to OH^-, see Appendix 1 (see digital version of Appendix 1 on the AAG website).

Data Analysis

For all the following computations and graphical presentations version 1.1.14 of the R (2017) package 'rgr' (Garrett, 2017) was employed. To graphically illustrate the interrelations between the geochemical data and the lithology a Principal Components Analysis (PCA) was undertaken following a centred log-ratio transformation (function 'gx.mva.closed'), see Figure 1 (function 'gx.rqpca.plot'), which was annotated (coloured text) with the lithological abbreviations outside 'rgr'. The end members and outliers in Figures 2 to 4 were similarly annotated. Functions in 'rgr_1.1.14', 'xyplot.tags' in conjunction with function 'gx.symm.coords.mat', can directly display plots tagged by text, such as lithological names.

The first principal component, PC-1, explains 74.7% of the total variability in the data set. High Si, Al and alkali metal felsic, quartzo-feldspathic, rocks are characterized by negative PC-1 scores, while femic, ferromagnesian mineral-rich, rocks high in Mg, Fe^3, Fe^2, Mn and Ti are characterized by positive PC-1 scores. In contrast, alkalic rocks with higher Ca, Na and P contents are characterized by negative PC-2 scores. The path from felsic intrusives, e.g., generally granitic, to femic rocks (gabbros and diorites) follows a 'NW' to 'SE' trend. Two Si deficient rocks, olivine- and pyroxene-rich peridotite, and nepheline- and alkali pyroxene-rich ijolite both plot as 'outliers' off-trend. The essentially mono-mineralic rock anorthosite, with dominant plagioclase feldspar, plots proximal to Al, K and Na close to the main trend in the data.

The default procedure in function 'gx.symm.coords.r' calculates Spearman correlation coefficients for the symmetric coordinates derived from the input data. Spearman ranked coefficients are preferred over Pearson product moment coefficients as they provide better estimates of correlation for data pairs that vary monotonically, i.e. the data points vary sympathetically or antipathetically, but not necessarily linearly. Furthermore, any monotonic transformation, e.g., logarithmic, has no impact on the Spearman coefficient as the ranks remain the same. For Exploratory Data Analysis (EDA) any systematic data relationship is of interest, even if it is curvilinear; should modelling be required linearizing transformations can be sought.

The correlation matrix (Table 1) contains two sets of Spearman coefficients, the upper triangle contains those based on the symmetric coordinates computed after Kynčlová *et al*. (2017), and the lower contains those based on the input data. Alternately, Pearson coefficients may be selected, and the further option exists to apply a logarithmic transformation to the input data, which has been common practice amongst applied geochemists.

Table 1. Spearman correlation coefficients for the Nockolds data set. Upper triangle based on symmetric coordinates, lower triangle based on raw data

	Si	Al	Fe^3	Fe^2	Mg	Ca	Na	K	Ti	Mn	P	OH
Si		0.87	-0.58	-0.38	-0.66	-0.55	0.56	0.74	-0.79	-0.40	-0.16	0.71
Al	-0.44		-0.54	-0.67	-0.82	-0.44	0.77	0.75	-0.80	-0.42	-0.27	0.60
Fe^3	-0.79	0.36		0.45	0.54	0.22	-0.22	-0.13	0.46	0.84	0.08	-0.19
Fe	-0.74	-0.04	0.80		0.93	0.51	-0.88	-0.59	0.67	0.41	-0.06	-0.23
Mg	-0.76	0.00	0.76	0.98		0.71	-0.86	-0.77	0.92	0.44	0.11	-0.30
Ca	-0.78	0.54	0.63	0.57	0.64		-0.32	-0.76	0.82	0.10	0.09	-0.30
Na	-0.13	0.77	0.20	-0.28	-0.30	0.21		0.64	-0.68	-0.25	0.01	0.47
K	0.66	0.06	-0.30	-0.56	-0.63	-0.63	0.29		-0.74	-0.30	0.47	0.41
Ti	-0.79	0.37	0.94	0.84	0.83	0.71	0.13	-0.36		0.41	0.15	-0.35
Mn	-0.77	0.19	0.84	0.76	0.70	0.37	0.06	-0.37	0.78		-0.21	-0.01
P	-0.40	0.40	0.75	0.49	0.50	0.62	0.30	0.00	0.79	0.37		-0.62
OH	-0.51	0.23	0.41	0.50	0.52	0.34	0.10	-0.48	0.54	0.53	0.17	

Discussion

Silicon (Si) is the dominant part in the data set with cation percentages varying from some 20%, ijolite and peridotite, to 34.5%, alkali granite (see Appendix 1). Reading down the first column of Table 1, the Spearman coefficients are all negative, but for K. As the dominant part (Si) increases most of the remaining parts have to decrease to maintain constant sum. Yet from the mineralogy of these rocks we know that Si, Al, Na and K increase together in felsic rocks as the amounts of quartz, and alkali feldspar increase, together with white micas (OH), at the expense of less Si-rich ferromagnesian minerals rich in Fe, Mg, Ti and Mn, such as dark micas and amphiboles that are more abundant in femic rocks.

This mineralogical reality is reflected in the Spearman coefficients based on the symmetric coordinates displayed across the first row of Table 1. The negative symmetric coordinate correlations for Mg, Fe^2, Fe^3, Ti, Mn and Ca reflect the sympathetic relationship between these elements in ferromagnesian minerals from amphiboles, through pyroxenes to olivines, as they increase in abundance in femic rocks. This increase is at the expense of quartz (Si), albitic (Na) and orthoclase (K) aluminosilicate feldspars, and is reflected in positive symmetric coordinate correlations between Si, Al, Na, K and OH, and, as a group, their negative correlations with Mg, Fe^2, Fe^3, Ti, Mn and Ca.

continued on page 6

Finally, a correlation coefficient that tells the geochemical truth... continued from page 5

Data inspection and interpretation is often facilitated and improved by graphical presentations, function 'gx.symm.coords. plot' undertakes that task. The classic example of problems with compositional data is the Harker diagram, which dates back to 1909, for plotting various oxides against silica. Silicon (Si) and Al are the dominant cation pairs for each of the lithologies in the Nockolds data except peridotite (Mg & Fe replace Al), and Al-rich ijolite and nepheline syenite. The plot for Si and Al is presented in Figure 2.

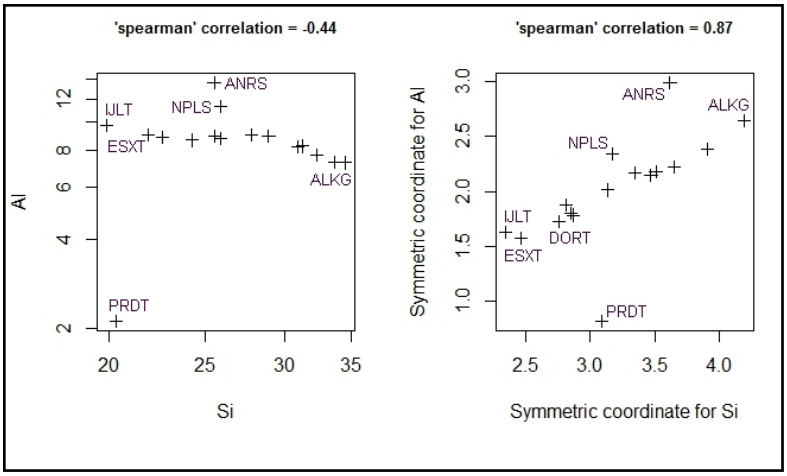

Figure 2. Plots of the Nockolds data for Si and Al as a pseudo Harker diagram (left) and as symmetric coordinates (right)

The Harker plot on the left shows the familiar negative relationship imposed by the compositional form of the data, with the mineralogical and geochemical outliers, peridotite in the lower left, and nepheline syenite and anorthosite at the top with highest Al. In contrast, the plot based on symmetric coordinates (Fig. 2, right) demonstrates sympathetically increasing Si and Al, with the ultramafic peridotite remaining an outlier at the bottom of the plot. The other two upper outliers are of interest, the most extreme is Al-rich anorthosite, and the less is nepheline syenite, which lies in the felsic to femic trend observed in the PCA (#14 in Fig. 1). The difference between the two plots is summarized in the differences between their Spearman correlations, -0.44 for the Harker plot and 0.87 for the symmetric coordinate plot, a convincing reversal. In this case the Pearson correlation is of interest. It is surprisingly positive 0.18 (with a logarithmic transformation) for the Harker plot, however, this is due to the influence of the high leverage outlier peridotite, and in view of the graphic (Fig. 2, left) totally misleading. The Pearson correlation for the symmetric coordinates is 0.69, essentially unchanged.

A similar reversal can be demonstrated with Ca and Na, the two cations in the anorthite-albite plagioclase solid solution series (Fig. 3).

continued on page 8

Note: *This EXPLORE article has been extracted from the original EXPLORE Newsletter. Therefore, page numbers may not be continuous and any advertisement has been masked.*

Finally, a correlation coefficient that tells the geochemical truth...

continued from page 6

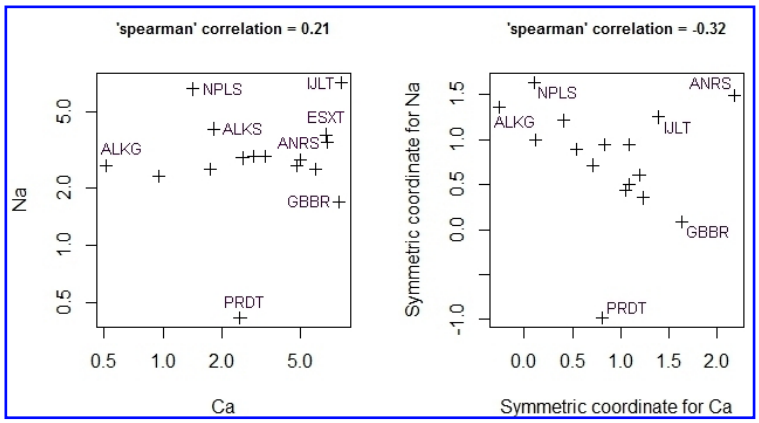

Figure 3. Plots of the Nockolds data for Ca and Na (left) and as symmetric coordinates (right).

The standard plot on the left demonstrates little variation in Na, exaggerated graphically by the presence of the low Na peridotite at the bottom. The two highest Na lithologies are nepheline syenite and ijolite, alkalic rocks. The expected antipathetic relationship between the two plagioclase end members, and the increase of anorthitic members in femic rocks versus the increase in albitic members in felsic rocks is not apparent due to the role of Si and Al as dominant parts in the rock compositions. The plot of the symmetric coordinates (Fig. 3, right) illustrates what we know as 'true' on the basis of mineral stoichiometry and petrology, a strong antipathetic Ca-Na relationship due to the plagioclase solid solution series and the observed mineralogical variations between felsic and femic rocks. Peridotite remains an outlier at the bottom of the plot, and the upper right-most outlier is anorthosite, and the less extreme is ijolite. Summarized numerically by the Spearman coefficients, the untransformed data are positively correlated, 0.21, and the symmetrically transformed data are negatively correlated, -0.32, as should be expected on geochemical grounds.

A final example is one involving K and Ti, a minor element, i.e. between 1 and 0.1% in the composition, which clarifies their relationship, Figure 4.

The standard plot on the left shows a generally antipathetic relationship between K and Ti. As to be expected as K-rich felsic rocks are poor in Ti bearing minerals such as biotite, ilmenite and rutile and femic rocks are rich in Ti-bearing biotites, amphiboles, and other ferromagnesian minerals, but poor in K-rich minerals. There are two outliers, low Ti alkali granite and high Ti essexite, a Si under-

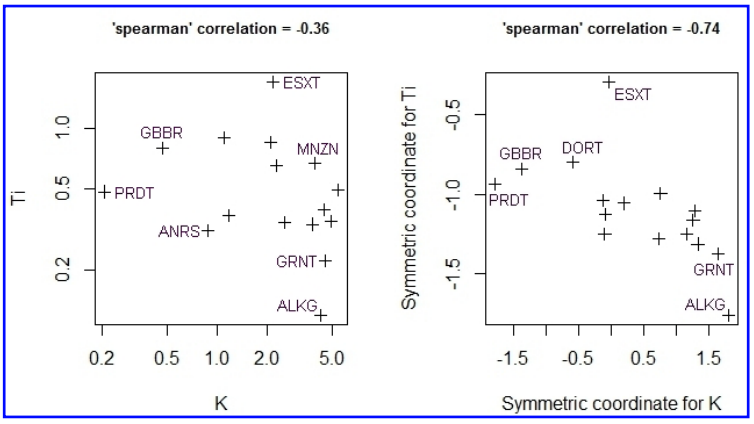

Figure 4. Plots of the Nockolds data for K and Ti (left) and as symmetric coordinates (right).

saturated rock dominated by plagioclase feldspar and pyroxene. The plot based on symmetric coordinates (Fig. 4, right) is much tidier, the main mass of the data plots within a more confined band due to the reduced influence of all the remaining parts in the total composition. The high Ti symmetric coordinate outlier, -0.30, is essexite which has the highest Ti cation percentage; the lithology in the lower right corner is alkali granite, which from its mineralogy of abundant orthoclase (K) and minimal biotite (Ti) plots as expected. Summarized numerically, the raw data Spearman coefficient of -0.36 has been improved to -0.74 through the symmetric coordinates removal of the effect of the competing parts in the composition on a part that is a minor/trace contributor to the composition.

The Nockolds data set contains only major (Si, Al, Fe, Mg, Ca, Na & K) and minor (Ti, Mn & P) elements. It is used here as an example because of the ease of its interpretation. Many researchers are under the wrong impression that compositional data effects only exist when working with major elements. It has often been assumed (including by the senior author in the past)

continued on page 9

Finally, a correlation coefficient that tells the geochemical truth... continued from page 8

that a simple logarithmic transformation of minor and trace element data is sufficient. The example of Ti above demonstrates that the effect is not restricted to major element concentrations. The dominantly trace element study of Norwegian soils by Reimann *et al.* (2017) demonstrates that equally strong effects are exhibited for trace elements. Compositional effects are present in water analyses where the concentrations are usually reported in μg/L, three orders of magnitude lower than ppm (μg/g), and they must be treated appropriately in order to obtain a correct representation of the interrelationships between the parts (Flem *et al.*, submitted). It is of no importance whether or not major elements are determined, the effect is inherent in the data – in their relative units.

Conclusions

It has been demonstrated how the use of symmetric coordinates leads to correlation coefficients that 'tell the truth' and provide numerical expression to our observations of the mineralogy of the igneous rock and the stoichiometry of their minerals. Furthermore, the graphical display of the symmetric coordinates greatly improves the ability to interpret the results in a geoscientific context. The example of Si and Al clearly demonstrates the advantage of Spearman correlations over Pearson correlations in this kind of exploratory (EDA) investigation by the reduction of the influence of high leverage outliers. Importantly, the results presented go beyond correlation analysis. They demonstrate that simple bivariate scatterplots are not 'simple' at all when working with compositional data. The true relations between two parts only becomes clear when their symmetric coordinates are studied.

The Nockolds data are simple in structure and the underlying petrology and mineralogy are well understood and this is the reason they are used here. Interpretation of the Reimann *et al.* (2017) exposition for C- and O-horizon soils from a Norwegian survey is far more complex, and compounded by major variability introduced by varying ratios of minerogenic and organic fractions within the individual soil samples.

Correlation coefficients are sometimes inferred to imply causal relationships between the variables, or parts for compositional data. This can be dangerous as both measures may be unrelated directly, but through a third measure, 'a lurking variable', that may, or may not, have been measured. The result of this is that the inferred causation can be false and conclusions drawn erroneous. Given this, it is even more important for scientists working with compositional data to numerically estimate and display bivariate relations without the influence of the compositional nature of their data.

It is to be hoped that this procedure of working with symmetric coordinates will be incorporated into the common software packages used by geochemists and other users of compositional data. To facilitate their use the R scripts for the three symmetric coordinate functions are included in digital Appendix data files 3, 4, and 5 on the AAG website and an example of their use in Appendix 2; and if R is unavailable or inappropriate the processing flow and logic can be translated into a more convenient language for the user.

Acknowledgements

The authors gratefully acknowledge the constructive suggestions of Chris Lawley and Pim van Geffen for improvements to the article.

References

AITCHISON, J. 1984. The statistical analysis of geochemical compositions. *Mathematical Geology*, **16**, 531-564.

AITCHISON, J. 1986. The statistical analysis of compositional data. Chapman and Hall, London, U.K., 416 pp.

CHAYES, F. 1960. On correlation between variables of constant sum. *Journal of Geophysical Research*, **65**, 4185-4193.

EGOZCUE, J.J., PAWLOWSKY-GLAHN, V., MATEU-FIGUERAS, G. & BARCELÓ-VIDAL, C. 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35**, 279-300.

FLEM, B., REIMANN, C., BIRKE, M., FILZMOSER, P. & BANKS, D. Submitted. Graphical statistics to explore the natural and anthropogenic processes influencing the inorganic quality of drinking water, ground water and surface water. *Applied Geochemistry*.

GARRETT, R.G. 2017. 'rgr': Applied Geochemistry EDA. https://cran.r-project.org/package=rgr .

KYNČLOVÁ, P., HRON, K. & FILZMOSER, P. 2017. Correlation between compositional parts based on symmetric balances. *Mathematical Geosciences*, **49**, 777-796.

NOCKOLDS, S.R. 1954. Average chemical compositions of some igneous rocks. Geological Society of America, Bulletin 65, 1007-1032.

PAWLOWSKY-GLAHN, V., EGOZCUE, J.J. & TOLOSANA-DELGADO, R. 2015. Modeling and analysis of compositional data. Wiley, Chichester, U.K., 272 pp .

PEARCE, T.H. 1970. A contribution to the theory of variation diagrams. *Contributions to Mineralogy and Petrology*, **19**, 142-157.

PEARSON, K. 1897. Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London, LX, 489-502.

R 2017. R Project for Statistical Computing.https://www.r-project.org/ (last accessed May 21, 2017)

continued on page 10

Finally, a correlation coefficient that tells the geochemical truth... continued from page 9

REIMANN, C., FILZMOSER, P., HRON, K., KYNČLOVÁ, P. & GARRETT, R.G. 2017. A new method for correlation analysis of compositional (environmental) data - a worked example. *Science of the Total Environment*, **607-608**, 995-971.

Appendix 1

The Nockolds igneous plutonic data set as used in the report

Lithology	Si	Al	Fe^3	Fe^2	Mg	Ca	Na	K	Ti	Mn	P	OH
ALKG	34.53	7.28	0.55	0.88	0.16	0.51	2.60	4.26	0.120	0.039	0.061	0.444
GRNT	33.70	7.33	0.60	1.30	0.31	0.95	2.29	4.53	0.222	0.046	0.079	0.500
QZMZ	32.33	7.74	0.85	1.76	0.60	1.75	2.49	3.80	0.336	0.046	0.087	0.510
GRDR	31.27	8.29	0.93	2.01	0.95	2.54	2.85	2.55	0.342	0.054	0.092	0.614
QRZD	30.93	8.23	0.95	2.66	1.17	3.32	2.89	1.18	0.372	0.062	0.092	0.651
ALKS	28.92	8.95	1.62	2.04	0.58	1.82	4.05	4.91	0.348	0.085	0.083	0.500
SENT	27.77	9.06	1.53	2.20	1.22	2.90	2.91	5.42	0.497	0.062	0.166	0.595
MNZN	25.88	8.77	1.80	3.56	2.21	4.83	2.60	3.88	0.671	0.101	0.192	0.566
MZDT	25.55	8.99	2.28	4.18	2.38	5.00	2.79	2.29	0.653	0.108	0.188	0.566
DORT	24.24	8.68	1.91	5.42	3.69	6.00	2.49	1.10	0.899	0.139	0.153	0.755
GBBR	22.61	8.91	1.78	6.16	4.86	7.91	1.68	0.46	0.791	0.139	0.105	0.604
PRDT	20.35	2.11	1.76	7.65	20.52	2.47	0.42	0.21	0.485	0.163	0.022	0.717
ANRS	25.50	13.61	0.58	1.13	0.50	6.88	3.46	0.88	0.312	0.015	0.048	0.595
NPLS	25.89	11.27	1.69	1.55	0.34	1.42	6.56	4.43	0.396	0.147	0.083	0.906
ESXT	21.92	9.03	2.53	4.62	2.93	6.78	3.78	2.19	1.684	0.124	0.209	0.916
IJLT	19.91	9.77	2.80	3.26	1.94	8.13	7.09	2.12	0.845	0.155	0.663	0.528

ALKG - Alkali Granite; GRNT - Granite; QZMZ - Quartz Monzonite; GRDR - Granodiorite;
QRZD - Quartz Diorite; ALKS - Alkali Syenite; SENT - Syenite; MNZN - Monzonite;
MZDT - Monzodiorite; DORT - Diorite; GBBR - Gabbro; PRDT - Peridotite;
ANRS - Anorthosite; NPLS - Nepheline Syenite; ESXT - Essexite; IJLT – Ijolite

Appendix 2

Example scripts for use with symmetric coordinates functions

It is taken that the data table in Appendix 1 (see digital version of Appendix 1 on AAG website) has been converted to a .csv file and imported into R as a data frame. Note that there can be no missing entries in the data table, if a value is missing the column must be deleted, or a suitable value imputed:

```
> nockolds <- read.csv("D:\\my data\\nockolds.csv")
```

To generate the correlation matrix with Spearman coefficients, upper triangle based on symmetric coordinates, lower triangle based on untransformed data, Table 1, the default:

```
> gx.symm.coords.r(nockolds)
```

To generate the correlation matrix with Pearson coefficients, upper triangle based on symmetric coordinates, lower triangle based on log transformed data:

```
> gx.symm.coords.r(nockolds, method = "pearson", log = TRUE)
```

To generate the Si-Al plots in Figure 2, note that Si is in the second column of the data frame and Al in the third:

```
> gx.symm.coords.plot(nockolds, 2, 3)
```

Similarly, for the Ca-Na plots in Figure 3, with Ca in the seventh column and Ca in the eighth:

```
> gx.symm.coords.plot(nockolds, 7, 8)
```