The 'rgr' package for the R Open Source statistical computing and graphics environment - a tool to support geochemical data interpretation

Robert G. Garrett
Geological Survey of Canada, Ottawa, Ontario, K1A 0E7, Canada
**Corresponding author (e-mail: garrett@NRCan.gc.ca)*

ABSTRACT: The development of interactive computer graphics to support applied geochemistry over the last 40 years at the Geological Survey of Canada (GSC) is briefly discussed. The loss of an interactive computing environment, IDEAS, in 1995 based on a DEC VAX computer largely negated nine years of work, though the experience gained was invaluable. The availability of the commercial S-PLUS package in a Windows PC environment led to the redevelopment of most of the functionality of IDEAS in the S language for statistical analysis and graphics. In 2006 a request from a sister federal government department for the S-PLUS software led to the decision to translate the S functions into R, an Open-Source implementation of the S language, and therefore free to the user. Since that time all development has been in R, resulting in the 2007 release to the public of a package of tools, 'rgr', to assist applied geochemists in interpreting their data. Subsequently, 'rgr' has been updated and extended. The move to Open Source R and the release of the 'rgr' package on the Comprehensive R Archival Network (CRAN) has made these tools, and their documentation, available for Windows, Unix and Mac computing environments. The paper outlines the features of 'rgr' and illustrates key graphic and tabular displays. Its functionality is reviewed in the context of earlier GSC interactive graphics packages.

SUPPLEMENTARY MATERIAL: this is available at http://www.geolsoc.org.uk/SUP18713

KEYWORDS: *applied geochemistry, mineral exploration, exploratory data analysis, EDA, QA/QC, multivariate analysis, software tools, R*

To be able to use the power of computers to undertake the analysis of geochemical survey data, and display the results in graphic form as diagrams and maps, was a dream of applied geochemists when computers first became available in the 1960s. Early attempts were constrained by the availability of output devices, i.e. line printers and static plotters, and batch processing. With the advent of interactive computing and personal computers in the 1970s and 1980s the dream was achievable.

At the Geological Survey of Canada (GSC) the first attempts were made in 1972 (Crain 1974; Garrett 1974), but the project failed for the lack of appropriate institutional computing power. A review of the Applied Geochemistry Subdivision's computer requirements was undertaken in 1979, during which the concept of an interactive graphics package to meet geochemists' needs was revived. Between 1980 and 1981 a functional specification of a system to meet the defined needs was prepared; the platform would have been a mini-computer. But, again, due to a lack of funds the project was set aside. In 1984 the GSC, together with the Department of Energy Mines and Resources' Earth Physics Branch and Computer Science Centre, acquired a DEC VAX 11/780 running the VMS operating system to meet the needs, amongst others, for interactive graphics. This enabled the development of IDEAS, the Interactive Data Exploration and Analysis System, with the assistance of summer and co-op programme students from the University of Waterloo, and Ottawa and Carleton Universities (Garrett 1988). Notably, IDEAS included many of the graphics displays that were being promoted for use in Exploratory Data Analysis (EDA) (Tukey 1977; Velleman & Hoaglin 1981).

IDEAS met many production needs, while SPSS and locally written Fortran programs met others, and served as a development platform for work on multivariate probability plots for outlier recognition and for allocation/classification procedures (Garrett 1989, 1990). However, in 1995, as part of a federal government programme review, the lease on the Department's two VAX 8700s was terminated. By that time some of the original Tektronix graphics terminals used with IDEAS had been replaced by PCs running emulation software, as well as Access, dBASE III and Rbase System V, all of which were being widely used on PCs to manage and compile data prior to processing them in IDEAS.

Thus in 1995 the interactive graphics facility was lost. Work at Bell Labs, Murray Hill, New Jersey, which commenced in 1976, had led to the development of the S language (Chambers 1977; Becker & Chambers 1984). S ran under Unix and was an

Geochemistry: Exploration, Environment, Analysis, Vol. **13**, 2013, pp. 355–378
http://dx.doi.org/10.1144/geochem2011-106
Published *Online First* on November, 15, 2013

internal tool for statistical analysis and data display used by Bell Labs staff. In 1981 AT&T, Bell Labs' parent company, introduced a generous licensing policy for the university community, and as a result S was rapidly adopted in many statistics departments as a working and development tool. A commercial version of S, S-PLUS, was developed by Doug (R.M.) Martin at the University of Washington, Seattle, and marketed by his wholly owned company Statistical Sciences, Inc. In 1993, Statistical Sciences became AT&T's exclusive licence holder and was then bought out by MathSoft and became its Data Analysis Products Division. This was followed by an aggressive marketing and cost-attractive campaign targeting government agencies and universities. Thus in 1994, knowing of the demise of the VAX computers at financial year's end, March 1995, a DOS version of S-PLUS was acquired. Then, in 1995, when S-PLUS 3.3 for Windows 95/NT was released, the annual upgrade and maintenance agreement provided a tool that could be used as a foundation for redeveloping IDEAS.

Between 1995 and 2007 some 100 'rg' functions were written at the GSC for the S-PLUS proprietary statistical software to support exploration and applied geochemical survey and research activities. The functions supported tasks, such as data QA/QC (Garrett & Grunsky 2003), data sub-setting, provision of EDA tools and graphical displays, graphics for geochemical map preparation, threshold selection, weighted sums (Garrett *et al.* 1980; Garrett & Grunsky 2001), and a variety of multivariate statistical analysis procedures. Most of these function scripts were written from 'scratch'. However, others were based on scripts shared within the S-user community on S-News (http://www.biostat.wustl.edu/s-news/). One of the uses of S-PLUS scripts was to prepare GSC Open File 5084 (Rencz *et al.* 2006) as part of a contract with Health Canada in support of risk assessments for their Federal Contaminated Sites Program. Following delivery of the Open File, Health Canada requested copies of the software that had been used to generate the report figures and tables. It was decided that it would be preferable to convert the S-PLUS functions into an R package in order to minimize computer software costs to Health Canada and their contractors, and to make the scripts more generally available.

THE R PROJECT

Development of R commenced in 1995 as an alternate implementation of the S language (Ihaka & Gentleman 1996). The initiative has been led by a five-person Board with international representation, supported by a 19-person international core development team (R-Project 2013). The R project is an official part of the Free Software Foundation's GNU project, and thus all products and packages held in the Comprehensive R Archival Network (CRAN) are Open Source and in the public domain. As a result, R has been widely adopted in the university community and has been, and continues to be, one of the statistics and graphics systems of choice for research and method development among statistical researchers. As new tools become available they can be prepared as packages (libraries) for distribution through CRAN (http://cran-r.project.org) and be shared among the science and engineering communities. Currently (27/08/2013) 4744 packages are available on CRAN to run on Windows, Mac and Unix platforms, and it is estimated that there are over 100 000 R users worldwide.

R is an object-oriented language. An object may be a script (processing instructions in a function, cf. a program), data, or output from a function (known as a 'saved object') that in turn may be a list containing numeric and other information generated by the function. The availability of 'saved objects' permits results to be passed from function to function for further processing or display.

For geochemical data the most useful construct is the data frame, essentially a table with row and column information. A data frame may contain numeric and text string data. In the latter case, the data are stored as 'factors' by which the data may be subdivided. Data, including 'factors', may also be stored as vectors and matrices, the latter like a data frame but lacking the row information. The advantage of a data frame is that its contents can be made easily accessible in an R session and the variables in the data frame accessed through their names in the column headings. R functions are available to move from one data construct to another.

Package 'rgr'

For the most part the conversion of S scripts to R scripts is straightforward. However, some 'rg' functions were encountered where, due to a few fundamental differences in the operations of S and R, some re-scripting was necessary. A total of 40 'rg' functions, along with six test data sets to support the examples embedded in the package, were made available on CRAN as 'rgr' version 1.0.1 in March 2007 (Garrett & Chen 2007). Over the next four years additional functions were moved from S-PLUS, or developed in R, for inclusion in 'rgr'. Additionally, a number of changes were made to the R scripts to keep up with evolving language and help-file requirements. Wherever possible existing functions in R libraries and packages were employed by writing 'wrappers' consistent with the 'rgr' style around them. The 'wrappers' executed these functions which, with appropriate defaults for applied geochemical data, were easy to use by geochemists. In October 2012 version 1.1.8 was placed on CRAN and included a range of multivariate tools that had not been in previous 1.0 versions. In March 2013 version 1.1.9 (Garrett 2013) contains 103 functions, selectively described in greater detail below, 15 test data sets, 118 hypertext help files accessible during 'rgr' use, and 3 other documents providing additional information for R and 'rgr' users and those wishing to work directly with the 'rgr' functions. Within the 'rgr' package there is also a printed manual that, together with Reimann *et al.* (2008) and the statistical functions in 'rgr' and R itself, provide a wide range of statistical and graphical tools to support applied geochemical data interpretation. The R scripts that generated the examples in Reimann *et al.* (2008) are available at http://www.statistik.tuwein.ac.at/StatDA/R-scripts. Furthermore, the 'StatDA' package that developed from those scripts is available on CRAN (Filzmoser & Steiger 2012).

In many 'rgr' functions the use of robust procedures is encouraged, for instance by the display or use of the median and Median Absolute Deviation (MAD) rather then the mean and standard deviation (SD). For multivariate data analysis two functions are present specifically to undertake robust estimation, and a third is present to facilitate the Graphical Adaptive Interactive Trimming (GAIT) of data sets (Garrett 1989). The reason for employing robust procedures is that most applied geochemical survey data are characterized by the presence of outliers; in fact, in mineral exploration the search is for the outliers. Robust estimators lead to improved estimations of the statistical descriptors for the core background data, i.e. mean and SD (variance) or means and correlations (covariances), and by viewing all the data in the context of these background estimators, the outliers become even more pronounced and easier to recognize.

The compositional nature of applied geochemical data

The analytical data that are the subject of mapping and interpretation in applied geochemistry are in the form of compositions (measurements that sum to a constant), whether 1, 100 % or 1 million ppm (mg/kg). This is true whether or not all the members of the composition are determined, or expressed as metals, or metalloids ignoring elements such as O, S, carbonate C, the halogens, etc. As some concentrations increase others must decrease to maintain the constant sum. Thus, measures of correlation and covariance are constrained, and displays such as ternary and Harker diagrams do not display the true relationships, particularly for major elements. However, these diagrams are widely used in lithogeochemical studies for classifying individual analyses into particular fields characteristic of known rock types. What must be remembered is that these diagrams are unsuitable for interpreting petrogenic processes. Similarly, in studies of processes with other sample media, for example, soils and sediments, such plots will not lead to correct interpretations. In the lithogeochemical field this was recognized by Pearce (1968) who promoted the use of ratios for process interpretation. Subsequently, the work of Aitchison (1984, 1986) provided tools for handling the closure issue characteristic of compositional data. The investigation of compositional data and the development of methods for working with them has been a topic of extensive research in the last decade: see, for example, the papers in Buccianti *et al.* (2006), Egozcue *et al.* (2003), Filzmoser *et al.* (2009*a*, 2010), and tools to investigate compositional data sets, for example, Boogaart & Tolosana-Delgado (2008), van den Boogaart *et al.* (2011) and Templ *et al.* (2011*a*, *b*).

For univariate compositional data the logistic transformation is appropriate (Filzmoser 2009*a*). In practice at trace element levels, in fact up to as high as 10%, there is little practical difference between calculations undertaken with a logarithmic or logistic transform once the results have been back-transformed to the original measurement units. The Pearson correlation coefficient between the log and logit transformations of 800 values simulated at even intervals between 1 ppb and 10% is 0.9999968 (despite recommendations that estimates be quoted to only 3 or 4 significant figures), see Figure S1 in the Supplementary material.

For bivariate compositional data the use of ratios is recommended if the data are to be investigated for the purpose of understanding the processes affecting the data. The choice of element for the divisor needs to be undertaken with care. In some studies it makes sense to use a 'conservative' element if alteration and soil forming processes are the focus, for example, Zr or Hf. Titanium is often used, but minerals such as rutile are susceptible to alteration, and Ti may also be present in ilmenite. If a 'conservative' element is not required, Al often serves the purpose. Whatever element is chosen it should not be close to its detection or quantification limit and should not be rounded down to less than two non-zero digits if it is less than one. The plotting of such ratios with logarithmic scaling is equivalent to the additive log-ratio of Aitchison (1986). For truly multivariate data, centred log-ratio and isometric log-ratio transformations are available (Aitchison 1984, 1986; Egozcue *et al.* 2003; Filzmoser *et al.* 2010). However, it must be remembered that the centred log-ratio transformation yields different estimates for different subsets of elements (subcompositions).

Package 'rgr' contains functions for undertaking logistic, and additive, centred and isometric log-ratio transformations, and the multivariate data analysis functions identified as **.closed** are written to carry out the necessary transformations and back-transformations. Bivariate data inspections are supported by functions **gx.vm**, **gx.sm**, **gx.plot2parts** and **gx.pairs4parts**. When the purpose of a geochemical interpretation is to identify outliers, i.e. pure Exploratory Data Analysis (EDA), the most useful graphics may arise from using no transformations, not even a traditional logarithmic transformation. That EDA techniques have been used successfully in mineral exploration without addressing the closure issue is confirmation that data closure may not be an essential consideration. Lastly, our mineralogical training has imbued us with the principles of stoichiometry, an inherently closed compositional concept. The conflict between this reality and the mathematical concepts of closure is a topic of continuing research (Grunsky *et al.* 2008; Grunsky & Bacon-Shone 2011). Users are encouraged to investigate the effects of closure on their data using 'rgr', 'compositions' (van den Boogaart *et al.* 2011) and 'robCompositions' (Templ *et al.* 2011*b*) to try alternate approaches to their data, particularly if their objectives are to identify and understand geochemical processes.

Contents of package 'rgr'

The contents of 'rgr' may be broken down by functionality into eight groups containing functions for:

(1) Univariate statistical graphics;
(2) Mapping;
(3) Plotting;
(4) Summary statistics;
(5) Bivariate and multivariate statistics;
(6) QA/QC support;
(7) Data conditioning; and
(8) Utility operations.

In the sections below, the various functions are demonstrated with data from:

(a) The 2000 and 2001 National Geochemical Reconnaissance (NGR) stream sediment (<177-μm fraction) surveys undertaken in New Brunswick, Canada, covering four 1:50 000 scale map sheets, for details see Friske *et al.* (2002, 2003); or
(b) The Canadian Maritimes segment of North American Soil Geochemical Landscapes Project (Friske *et al.* 2011; Rencz & Kettles 2011), specifically the US-EPA 3050B aqua regia variant (US-EPA 1996) analyses of the <2 mm fraction of the 0–5 cm soils.

Univariate statistical graphics

The basic tools of statistical graphics for univariate, one-variable-at-a-time, data analysis are histograms, empirical cumulative distribution functions (ECDFs), boxplots of various kinds, and cumulative probability plots (CPPs) – the physical scientist's and engineer's version of the statistician's Q-Q or Q-normal plot. The function **shape** presents these four plot types in a single display (Fig. 1) with the Tukey boxplot (Tukey 1977) as the default rather than a box-and-whisker plot (Garrett 1988). Figures 1 to 3 are plots of untransformed As data; these clearly illustrates the presence of high concentration outliers, which a plot of the data with logarithmic scaling does not (see Fig. S2). Each of the component displays of **shape** may be plotted individually if required. Commonly, data interpretation benefits if the data can be subdivided on the basis of some categorical variable, such as some field data-related property, and displayed as Tukey boxplots, **tbplots** (Fig. 2), or box-and-whisker plots, **bwplots**. These displays are extremely effective and both 'rgr' functions allow extensive cosmetic enhancements in order to generate document-ready graphics which may be saved in a variety of graphical formats for

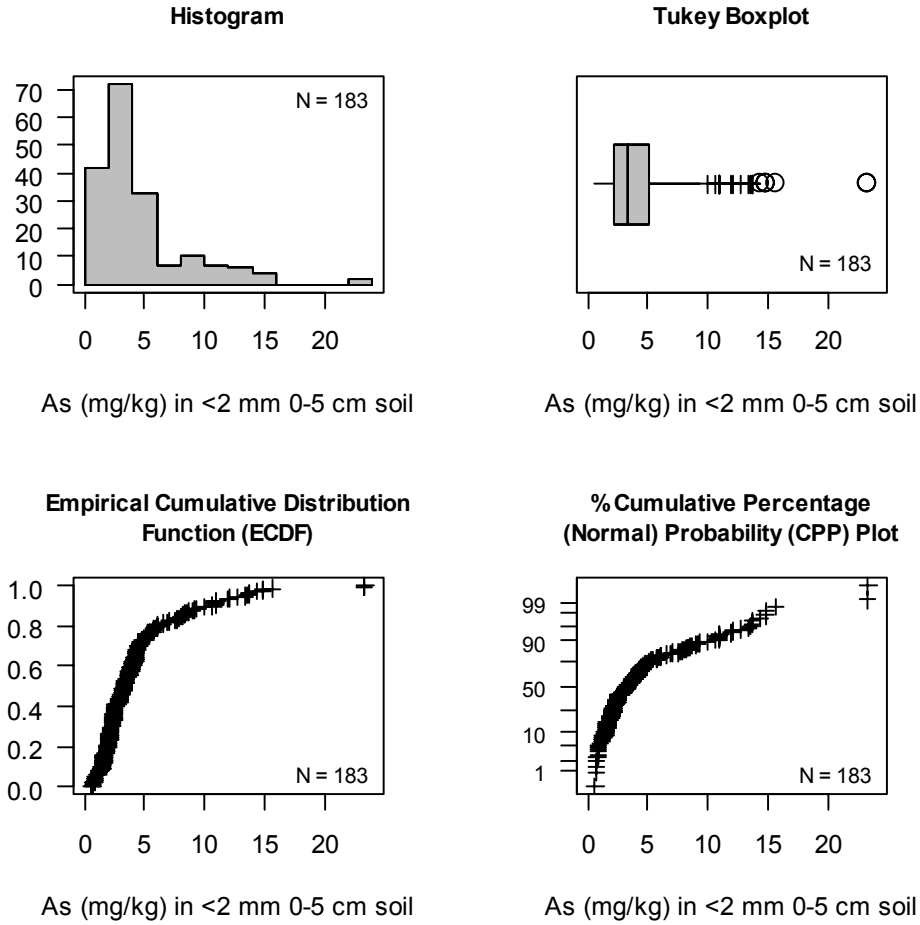

Fig. 1. Example of function **shape** – As (mg/kg) in <2 mm 0–5 cm soil, Maritime Provinces.

inclusion in reports, etc. One of the most insightful displays for studying univariate data in detail is the cumulative probability plot (cf. Q-Q plot), and 'rgr' has a function, **gx.cnpplts**, for plotting the data for up to nine data subsets (Fig. 3). The selection of colours and symbols with which to display the subsets may be changed if the defaults are unsuitable with function **gx.cnpplts.setup**. A natural extension of comparing probability plots is the Kolmogorov–Smirnov test for the comparison of any two data subsets (see, for example, Reimann *et al.* 2008). Function **gx.ks.test** undertakes this test and the results are displayed as an ECDF where no assumptions are made as to the underlying data distributions and the test is for the two data sets being drawn from the same underlying distribution (Fig. S3).

On some occasions the data sets for comparison may be determinations of different elements, or the same element by different methods of analysis. For such tasks there are functions **tbplots.by.var** (Fig. S4), **bwplots.by.var** and **gx.cnpplts**.

It is common practice to place an inset diagram on a map displaying a histogram and a cumulative probability plot together with some summary statistics, or to use a similar display as a report figure. Function **inset** prepares such a graphic (Fig. 4), and function **inset.exporter** automates the production of these displays for use on a series of geochemical maps. In both instances the graphics file may be saved in any of the graphics formats supported by R.

Mapping

A statistical graphics package like 'rgr' cannot replace a Geographic Information System (GIS). GISs permit complex operations in geographic space, such as searching for patterns of spatial coincidence, and the preparation of cartographic-quality map displays. For those interested in integrating R with GIS capabilities, attention is drawn to GRASS. GRASS is Open Source GIS software based on the Geographic Resources Analysis Support System developed by the U.S. Army Construction Engineering Research Laboratories (CERL) between 1982 and 1995. Subsequently, GRASS development and maintenance was taken over by its user community (GRASS 2011). The R packages **GRASS** and **spgrass6** are available on CRAN and provide interfaces between R and GRASS 5.0 and GRASS 6.0 (and greater), respectively (Bivand *et al.* 2008; Hengl 2009). Alternatively, GRASS functionality and an R interface are available in the Open Source Quantum GIS desktop software (QGIS 2011).

However, despite not having the functionality of a GIS, a series of functions is available in 'rgr' to take a 'quick-look' at data distributions in geographic space. This requires the user to provide spatial coordinates for the data sites; 'rgr' contains no spatial coordinate transformation functions, and any arbitrary rectangular coordinate system may be used. For small and medium sized areas it is convenient to use UTM coordinates expressed for a single UTM Zone; for larger areas Lambert Conformal or Albers Equal Area coordinates may be used.

Four spatial display functions are available, **map.eda7**, **map.eda8**, **map.tags** and **map.z**. The first two functions subdivide the data into Tukey boxplot-based groups and percentiles respectively, and may be displayed in colour or grey-scale. Thus for the Tukey boxplot display, the data are divided into seven groups, the central 50% of the data (the

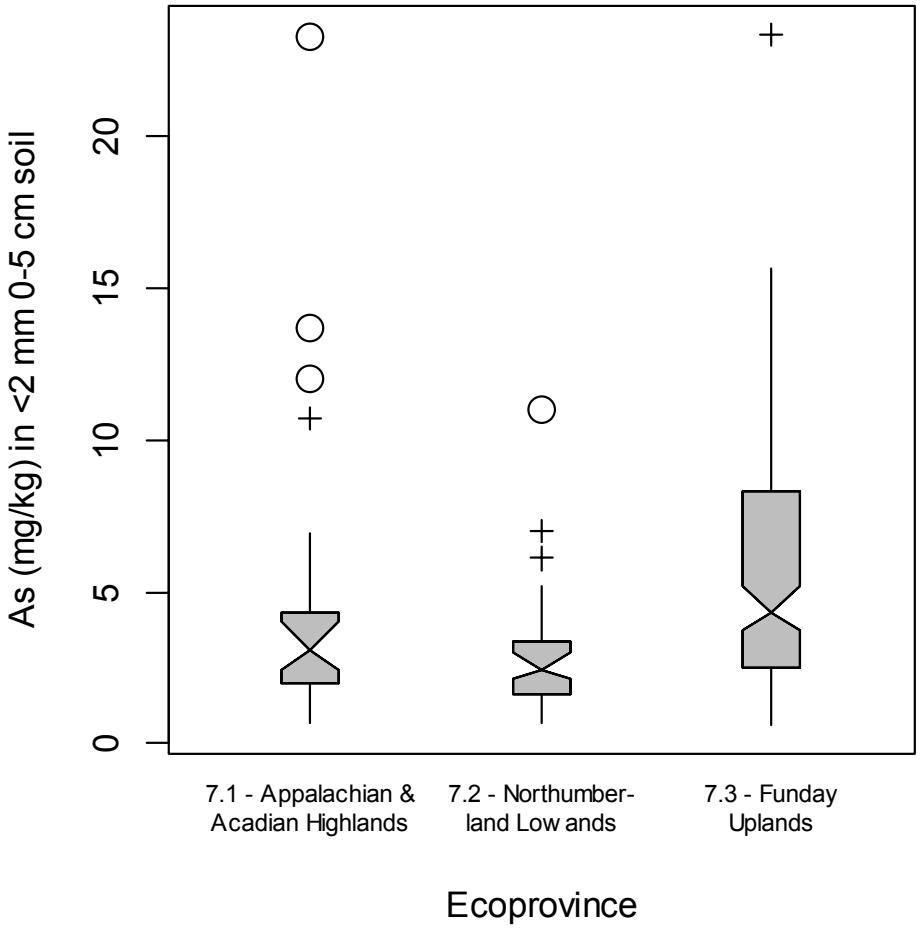

Fig. 2. Example of Tukey boxplots from function **tbplots** - As (mg/kg) in <2 mm 0–5 cm soil, Maritime Provinces, subdivided by Ecoprovince. The notches indicate the 95% confidence intervals on the medians.

box), the high and low 'background' data (the data represented by the whiskers in a boxplot) and the upper and lower near and far outliers (Fig. S5). The percentile-based display divides the data into eight groups with boundaries at the 5th, 10th, 25th, 50th, 75th, 90th and 95th percentiles (Fig. S6). In both displays, squares of various sizes are used to display the groups above the median, and circles the groups below it, with symbol size increasing away from the median, or a '+' for the central 50% group. Function **map.z** displays the data as a set of increasing sized circles, whose absolute size is controlled by a user-set parameter, **sfact**, with a default value of 2.5. The rate at which the circles increase in diameter is controlled by a user-set parameter, **p**, varying about 1, usually in the range of 0.2 to 5, with a default value of 0.5 (Fig. 5), a value of 1.0 results in a linear rate of increasing diameter across the data range. For values of **p** less than 1.0 the symbol size initially rises rapidly above the minimum value, and conversely for values above 1.0, function **syms.pfunc** provides a display of the effects of changing **p**. The specific calculations are in function **syms** and described in the online hypertext help and printable manual. Legends may be optionally added to all these maps indicating the range of values that each symbol represents, or the size of the proportional symbol for the minimum and maximum values and the three quartiles. For the percentile-based map the legend can optionally display the percentiles, or the values of those percentiles (Fig. 5). If required the data may be transformed prior to diameter calculation, for example, to logarithms or square roots, additionally an upper and/or lower limit may be set beyond which all plotted values result in the same sized symbol. This latter feature is provided to enable a useful display range for the majority of the data – once observations are anomalously high or low it may not matter for the required display, and they may be displayed similarly sized. Lastly, function **map.tags** simply posts a value at the geographic location of the sample site, and can be used to plot sample identification numbers (IDs). This latter function is not suitable for spatially dense data because the size of the display would result in overplotting.

Fractal geometry has proven to be a useful tool for investigating possible threshold and other data-grouping boundaries (Cheng *et al.* 1994; Cheng & Agterberg 1995). The Concentration-Area (C-A) procedure is available as function **caplot**, and in addition to the C-A plot a coloured or grey-scale interpolated map is displayed (legend-less) by the procedure. Optionally the C-A plot may be displayed accumulating from the highest to the lowest values (i.e. the traditional display), or vice versa. The latter is useful when the different fractal patterns extend over large portions of the study area as in some environmental contamination studies (see Reimann *et al.* 2008). Figure S7 displays the C-A plot for the As data in the Maritime Provinces. The data set is not ideal, as in creating the limiting boundary (a convex hull that includes all data points) large areas of 'no data' are included in the Bay of Fundy and between northern Nova Scotia and New Brunswick. In fact, in this particular instance a C-A plot provides no additional information than can be gained from traditional cumulative probability plot inspections (see Figs 1 & 3).

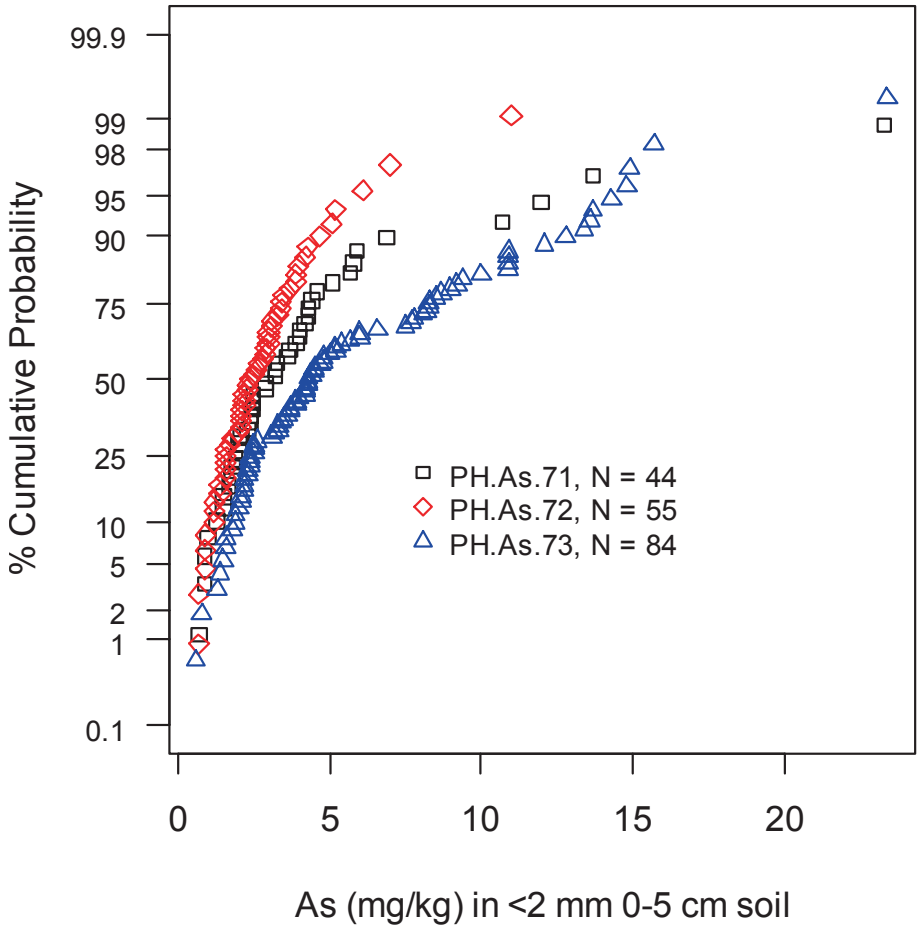

Fig. 3. Example of function **gx.cnpplts** for plotting cumulative probability plots for up to nine data subsets - As (mg/kg) in <2 mm 0–5 cm soil, Maritime Provinces, subdivided by Ecoprovince. See Figure 2 for identification of Ecoprovinces.

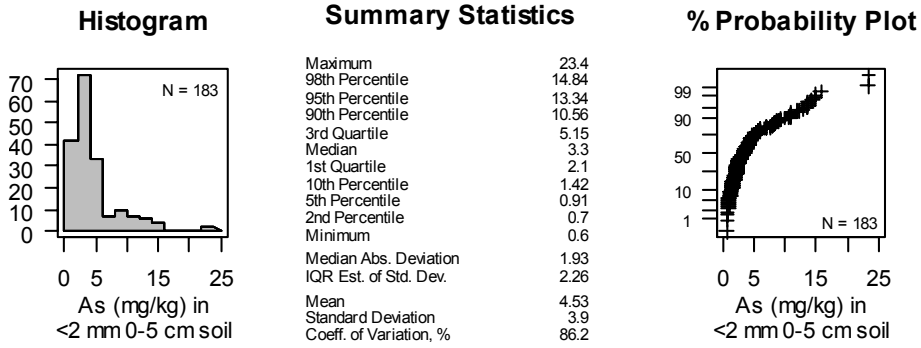

Fig. 4. Example of function **inset** – As (mg/kg) in <2 mm 0–5 cm soil, Maritime Provinces.

Bivariate data plotting

In spatial mapping a variable is displayed at its location in geographic space. Four plotting functions, **xyplot.eda7**, **xyplot.eda8**, **xyplot.tags** and **xyplot.z**, that are similar in style to the four mapping functions described in the previous section, are provided for use in the geochemical space. These are particularly useful in displaying trace element data in the context of major element data, for example, As in the context of Fe, and organic carbon for soils; in the examples below, results from the proportional symbol version, **xyplot.z**, are displayed. These plots are essentially enhanced Harker diagrams as used by petrologists, and are simply useful data displays that present data in a familiar context (Fig. 6). However, they are burdened with all the problems of Harker diagrams with respect to data closure (Aitchison 1984, 1986; Filzmoser *et al.* 2010) and their inability to support insightful process-related data interpretation. The closure issue can be overcome by plotting the ratios of major components to another element in the composition with log scaling, essentially an Aitchison additive log-ratio transformation (Fig. 7). A comparison of the two displays is informative, and indicates that: (1) the inverse relationship between Fe and organic carbon in Figure 6 is only part of the story; (2) in reality the Fe content remains essentially constant in the minerogenic fraction of the soil, as measured by Al, as organic C increases, increasing only slightly with respect to decreasing organic carbon; and (3) high As levels are associated with higher Fe in the soils. In Figures 6 and 7 the As data are plotted without transformation; Figures S8 and S9 in

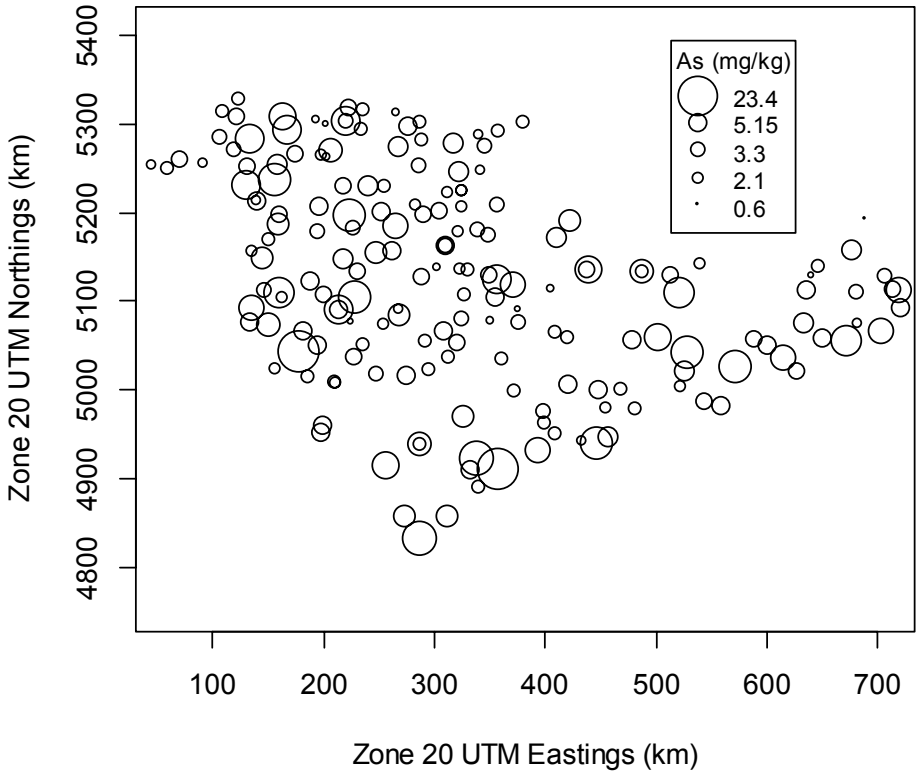

Fig. 5. Example of function **map.z** – As (mg/kg) in <2 mm 0–5 cm soil, Maritime Provinces.

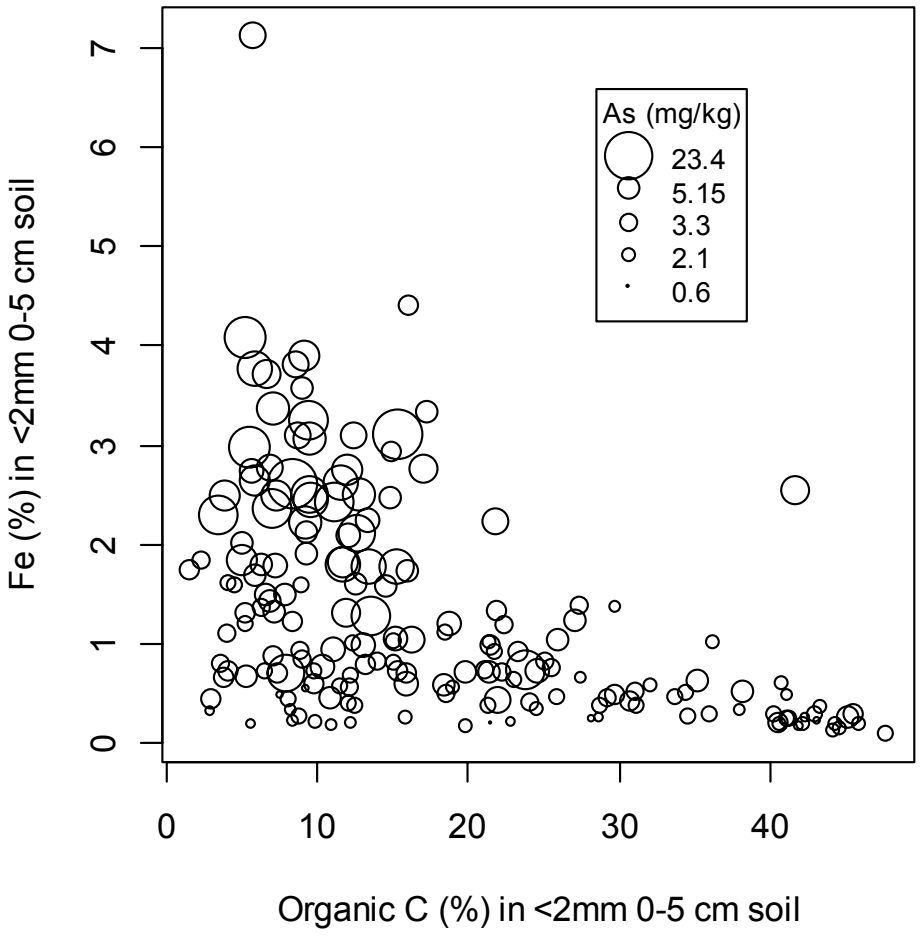

Fig. 6. Example of function **xyplot.z** – As (mg/kg) in <2 mm 0–5 cm soil, Maritime Provinces, plotted as a traditional Harker diagram in the Fe – organic C framework.

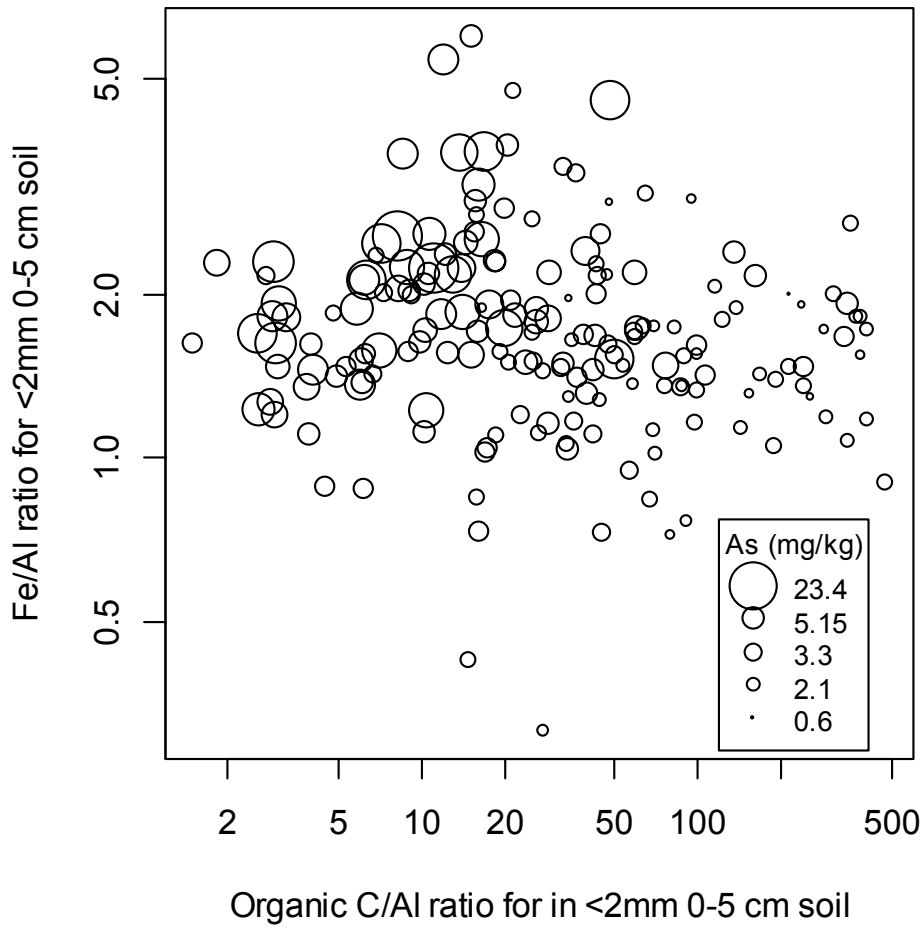

Fig. 7. Example of function **xyplot.z** – As (mg/kg) in <2 mm 0–5 cm soil, Maritime Provinces, plotted into an additive log-ratio framework to remove the effects of closure on Fe and organic C concentrations.

Supplementary material display the As plotted on a logarithmic scale, and as the As/Al ratio, respectively. Plotting As on a logarithmic scale (Fig. S8) confirms that As is preferentially located in the minerogenic component of the soil, because of the equivalence of scaling after 10%. However, plotting the log of the As/Al ratio, the full additive log-ratio treatment (Fig. S9), indicates higher As/Al levels across the whole range of the data. Another insightful way to study the relationship between Fe, Al, organic carbon and As is with function **gx.pairs4parts (**Fig. 8). The displays in the upper triangle are log-log plots, those in the lower triangle are Tukey boxplots of the isometric log-ratio (ilr) for each plotted pair, and the number above each boxplot is the robust ilr stability (Filzmoser *et al.* 2010). The ilr stability measure quantifies the constancy of the log-ratio of the pairs, referred to as 'parts' in compositional data analysis. If two-element 'parts' have a systematic linear relationship, despite the effects of dilution by other members of the composition, the log-ratio will have a small range and the stability measure approaches 1. It can be thought of as a measure of linear correlation; however, it only has positive values, and hence in 'rgr' is referred to as a stability. A more detailed display for any pair is available with function **gx.plot2parts** (Fig. S10), which displays the Fe–organic carbon relationship. The ilr stability measure is low (0.2), indicating the log-ratios vary widely and that there is an inconsistent relationship between the two 'parts'; however, that inconsistency is a key feature of the soil geochemistry. These examples provide a demonstration of the complexities that arise between geochemists' knowledge of stoichiometry and surficial geochemistry, the sequestration of As with Fe-oxyhydroxides, and the mathematical approach to compositional data sets. Functions to display ternary diagrams are not present in 'rgr', though they were in the previous S-PLUS library. Those interested in using and investigating ternary diagrams for compositional data sets in the R environment should use 'compositions' (van den Boogaart & Tolosana-Delgado 2008; van den Boogaart *et al.* 2011) and robCompositions (Templ *et al.* 2011*a*, *b*). Similarly, those wishing to use petrochemical displays in the context of established petrological classifications should investigate the Geochemical Data Toolkit (GCDkit) package (Janoušek *et al.* 2006, 2011). Note, however, that GCDkit is not available on CRAN and therefore is not available for Mac and Unix operating systems, though it may be run via a Windows emulator (Janoušek *et al.* 2011).

Attention is drawn to the base R function **pairs** that plots a matrix of x-y scatterplots for a data matrix. Depending on the data displayed there may, or may not, be a closure issue when it comes to interpretation of the plots (Filzmoser *et al.* 2010). However, **pairs** is a useful EDA graphical display, and **gx.pairs4parts** has been written for use with compositional data.

Summary statistics

All summary statistics are computed with a single function, **stats**, and the required estimates are then extracted from the saved object and displayed as required, for example, by **inset** (Fig. 4). Tabular displays are generated by functions **gx.summary1** and **gx.summary.2** for single specified numeric variables

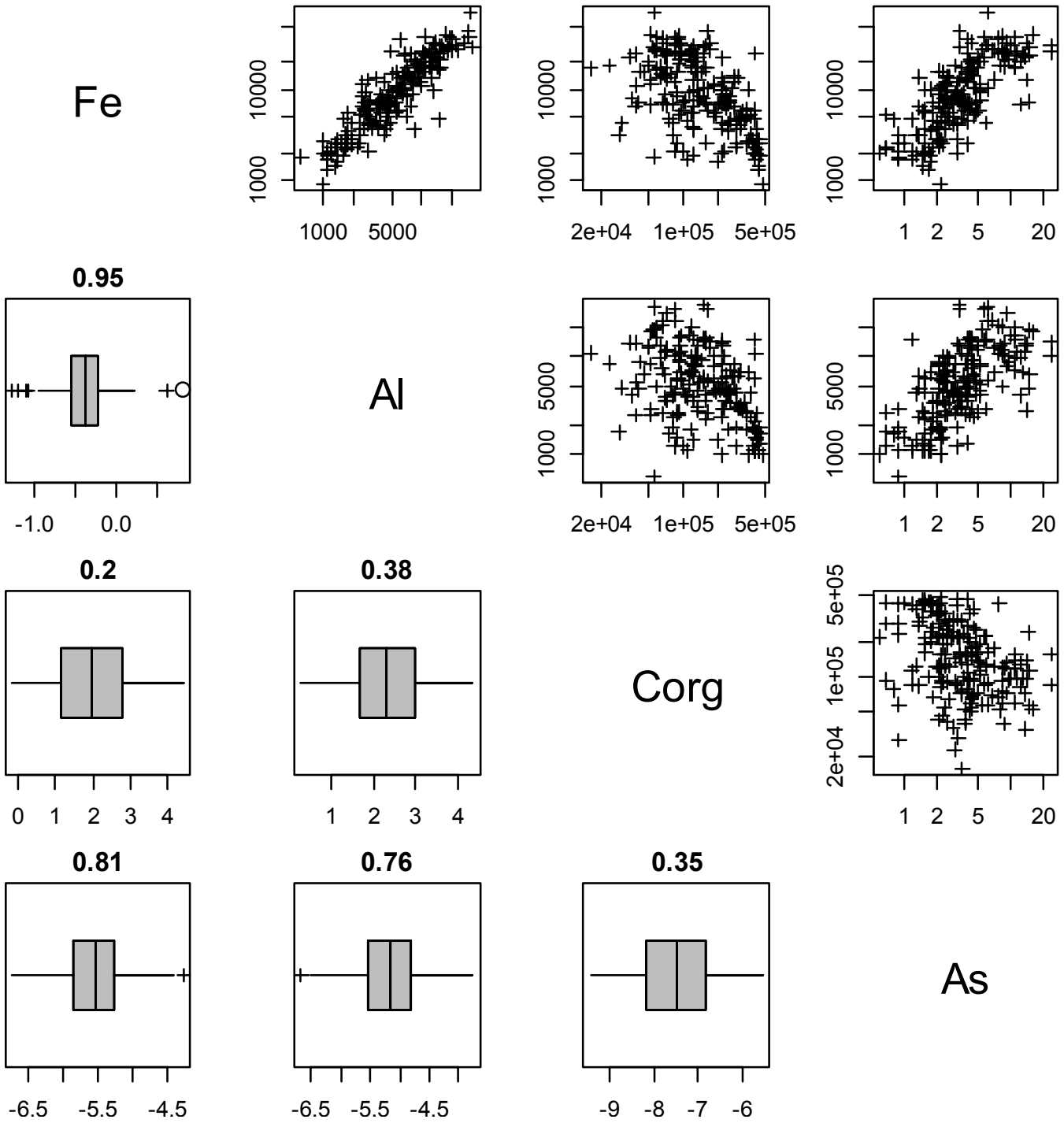

Fig. 8. Example of function **gx.pairs4parts** for the study of Fe, Al, organic C and As relationships; plots in the lower triangle are Tukey boxplots of the ilr transforms of the pairs in the upper triangle. The boxplot titles are the respective robust ilr stability measures.

(see Tables 1(A) and 1(B), respectively). The former displays a one-line summary consisting of the data set size, minimum, maximum, quartiles, MAD and an Interquartile Range-based estimate of the SD, mean, SD, coefficient of variation (CV%), standard error of the mean (SE), and the 95% confidence bounds on the mean. The function **gx.summary.2** displays a more extensive eight-line summary, with the addition of the geometric mean and its 95% confidence bounds, the mean and SD for the log10 transformed data, the 95% confidence bounds on the median, and the 2nd, 5th, 10th, 90th, 95th and 98th percentiles.

The estimates in **gx.summary1** may be displayed for selected variables from a data matrix with function **gx.summary.mat** by specifying the name of the matrix and the columns for which estimates are required (Table 1(C)). Similarly, in function **gx.summary.groups** specifying the name of a variable and a factor (the grouping variable, for example, a character string for soil or rock type), results in a table being displayed of statistical estimates subdivided by the unique character strings present in the factor variable (Table 1(D)).

Function **framework.summary** performs in a similar way to **gx.summary.groups** but outputs the resulting table to a

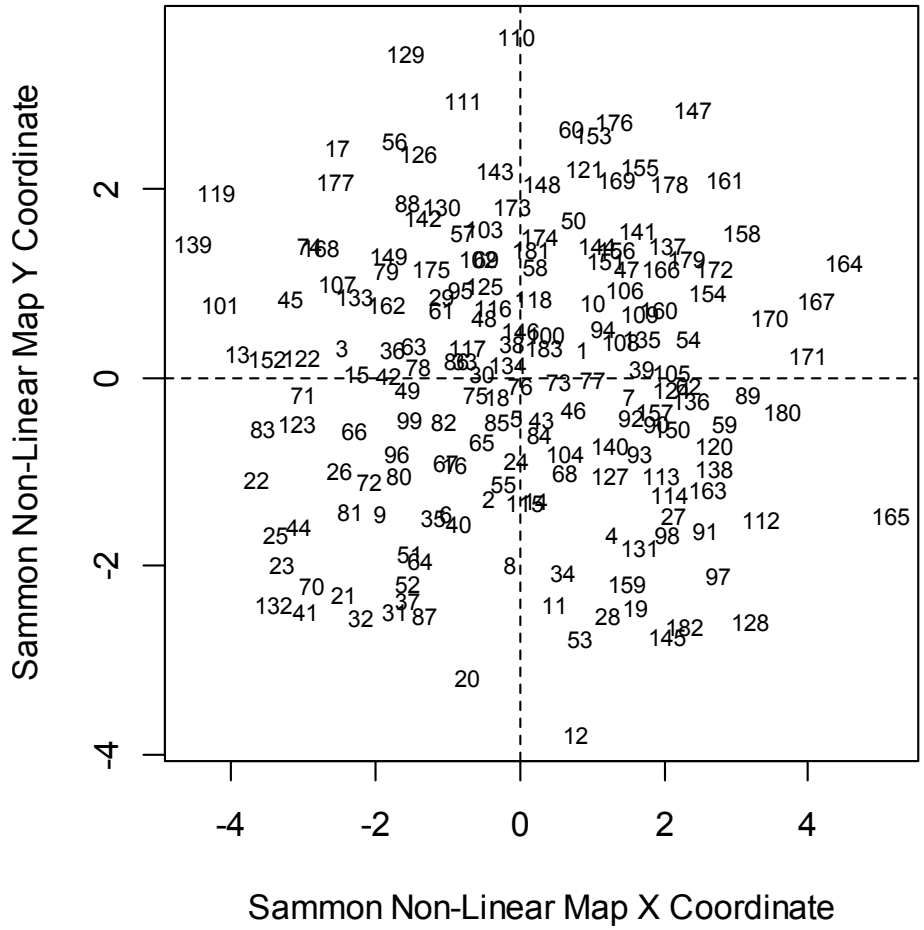

Fig. 9. Example of the results from function **gx.2dproj**, plotted by **gx.2dproj.plot**, applied to the major and trace element data for 0–5 cm <2 mm soils, Maritime Provinces, following an isometric log-ratio transformation.

csv file. The table includes data subset size, the percentiles displayed by **gx.summary.2**, the 95% confidence bounds on the median, non-parametric estimates of spread (MAD, etc.), the mean, SD and CV%. The resulting table can be displayed with a spreadsheet program and imported into a GIS. In the former instance it can be edited for use in a report or publication. In the latter, when the data subsets are geographic entities, geochemical maps may be prepared where the entities are displayed according to a selected statistical estimator, for example, the median elemental content (Fig. S11, prepared with the ESRI ArcMap GIS).

The statistical estimation of thresholds is supported by function **fences** which displays all the estimators discussed in Reimann *et al.* (2005) for a single specified variable. Function **fences.summary** operates in a manner similar to **gx.summary.groups** and generates 'fences' for all the groups (data subsets), which are uniquely identified by the grouping variable (factor) (Table 2). The 'fences' are estimated as the 2nd and 98th percentiles, mean ± 2SD, median ± 2MAD, and the inner fence estimated as for a Tukey boxplot, i.e. all computations without transformation and with log and logit transformations of the data. This leads to ten estimates for the upper limit of normal background variation (Table 2), and, needless to say, few are identical; at trace element levels log and logit transformations yield very similar, if not identical, estimates. Therefore, the selection of thresholds has to be based on careful graphical inspections of the data in both geochemical (see Fig. 3) and geographical spaces (Reimann *et al.* 2005) before geochemically defensible estimates can be made. The output from **fences.summary** may be optionally saved to a tab-delimited text file, in which case there is no screen display. By virtue of the tab delimitation, the saved text file can be opened with a spreadsheet program for display in tabular form and quickly edited into a publication-ready table.

Bivariate and multivariate statistics

Three functions support traditional bivariate statistical operations, **gx.pearson**, **gx.spearman** and **gx.rma**, and two functions, **gx.vm** and **gx.sm**, support compositional data analysis. The first two simply repackage the Pearson product moment and Spearman rank correlation matrix functions in R, placing the correlation coefficients in the upper right triangle of the displayed matrix, and placing the significance of the coefficients not being due to chance (Ho: coefficient = 0) in the lower left triangle. Both functions permit the data to be log or centre-log-ratio (Aitchison 1986) transformed prior to computation. However, the application of a centred log-ratio transformation does not completely solve the problem of closure. Different subcompositions, i.e. sets of different measured variables, will yield different correlation matrices between the same pairs of variables. For this reason traditional correlation analysis needs to be undertaken with great caution; some would argue that it should not be undertaken at all. For compositional data analysis, function **gx.vm** displays the Aitchison Variation Matrix (Aitchison 1984, 1986), and function **gx.sm** provides a robust equivalent employing the Filzmoser *et al.* (2010) stability measure. However, it should be remembered

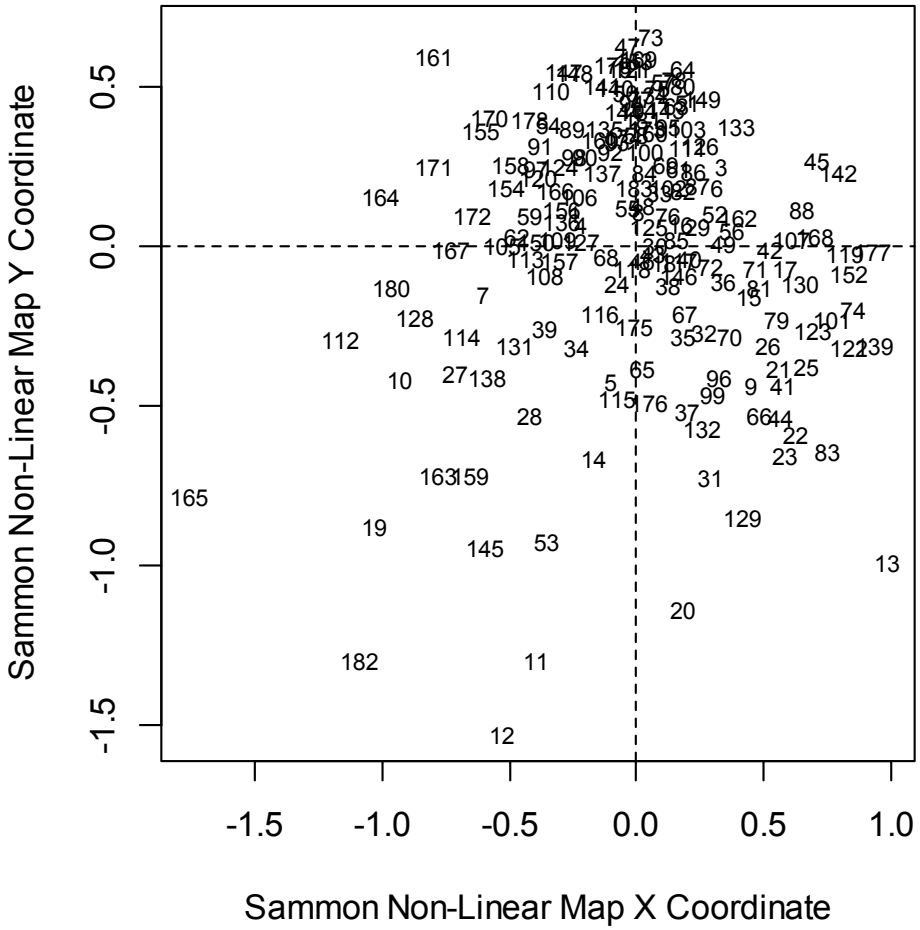

Fig. 10. Example of the results from function **gx.2dproj**, plotted by **gx.2dproj.plot**, applied to the major and trace element data for 0–5 cm <2 mm soils, Maritime Provinces, following a range transformation.

that function **gx.sm** is still based on linear relationships, and, although robust, is not an equivalent to Spearman rank correlations, which are good measures of non-linear but monotonic relationships. Functions **gx.pearson** and **gx.spearman** are provided in 'rgr' as there are instances where compositional values may be studied in relation to some non-compositional value, for example, sample depth or distance from a point source; or relationships within the same subcomposition may be compared between different studies.

Function **gx.rma** estimates the coefficients, and their 95% confidence bounds, for the Reduced Major Axis (RMA), a.k.a. Orthogonal Regression, for the situation when both the variables are independent and drawn from different compositions (Miller & Kahn 1962). For other regression tasks, the base R functions perform both Ordinary and Non-Linear Least-Squares and Robust Regression modelling. The RMA is appropriate for studying situations where the same variable has been measured by two independent methods on the same physical samples, and the Reduced Major Axis may be added to a plot of the two data sets. Table 3 and Figure S12 provide an example of Ni determined in the same samples of <2 mm 0–5 cm soil by aqua regia and 4-acid digestions. As the confidence limits on the slope and intercept estimates include 1 and 0, respectively, it can be inferred that the two sets of analyses are indistinguishable from one another at the 95% confidence level.

Weighted sums (Garrett *et al.* 1980; Garrett & Grunsky 2001) generate a single number that represents the combination of a number of variables, either a sum of the variables or a ratio if the data are logarithmically transformed. Weighted sums quantify a knowledge-based relationship known to the geochemist. Techniques such as Principal Components Analysis (PCA) are data-based and if a particular multivariate relationship is rare, reflecting a very small amount of the total variability, it will most likely be buried in the 'noise' of the components that reflect only small amounts of the total variability. In weighted sums, the geochemist specifies the relative importance of the variables or informative ratios, and whether high or low values of the variable are favourable or unfavourable to the outcome, using positive or negative signs, respectively. Weights are then computed such that when the data for the variables are transformed to mean zero and SD one and multiplied by the weights, the resulting weighted sums are also mean = 0 and SD = 1 distributed. Thus the resulting weights may be interpreted probabilistically as standard normal deviates. The procedure benefits from using the median and MAD, rather than the mean and SD, for the estimates of background for each variable; this has the effect of accentuating the differences from background. The function **wtd.sums** carries out all the necessary computations with the user, providing the variables and their relative importance with a default use of the medians and MADs. The results are saved in an object and may be inspected and plotted as required in either the geochemical or geographical space.

There are numerous multivariate data analysis tools in R. The 'rgr' package contains procedures for undertaking four

Table 1. *Examples of summary statistics output using trace element data for 0–5 cm <2 mm soil, Maritime Provinces – (A) function* **gx.summary1***; (B) function* **gx.summary2***; (C)* **gx.summary.mat***; and (D)* **gx.summary.groups** *subdivided by Ecoprovince*

(A)
> gx.summary1(As,"As (mg/kg) in <2 mm 0-5 cm soil")
Summary Stats for: As (mg/kg) in <2 mm 0-5 cm soil
N NA - Min Q1 M Q3 Max - MAD IQR_SD - Mean SD CV% - SE & 95% CLs on Mean
183 0 - 0.6 2.1 3.3 5.15 23.4 - 1.927 2.261 - 4.528 3.903 86.2% - 0.2885 3.959 <-> 5.097
(B)
> gx.summary2(As,"As (mg/kg) in <2 mm 0-5 cm soil")
Extended Summary Stats for: As (mg/kg) in <2 mm 0-5 cm soil
N and number of NAs: 183 & 0
Arithmetic Mean and 95% CLs: 4.528 & 3.959 <-> 5.097
SD and CV%: 3.903 & 86.2%
Geometric Mean and 95% CLs: 3.403 & 3.053 <-> 3.793
Log10 Mean and SD: 0.5319 & 0.3232
Median and 95% CLs: 3.3 & 2.9 <-> 3.9
MAD and IQR estimates of SD: 1.927 & 2.261
Percentiles: Min 2 5 10 25 - 50 - 75 90 95 98 Max
0.6 0.764 0.91 1.42 2.1 - 3.3 - 5.15 10.56 13.34 14.84 23.4
(C)
> gx.summary.mat(PH.Majors.TEs.EcoP,c(10,12,15,4),"As, Cu, Zn & Fe in (mg/kg) in <2 mm 0-5 cm soil")
Summary Stats for As, Cu, Zn & Fe in (mg/kg) in <2 mm 0-5 cm soil
N NA - Min Q1 M Q3 Max - MAD IQR_SD - Mean SD CV% - SE 95% CI on Mean
As: 183 0 - 0.6 2.1 3.3 5.15 23.4 - 1.927 2.261 - 4.528 3.903 86.2% - 0.2885 3.959 <-> 5.097
Cu: 183 0 - 1.31 4.995 7.58 10.54 73.26 - 3.944 4.111 - 8.405 6.235 74.18% - 0.4609 7.496 <-> 9.315
Zn: 183 0 - 6.3 25.3 40.9 66.1 245.2 - 28.76 30.25 - 51.43 37.55 73.02% - 2.776 45.95 <-> 56.9
Fe: 183 0 - 900 4350 8300 17900 71300 - 8154 10040 - 12520 10980 87.69% - 811.7 10920 <-> 14120
(D)
> gx.summary.groups(EcoP,As,"As (mg/kg) in <2 mm 0-5 cm soil grouped by Ecoprovince")
Summary Stats for As (mg/kg) in <2 mm 0-5 cm soil grouped by Ecoprovince subset by EcoP
N NA - Min Q1 M Q3 Max - MAD IQR_SD - Mean SD CV% - SE 95% CI on Mean
7.1: 44 0 - 0.7 1.975 3.05 4.325 23.3 - 1.853 1.742 - 4.12 4.046 98.2% - 0.61 2.89 <-> 5.351
7.2: 55 0 - 0.7 1.6 2.4 3.4 11 - 1.186 1.334 - 2.805 1.755 62.54% - 0.2366 2.331 <-> 3.28
7.3: 84 0 - 0.6 2.475 4.3 8.3 23.4 - 3.113 4.318 - 5.869 4.364 74.36% - 0.4762 4.922 <-> 6.816

tasks that have been found useful in geochemical investigations, and for displaying the results in ways familiar to geochemists:

(1) Projection Pursuit, where p-space data are projected onto a 2-d plane for display;
(2) Principal Component Analysis;
(3) Multivariate Chi-square probability plots; and
(4) Allocation of data into the most similar of various subpopulations.

Within the latter three general areas there are functions that allow for the fact that data may be compositional, i.e. closed, and sum to a constant, such as geochemical analyses. For projection pursuit procedures, the data may be appropriately transformed to allow for closure prior to computation. An outstanding research issue is the development of heuristics to inform if trace element data sets representing small proportions of the total composition, say less than 1%, should be transformed to allow for closure, or if in those cases some other transformation, for example, logarithmic, may be adequate or preferable.

A key issue in applying multivariate statistical procedures is to ensure that there are sufficient observations for the number of variables. An analogy is that you can fit a straight line through any two points, or a plane to three points, perfectly, no matter how unrepresentative those points are of the underlying data distribution from which they are drawn. Several of the multivariate analysis functions display warning messages if the number of observations is, or falls below, five times the number of variables; even more dire warnings are displayed if the ratio falls below three to one. There is no generally accepted statistical rule to define the critical point below which the ratio of observations to variables should not fall, and some statisticians argue for ratios as high as eight or nine to one (Garrett 1993). If a data set has too few samples for the number of variables, not an unusual occurrence with today's ICP–AES and ICP–MS analytical techniques, consideration should be given to using only a subset of the variables for analysis. For instance, are all the rare earth elements (REEs) required? Would a reliably and commonly understood light REE and heavy REE suffice? Often an initial univariate inspection of the data with function **shape** will indicate if the data possess any interesting features, or if some are just bland Gaussian distributions. In the latter case they could be candidates for removal prior to multivariate data inspection. Similarly, the inspection of ratios and displaying the bivariate relationships with the R **pairs** or **gx.pairs4parts** functions can help identify variables with interesting off-axis behaviour (Garrett 1993) for retention.

Unless otherwise stated, in the following examples a subset of the North American Soil Geochemical Landscapes Project US-EPA 3050B aqua regia variant data for the <2 mm fraction of the 0–5 cm soils have been used. The 14 elements in the subset (N = 183) are: Al, Ca, Fe, K, Mg, Mn, Na, Organic C (Corg), As, Cd, Cu, Ni, Pb and Zn, all expressed in mg/kg.

Projection pursuit methods are useful for data inspection to determine if the data fall into disjoint clusters, in which it might be beneficial to divide the data into subsets, and to identify outliers. Function **gx.2dproj** provides four different projection pursuit solutions:

(1) Sammon's Non-Linear Mapping (Sammon 1969; Garrett 1973; Howarth 1973);
(2) Multidimensional Scaling, also known as Principal Coordinate Analysis (Venables & Ripley 2001);
(3) Non-metric Multidimensional Scaling (Cox & Cox 2001); and
(4) Independent Component Analysis (Hyvarinen & Oja 2000).

The functions to undertake these operations are available through packages 'MASS' and 'fastICA' available on CRAN; 'rgr' function **gx.2dproj** is a 'wrapper' that simplifies investigation of the various project pursuit procedures and their display. Function **gx.2dproj.plot** is provided to give additional display options. Unfortunately the function that undertakes the Friedman & Rafsky (1981) Minimum Spanning Tree procedure (Garrett 1983; Reimann *et al.* 2008) is available in the S-PLUS version of 'MASS' but not in the R implementation. Additional tools for cluster analysis within R and for geochemical data are available in the clustTool package (Templ *et al.* 2008; Templ 2010).

As an example of projection pursuit, the results of the Sammon's Non-Linear Mapping procedure, following an isometric log-ratio transformation in recognition of the compositional

Table 2. *Example of function* **fences.summary** *- As (mg/kg) in 0–5 cm <2 mm soils, Maritime Provinces, subdivided by Ecoprovince*

> fences.summary(EcoP,As,file="M2007_PH")

Variable As subset by EcoP - output will be in M2007_PH_EcoP_As_fences.txt

As [7.1] : N = 44 NAs = 0						2%ile = 0.872	98%ile = 15	
	Mean	SD	Median	MAD		Mean±2SD	Med±2MAD	Tukey Fences (actual)
	4.12	4.05	3.05	1.85	+	12.2	6.76	7.85 (6.9)
					−	−3.97	−0.656	−1.55 (0.7)
Log10	0.489	0.318	0.484	0.268	+	13.3	10.5	14.0 (13.7)
					−	0.714	0.887	0.609 (0.7)
Logit	−12.7	0.731	−12.7	0.617	+	13.3	10.5	14.0 (13.7)
					−	0.714	0.707	0.609 (0.7)
As [7.2] : N = 55 NAs = 0						2%ile = 0.716	98%ile = 6.93	
	Mean	SD	Median	MAD		Mean±2SD	Med±2MAD	Tukey Fences (actual)
	2.81	1.75	2.4	1.19	+	6.31	4.77	6.1 (5.2)
					−	−0.704	0.0278	−1.1 (0.7)
Log10	0.379	0.247	0.38	0.243	+	7.47	7.35	10.5 (7)
					−	0.766	0.784	0.517 (0.7)
Logit	−12.9	0.569	−12.9	0.559	+	7.47	7.35	10.5 (7)
					−	0.766	0.768	0.517 (0.7)
As [7.3] : N = 84 NAs = 0						2%ile = 1.13	98%ile = 15.2	
	Mean	SD	Median	MAD		Mean±2SD	Med±2MAD	Tukey Fences (actual)
	5.87	4.36	4.3	3.11	+	14.6	10.5	17.0 (15.7)
					−	−2.86	−1.93	−6.26 (0.6)
Log10	0.655	0.324	0.633	0.393	+	20.1	26.3	51.0 (23.4)
					−	1.01	0.703	0.403 (0.6)
Logit	−12.3	0.747	−12.4	0.905	+	20.1	26.3	51.0 (23.4)
					−	1.01	0.966	0.403 (0.7)

Note: The numbers quoted as (actual) for the Tukey Fences are the actual data values immediately within the calculated fences

Table 3. *Example of function* **gx.rma** – *The relationship between Ni determined after aqua regia and 4-acid digestions*

> gx.rma(log10(Ni.2mm.3050B),log10(Ni.2mm.4acid))

Reduced Major Axis for log10(Ni.2mm.3050B) and log10(Ni.2mm.4acid)

	log10 (Ni.2mm.3050B)		log10 (Ni.2mm.4acid)
Means =	1.464		1.489
SDs =	0.2833		0.277
Corr =	0.8816		
N =	79		
		SE	95% CLs
Slope =	0.9779	0.05192	0.8745<->1.081
Intercept =	0.05781	0.0774156	-0.09631<-> 0.2119

form of the data, are presented in Figure 9 where the plotted numbers are the data row identifiers. This display does not lead to the identification of any extreme outliers. In this respect the log-ratio transform, while being appropriate on statistical grounds, fails as an EDA tool. It is common practice in cluster analyses to transform the variables so that each has similar weight by scaling the data to range in value from 0 to 1, corresponding to the data minima and maxima. Figure 10 presents the result of using that procedure, and extreme outliers from the main mass of the data can be identified by their data row numbers. Function **gx.2dproj** includes, in addition to range transformations, options for log transformations and standard (mean and SD) and robust (median and MAD) normalization, and logit transformations may be made externally within 'rgr'. Figures S13 and S14 show similar plots for a log followed by range transformation, and a logit followed by range transformation, respectively. The logit transformation was employed as organic carbon levels approach 50%. Neither display leads to any improvement in outlier identification. In comparing all four plots, the same individuals appear close to the perimeter of the display, but it is with the range-transformed raw data that the outliers are most obvious.

Principal Components Analysis (PCA) and Factor Analysis (FA) have a long history of application in the earth sciences. One of the earliest applications was in litho-stratigraphy by Krumbein & Imbrie (1963), and both PCA and FA are discussed in the context of applied geochemistry in Reimann *et al.* (2008). Only functions for PCA are included in 'rgr'; functions for FA are available in R, as are alternate functions for PCA. The functions **gx.mva**, **gx.robmva** and **gx.robmva.closed** in 'rgr' follow the R script for R-Q analysis published by Grunsky (2001) that simultaneously estimates the loadings of the variables on the Principal Components (PCs) and the scores of the individuals (samples) on the PCs. Figure 11 presents the PC loadings with absolute values >0.3 (function **gx.rqpca.loadplot)**, together with the cumulative proportions of the variability explained, undertaken with function **gx.mva** following a centred log-ratio transformation in recognition of the compositional form of the data. It is important to note that loadings on PC-1 are both negative and positive, facilitating interpretation. In this instance negative PC-1 loadings relate to femic mineral-related elements, and positive PC-1 loadings relate to felsic and carbonate parent materials, together with an organic carbon-Cd-Pb-Zn, association. When closure is not allowed for and either no transformation or a logarithmic transformation is employed all the loadings, in this instance, are negative, making interpretation more difficult (Figs S15 and S16, respectively). Additionally, there is little difference between the pattern of loadings in the first two components, but a greater proportion of the variability is contained in fewer components

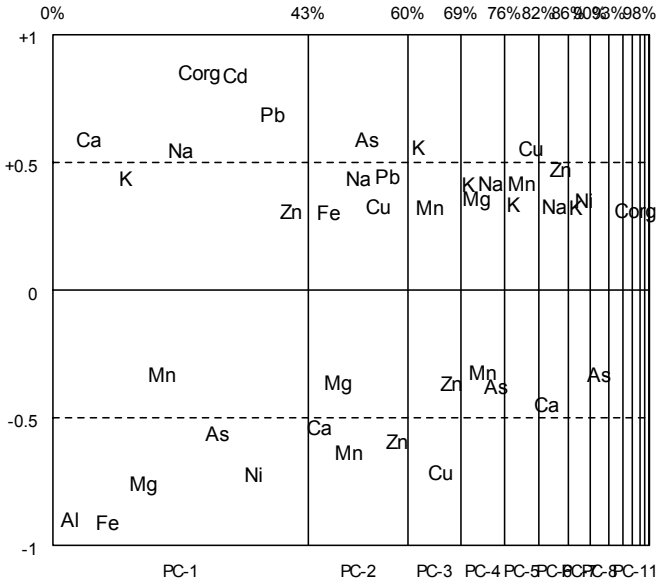

Fig. 11. Graphical display by function **gx.rqpca.loadplot** of PC loadings for the Maritime Provinces major and trace element suite following a centred log-ratio transformation.

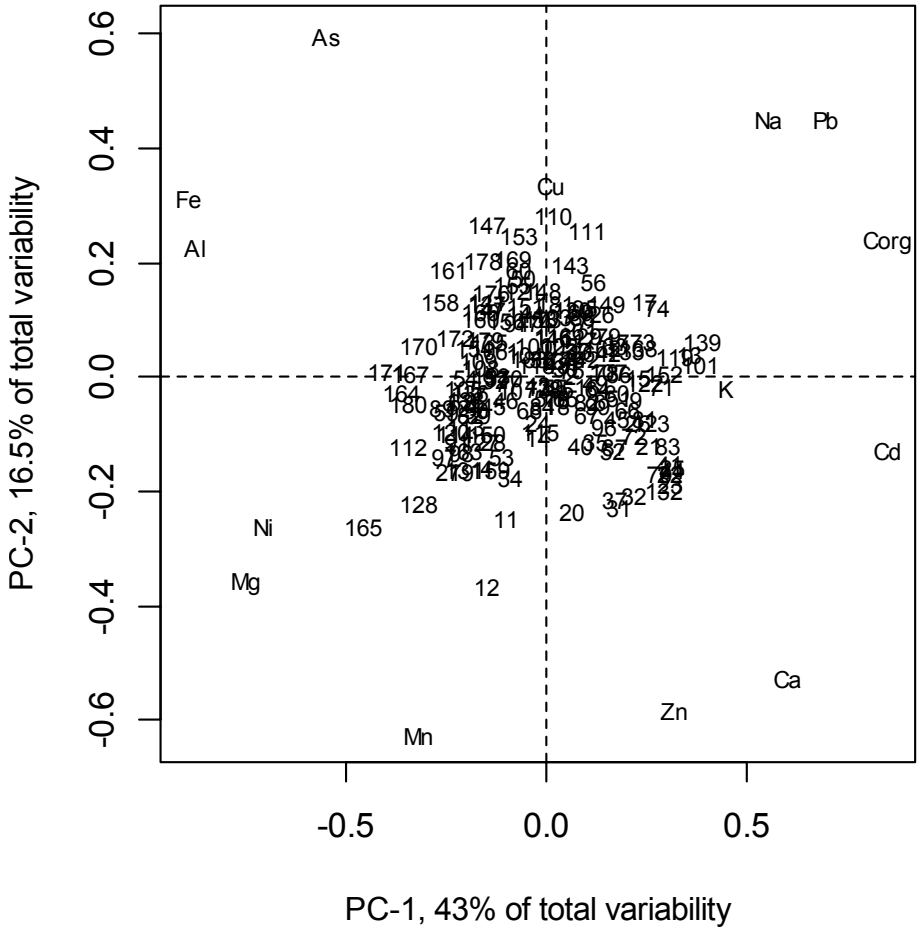

Fig. 12. Biplot display of the scores with function **gx.rqpca.plot** of the observations, identified by data row number, and the loadings in the PC-1 – PC-2 space.

following a logarithmic transformation. Function **gx.rqpca.screeplot** displays the traditional screeplot (Fig. S17), showing that the first four components contain 76% of the data that have been log-ratio transformed (also shown in Fig. 11). Function **gx.rqpca.plot** provides a biplot display of both the element loadings and the scores of the observations in the Principal Component space (Fig. 12), in this case with the observations identified by their data row numbers, which per-

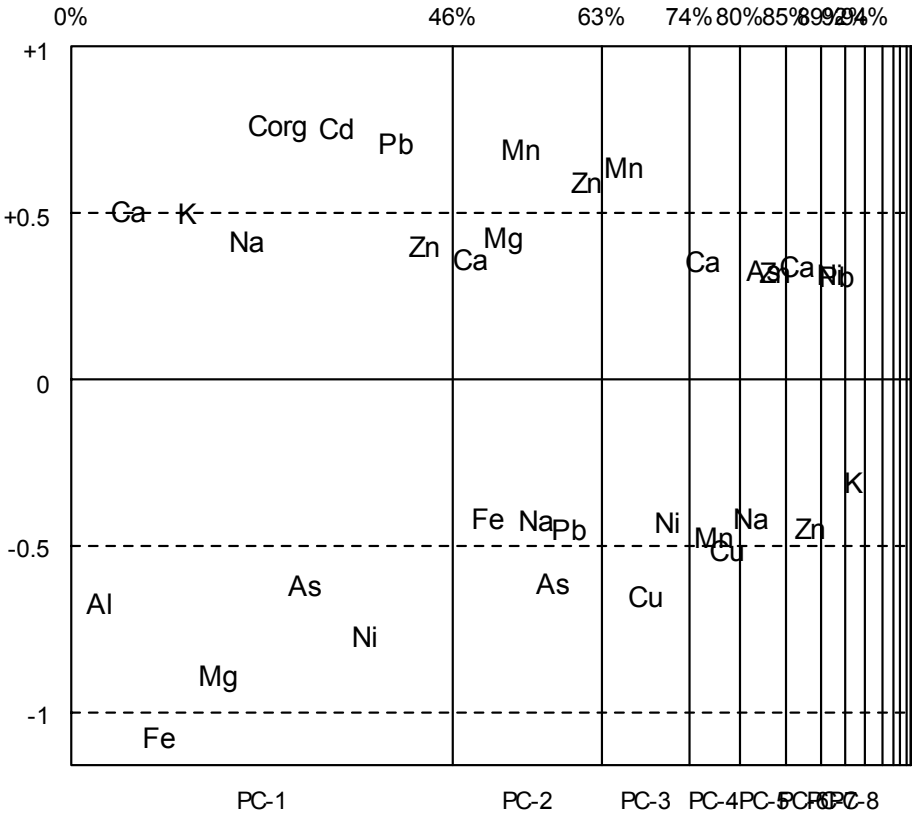

Fig. 13. Graphical display by function **gx.rqpca.loadplot** of robust PC loadings estimated with function **gx.robmva.closed** for the Maritime Provinces major and trace element suite employing an isometric log-ratio transformation.

mits the identification of any outliers. The inverted 'U' shape of the central cluster reflects a continuum of 0–5 cm soil samples from organic C-rich (positive PC-1) to mineral-rich (negative PC-1). Function **gx.rqpca.plot** permits the observations to be displayed in the context of any component pair that may be of interest in the search for outliers.

The search for outliers is more effective if the principal components are computed robustly with function **gx.robmva.closed** for compositional data. Robust estimation eliminates the effects of the outliers in the generation of the PC model, with the result that it better reflects the interrelationships in the majority background data (Filzmoser *et al.* 2009*b*). The Rousseeuw & Van Driessen (1999) minimum covariance determinant (mcd) is the default for robust covariance estimation; the minimum volume ellipsoid (mve) (Rousseeuw 1986) or user-provided weights may also be employed. Figure 13 displays the resulting loadings, where it can be seen that the first two include 63% of the core, background, variability. When the robust scores on the PCs are estimated it is in the context of the variability in the background model; therefore, any outliers become more distant and easier to recognize, Figure 14. The scores on these two components make up 91% of the total variability in the scores. This does not mean that there is no additional interesting variability to be sought, just that it has less range. Figure S18 is a plot of the scores on robust components 3 and 4, where outliers, candidates for further evaluation, are clearly evident.

Two additional functions are available for displaying and operating on PCA results. Function **gx.rqpca.print** permits loadings and scores matrices to be displayed and optionally saved as csv files; and function **gx.rotate** undertakes a Kaiser Varimax rotation (Kaiser 1958) on the PCs, after which any of the display functions may be used to display the results of rotation.

The three **mva** (multivariate analysis) functions described above for PCA also undertake the computations for displaying Chi-square plots of Mahalanobis distances (Gnanadesikan 1977) and store the results in the saved objects for display. These plots are the multivariate equivalent to the univariate cumulative probability plots used extensively by applied geochemists in data interpretation and outlier detection (Tennant & White 1959; Sinclair 1974, 1976; Stanley 1986).

The computations for Mahalanobis distances are only undertaken if the data are non-singular; if the data are singular an appropriate message is displayed and Mahalanobis distances are not estimated. Singularity arises with compositional data sets and when certain robust estimators are used. When compositional data are being investigated using robust estimation procedures, function **gx.robmva.closed** must be used. This procedure employs an isometric log-ratio transformation (ilr) (Egozcue *et al.* 2003; Filzmoser *et al.* 2009*b*), and involves additional internal computation hidden from the user. Chi-square plots are displayed with function **gx.md.plot** from the saved objects from the three **mva** functions.

Figure 15 displays the Chi-square plot for the same major and trace elements used to demonstrate the projection pursuit and PCA procedures following the application of an ilr transformation with function **gx.mva**. The use of the ilr transformation reduces the number of variables by one, explaining the y-axis title indicating 13 degrees of freedom. The data are clearly drawn from more than one multivariate normal distribution, as indicated by the flexure, and there are 8 outliers, possibly as many as 20. The identity of the potential outliers and their probabilities of membership in the data set can be displayed with function **gx.md.print**, and saved as a csv file for subsequent use if required.

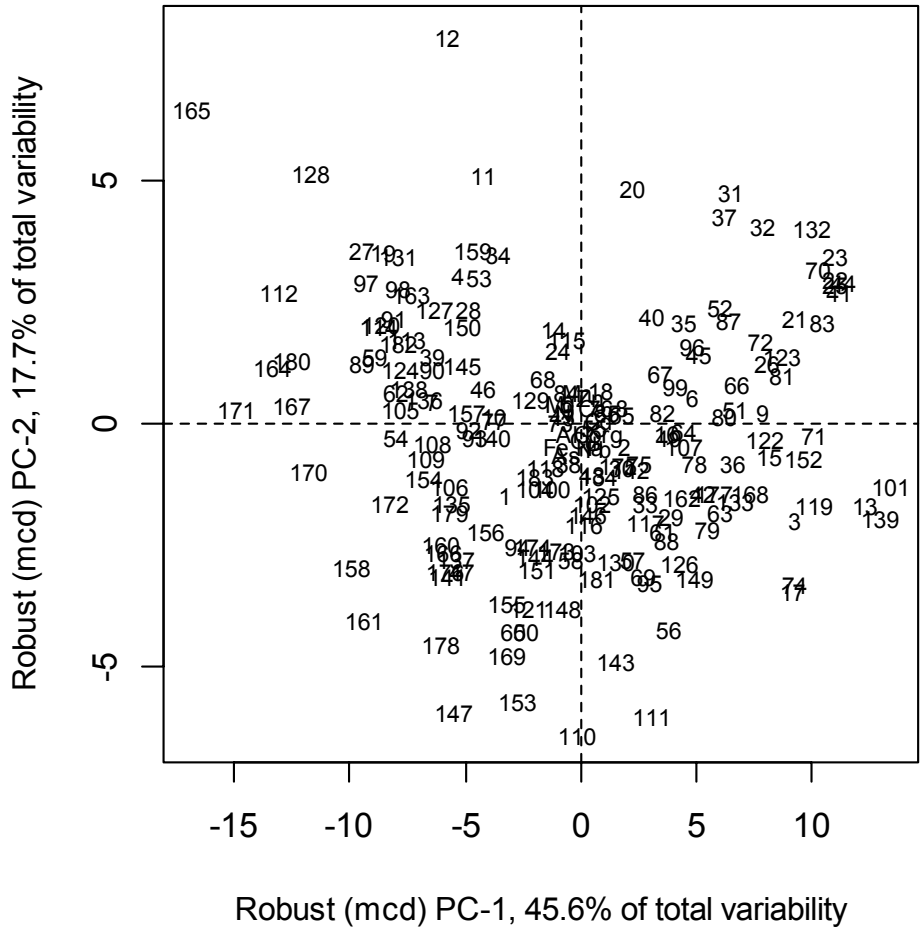

Fig. 14. Biplot display with function **gx.rqpca.plot** of the robust scores of the observations estimated with function **gx.robmva.closed**, identified by data row number, and the loadings in the PC-1 – PC-2 space (hidden in the central cluster).

When seeking to identify outliers it is recommended that robust estimators are used to better define the statistical parameters of the core background data. Figure 16 displays the results of applying function **gx.robmva.closed** to the data. By employing robust estimation with the default mcd procedure (Rousseeuw & Van Driesen 1999), two Chi-square plots are generated; the plot for the robustly estimated core data set of 98 individuals is displayed on the left, and the plot for all the data using the core data estimators is displayed on the right. An inspection of Figure 16 (right) indicates that there are ten clear outliers, and evidence of two populations in the data. The core data (Fig. 16 left), with a population size of 98, most likely represents a single multivariate normal population. Again, the identity of the potential outliers and their probabilities of membership in the data set can be displayed with function **gx.md.print**, and saved as a csv file for subsequent use if required.

A comparison of the outliers recognizable in the plots of robust PCs 1 versus 2 and 3 versus 4 (Figs 14 and S18) and the most extreme outliers identified by the robust Mahalanobis distance procedure (Fig. 16 right) was made. Of the 10 extreme robust Mahalanobis distance outliers, 8 occur as outliers in the PC plots; of the next 14 greatest Mahalanobis distances, 11 occur as outliers in the PC space.

A Graphical Adaptive Interactive Trimming (GAIT) procedure for sequentially trimming outliers from a Chi-square plot was proposed by Garrett (1989). The GAIT procedure is implemented through functions **gx.md.gait** for open, and **gx.md.gait.closed** for compositional, data sets, respectively. Again, the ilr transform (Egozcue *et al.* 2003; Filzmoser *et al.* 2009*b*) has to be used to permit estimating Mahalanobis distances robustly. The GAIT procedure progresses iteratively, with information (0 or 1 weights indicating trimmed outlier or member of the core, respectively) being passed to the subsequent step through the saved object from the prior step. As each step proceeds, two Chi-square plots are displayed by **gx.md.plot**, the plot for the trimmed data set (left side), and the plot of all the observations with respect to the remaining core subset at that point in the GAIT procedure (right side). The procedure continues until the Chi-square plot of the remaining core subset is reasonably linear, indicating the likelihood of a single multivariate normal distribution. The results in the final saved object from the GAIT procedure can be displayed and saved as required by functions **gx.md.plot** and **gx.md.print**. It is essential to start the GAIT procedure off robustly. Therefore, two options are available: firstly, to use the Minimum Covariance Procedure of Rousseeuw & Van Driessen (1999) as the default procedure in the **mva** functions; or secondly, to employ a multivariate trim (mvt) and remove a percentage of the most extreme Mahalanobis distances of the initial data set, often in the range of 10 to 25%. The actual percentage is determined by trial and error through repeated executions of the GAIT functions before actually starting to sequentially trim.

Figures 17 and 18 display the results of applying **gx.md.gait.closed** to the 14-element major and trace element data set. Figure 17 presents the Chi-square plots resulting from a 15% mvt of the extreme individuals identified by classical non-robust estimation. The left panel of Figure 17 indicates that a further three individuals could be trimmed; the result of

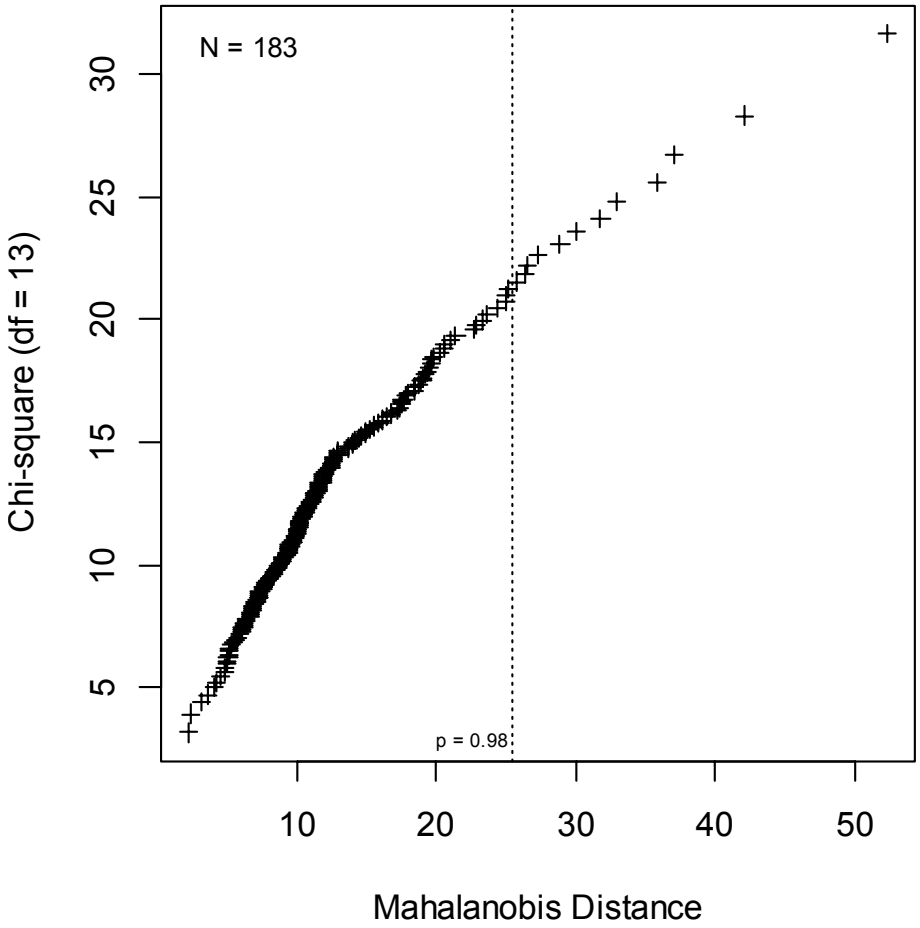

Fig. 15. Chi-square plot displayed with **gx.md.plot** from the output saved from **gx.mva** applied to the element major and trace element data for the 0–5 cm <2 mm soils, Maritime Provinces, following an isometric log-ratio transformation.

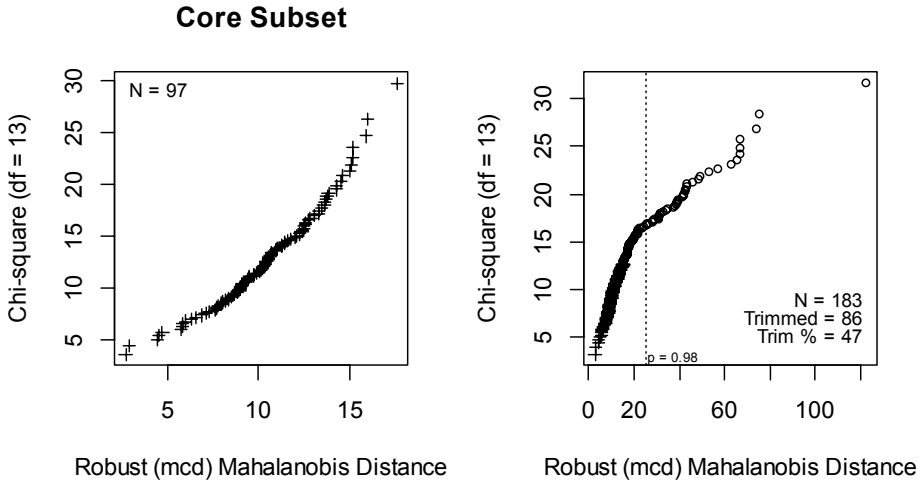

Fig. 16. Chi-square plots displayed with **gx.md.plot** from the output saved from **gx.robmva.closed** applied to the major and trace element data for the 0–5 cm <2 mm soils, Maritime Provinces.

doing so is presented in Figure 18. The result of this exercise is the removal of 30 multivariate outliers, and the creation of a core background data set of 153 individuals. The core data set is not a single population, as indicated by the flexure and lower density of points at the upper end of the Chi-square plot. However, it serves to separate the most extreme multivariate observations from the remaining data. If desired the GAIT procedure can be continued, or the previous step returned to and a greater number of observations trimmed. The group into which an individual observation falls is indicated by a weight (a list of the **wts** is in the saved object from the procedure). Weights of 0 and 1 indicate an outlier or a member of the core background data, respectively.

A comparison of the 30 outliers from the GAIT procedure with an mvt start and those from the mcd-based **gx.robmva.closed** procedure reveals gross differences. Of the top ten

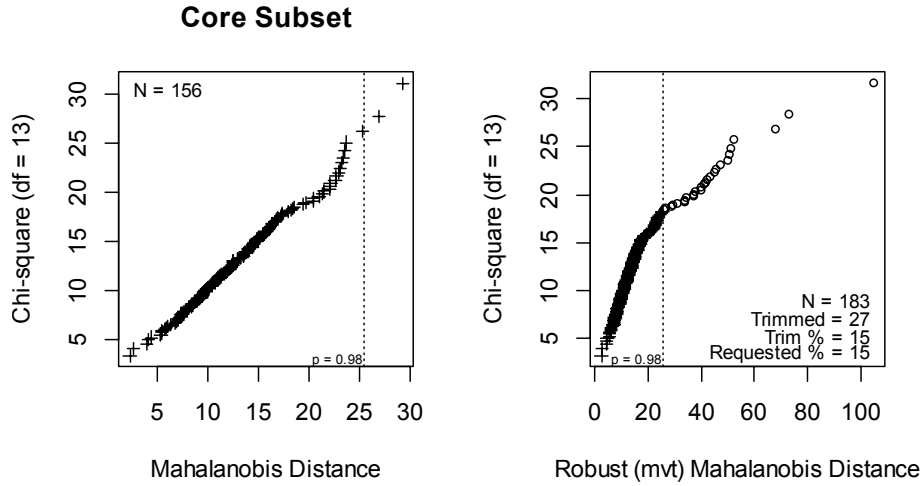

Fig. 17. Initial 15% Multivariate Trim (MVT) step for a GAIT investigation of the major and trace element data for 0–5 cm <2 mm soil, Maritime Provinces, following an isometric log-ratio transformation.

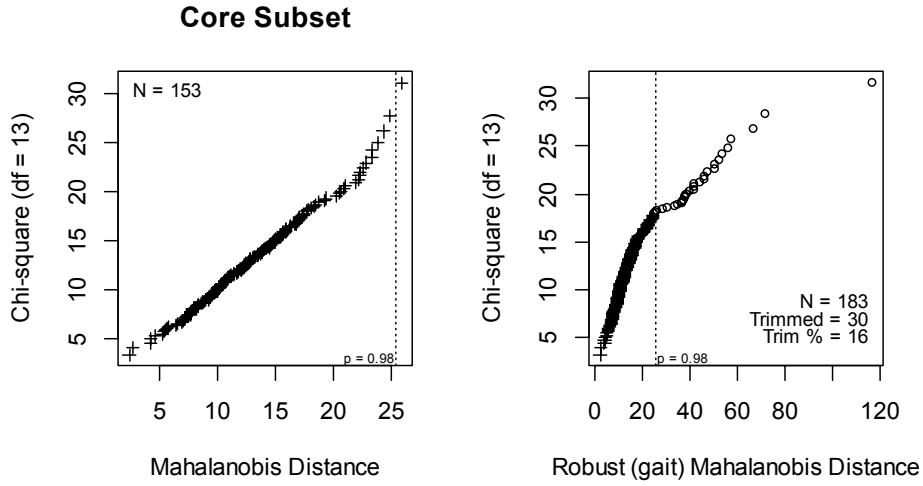

Fig. 18. Results of an additional GAIT step to remove a further three observations from the core data set.

Table 4. *Summary statistics for As (mg/kg), estimated with* **gx.summary1***, for the core background (N = 98) data subset identified by the* **gx.robmva.closed** *procedure*

> gx.summary1(As[PH.Majors.TEs.robmva.closed$wts==1])

Summary Stats for: As[PH.Majors.TEs.robmva.closed$wts == 1]

N	NAs	Min	Q1	M	Q3	Max	MAD	IQR_SD	Mean	SD	CV%	SE & 95%	CLs on Mean
98	0	0.6	2.2	3.4	5.2	15.7	2.15	2.224	4.653	3.713	79.8%	0.3751	3.909<->5.398

mcd identified outliers (Fig. 16 right) only three are identified by the GAIT procedure. From this it is deduced that it would have been important to continue the GAIT procedure to the point where the core data subset more clearly represented a single background population.

The function **gx.summary1** may then be used to generate summary statistics for As in the background data subset identified by **gx.robmva.closed** (see Table 4). Alternately, the weights can be bound to an existing data frame or matrix using the **cbind** R 'primitive' and the new object saved for further processing. It can immediately be seen in Figure 19 that the core background data still contain at least 12 As outliers. This particular example demonstrates an instance of a simple univariate procedure providing more insight than a more complex multivariate procedure. Figures 16 and 18 indicate that the data are polypopulational, as do the cumulative probability plots for As when they are subdivided by Ecoprovince (see Fig. 3). One of the gross As outliers discernable in Figure 3 is swamped by the data for the additional 13 major and trace elements and is not extreme in the multivariate context. The availability of prior ancillary information, such as Ecoprovinces, permits cumulative probability plots to be prepared supporting the hypothesis that there are three discrete data sets, and that upper limits of background variation (thresholds) can best be selected directly from those plots. Where no ancillary data are available multivariate data analysis tools may provide useful support to data interpretation. However, the simple approaches should always be investigated first, and they may prove to be the most effective.

Where a multivariate data set can be subdivided into a number of background populations and where core groups

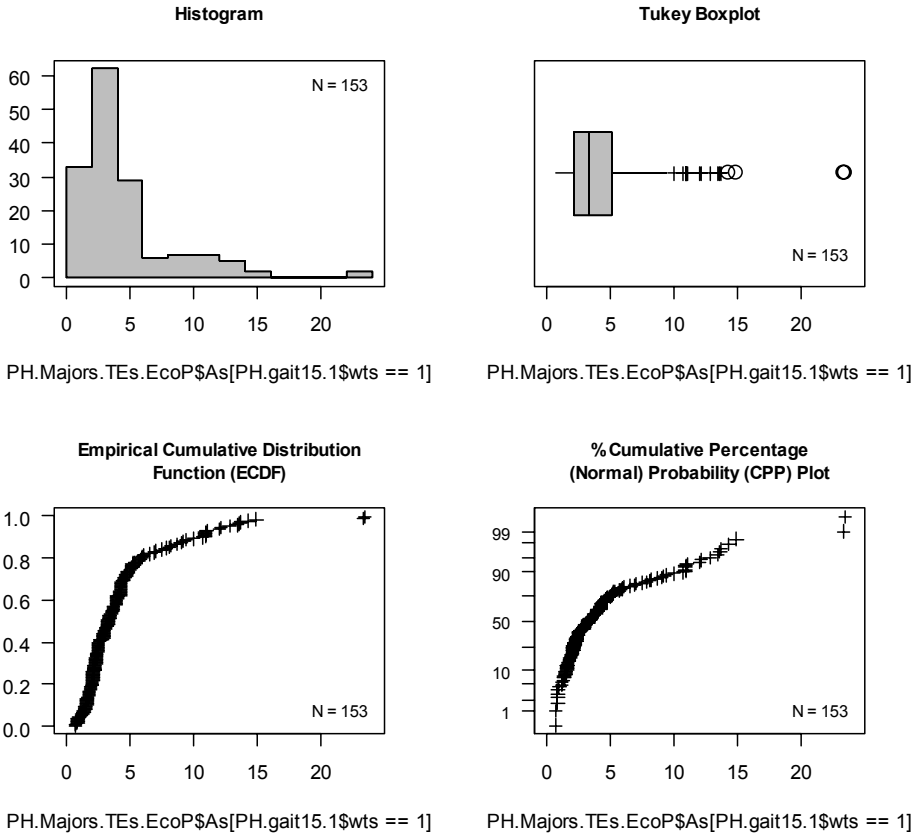

Fig. 19. Function **shape** used to display the As (mg/kg) data for the 0–5 cm <2 mm soil subset identified as background by the **gx.robmva.closed** procedure. Note the use of a data conditioning procedure available in R to rapidly produce a plot for inspection. Only those data for As are plotted where the weights (**wts)** from the **gx.robmva.closed** procedure are equal to 1.

of observations can be identified that characterize those populations, all of the observations in the original data set can be allocated to one of the background populations, or identified as a multivariate outlier with respect to the background populations (Garrett 1990). Functions **gx.mvalloc** and **gx.mvalloc.closed** perform the allocation procedure for open and compositional data, respectively. The functions are passed the total data set, the saved objects from the last GAIT or robust Mahalanobis distance investigations for each background population, and a cut-off value for the probability of group membership. Any observation (sample) with a probability of group membership in all background groups of less than the cut-off is allocated to an 'unknown' group. It is these individuals that are of the greatest interest as they identify those observations that have been most affected by non-background processes, whether they are unrecognized background populations, reflect the local presence of unusual primary or secondary processes, or are related to the presence of ore minerals. The allocation results and probabilities of group memberships may be displayed with function **gx.mvalloc.print** and saved as a csv file for any subsequent processing. Examples of the allocation procedure using synthetic and lithogeochemical data are provided in the 'rgr' manual.

The projection pursuit, **mva**, GAIT and allocation functions all store their results in saved objects. This makes 2-d coordinates, PCA scores, Mahalanobis distances and associated probabilities immediately accessible for display within 'rgr'. Using R 'primitives', the desired results can be extracted and appended to their parent data sets, thus enabling various plots and maps to be prepared using appropriate R and 'rgr' display functions.

QA/QC support

QA/QC support may be approached three ways:

(1) Inspecting analytical data for long-term continuity using control reference materials and estimating analytical precision and accuracy;
(2) Inspecting data for analytical duplicates to ensure satisfactory agreement, and displaying the duplicates as Thompson–Howarth plots; and
(3) Determining if the data are 'fit for purpose' using Analysis of Variance (ANOVA) procedures to estimate if sampling and/or analytical variability is sufficiently small relative to the variability between the observations in the study.

Control reference data should be plotted using the **crm.plot** function, with the data in sequence of analysis, in order to detect any changes in laboratory procedure with time. Where a mean and SD for the reference material are available, the mean and tolerance bounds, expressed in probability terms, may be plotted as horizontal lines (Fig. 20a). If the tolerance bounds are expressed as a percentage of the reference value, horizontal lines can similarly be added to the plot (Fig. 20b). Should any analyses fall outside the chosen tolerance limits, especially if several analyses occur in a continuous run, some follow-up action is required to determine the seriousness of the issue and what action should be taken. In the above example the new data all fall within acceptable limits; however, there is a slight upward shift averaging about 2 mg/kg. Some geochemists may choose to subtract 2 mg/kg from the project Cu values before proceeding further in data analysis or compilation. The function also displays the mean and relative SD (CV%) for the control reference material analyses to estimate precision.

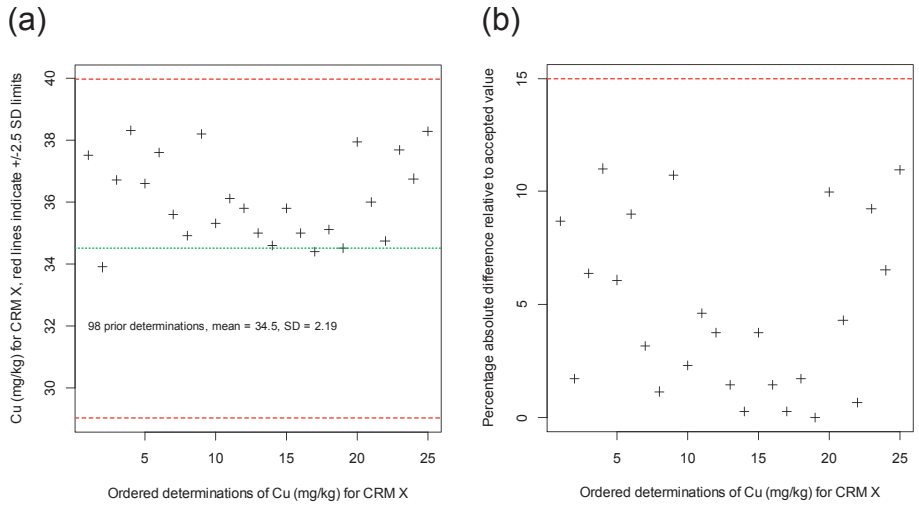

Fig. 20. QA/QC plots for control reference material (GSC NGR CRM-X) analyses for Cu (mg/kg) acquired during the 2000 and 2001 NGR stream sediment surveys in New Brunswick. The X-axis uses a R construct to generate an index number for the order in which the CRM-X data appear in the data frame. Plots generated by function **crm.plot**.

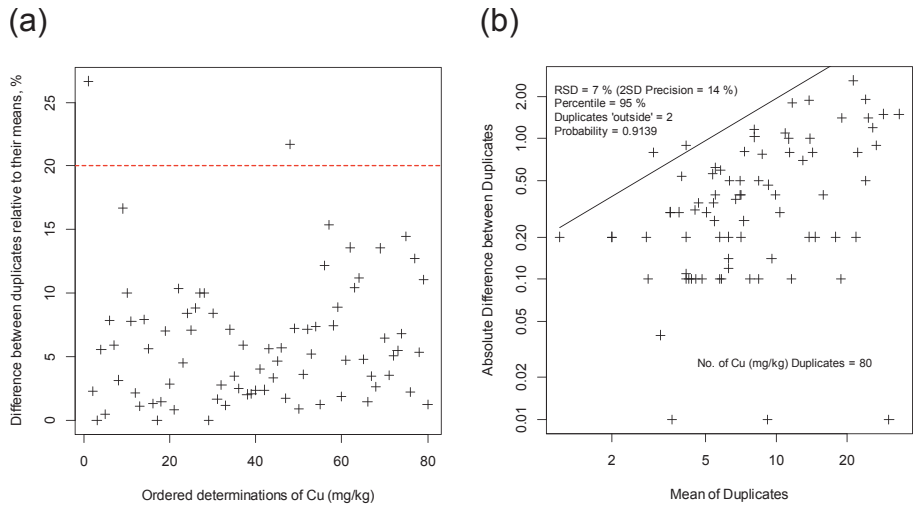

Fig. 21. QA/QC plots for blind (analytical) duplicate analyses for Cu (mg/kg) acquired during the 2000 and 2001 NGR stream sediment surveys in New Brunswick. Left plot generated by function **ad.plot2**, and right plot (Thompson–Howarth) by function **thplot2**. The X-axis uses a R construct to generate an index number for the order in which the analytical duplicates (RS = 8) appear in the data.

Duplicate analyses can be inspected in a generally similar way using the **ad.plot1** or **ad.plot2** functions, by plotting the differences between the duplicates against their order of analysis. In this instance the difference between the two analyses is expressed as a percentage of their mean, and tolerance levels, expressed as percentage differences, are plotted as horizontal lines on the plot for the ranges they apply to (Fig. 21a). Alternately, the duplicates may be plotted as Thompson–Howarth plots (Thompson & Howarth 1973, 1978; Garrett & Grunsky 2003) with functions **thplot1** or **thplot2**, depending on whether the data are tabulated in columns or rows, respectively, with a line plotted to indicate the limit below which 95% of the duplicates plotted should fall for a given precision (Fig. 21b). For this example a precision (RSD) of ±7% was selected, and the Thompson–Howarth procedure indicated that there was a better than 91% probability that the ±7% target was met. As with control reference samples, any runs or grouping of duplicates that fall out of tolerance should be investigated to determine what remedial action should be taken.

Rather than comparison against external benchmarks, i.e. accuracy and precision, the real test is for whether the data are 'fit for purpose'. This requires that the measurement errors due to local sampling and analytical variability are significantly smaller than the data variability between the observations (i.e. geochemical samples) across the survey area. Where analytical or sampling duplicates are available, the latter estimating the combined sampling and analytical variability, functions **anova1** or **anova2** may be used to carry out the F-test for fitness of purpose, and estimate the analytical precision in the case of analytical duplicates. Additionally, the ANOVA calculations permit the percentage of the total variability due to 'between observation' and 'within observation' components to be estimated (Table 5). The ANOVA for analytical duplicate determinations of Cu indicates that in excess of 99% of the variability in the duplicate analyses is between the duplicate pairs and less than 1% is due to analysis – a very favourable outcome. Additionally the precision (RSD) of the analysis is estimated as about 5%. In the case of field sampling duplicates this is particularly useful and informative for helping decide

Table 5. *One-way ANOVA for analytical duplicate determinations of Cu (mg/kg) from the 2000 and 2001 NGR stream sediment surveys in New Brunswick*

> anova2(Cu,ifalt=T)

Combined Sampling and Analytical, or Analytical Variability, Study, Utilizes Field Sampling or Laboratory Duplicates. In ANOVA Tables, the variability:

Between would be between sampling sites or analysed samples, and

Within would be at sampling sites or due to duplicate analyses

Two-Way Random Effects Model for Cu					
Source	**SS**	**df**	**MS**	**F**	**Prob**
Between	**8919.8**	**79**	**112.91**	**223.92**	**0.0088**
Within	**20.736**	**1**	**20.736**	**1510.38**	**0**
Residual	**1.0846**	**79**	**0.013729**		
Total	**8941.6**	**159**	**56.236**		
One-Way Random Effects Model for Cu					
Source	**SS**	**df**	**MS**	**F**	**Prob**
Between	**8919.8**	**79**	**112.91**	**413.94**	**0**
Within	**21.821**	**80**	**0.27276**		
Total	**8941.6**	**159**	**56.236**		
Source	**MS**	**Var Comp**	**%age**		
Between	**112.91**	**56.318**	**99.5**		
Within	**0.27276**	**0.27276**	**0.5**		
		56.591			
Summary Statistics for Cu					
Grand Mean = 10.027			**Variance = 56.236**		
'Error' S^2 = 0.27276			**Std. Dev. = 0.52227**		
'Error' RSD% = 5.2					
Miesch's V = 206.47			**Vm = 412.94**		

which variables merit plotting as maps or along traverses. However, if duplicates are generated appropriately a more informative procedure is available.

Preferably the analytical duplicate should be split from one of the field duplicates and the best practice is to use function **gx.triples.aov** (see Garrett (in press) for details). In such cases F-tests are undertaken for the local sampling variability relative to the analytical variability, and the between-observation, broader-scale, variability and the local sampling variability. The results can be useful for indicating where modifications to the field sampling or analytical protocols would improve data quality. For example, a high proportion of the variability at the analytical level would indicate that benefits would likely accrue from a more precise analytical procedure, perhaps one with a lower detection limit or use of larger sample aliquots. Where a high proportion of the variability is at the local sampling level, composite samples, which would have to be collected at all sites, could prove beneficial to survey quality.

The above ANOVA procedures are sensitive to heteroscedasticity, i.e. lack of homogeneity of variance. Homogeneity of variance is an underlying assumption to least-squares procedures, of which ANOVA is one, and requires that the variance of subsets of the data taken across the range of the data are relatively constant. With geochemical data ranging across more than 1.5 orders of magnitude this is unlikely. A standard procedure in such instances is to log-transform the data (Weisberg 1985). There is also the matter of the compositional nature of geochemical analyses; as has been noted previously, at concentrations less than 10% the scaling changes accomplished with logarithmic and logit transformations are linear, and therefore the partition of the variance between the various sources will be similar whichever transformation is used. Table 6 presents the results of the variance component estimation and F-tests for the Cu analytical duplicate data from Table 5. As can be seen, the logarithm and logit transformed results are identical,

Table 6. *Comparison of ANOVA results for duplicate determinations of Cu following different data transformations to accommodate the compositional nature of the data*

	Variance Components as %		
Transformation	Between Analyses	Within Analyses	F-test
None	99.5	0.5	413.9
Log10	99.4	0.6	331.5
Logit	99.4	0.6	331.5

as might be expected at tens of mg/kg levels, and no transformation underestimates very slightly the within-analyses variance. A similar study for combined field and analytical triplicates is presented in Garrett (in press).

Data conditioning

In some respects it could be argued that this section should be placed earlier as it contains some of the less glamorous but essential functions. These include tools for dealing with 'less than detection level' values, situations where there are no data, and transformations.

By convention, package 'rgr' treats negative numbers as indications of measurement levels below the method detection limit, thus -2 indicates <2. This is acceptable for geochemical analyses, but may limit the use of 'rgr' by physical scientists who routinely encounter negative values. In such cases, potential 'rgr' users would have to pre-process their data sets externally to 'rgr' prior to importing them. Two functions handle non-detects recorded as negative numbers, **ltdl.fix** and **ltdl.fix.df**, the former for use with vectors and the latter with data frames. Both these functions contain 'switches' permitting the replacement of zeros or coded values such as -9999 with NAs (the standard S and R indication of missing data, indicating Not Available). These features are present because some software packages replace blanks with zeros when exporting data files, and some data providers explicitly code the absence of data with a character string like -9999. A task for the user is to determine the meaning of any zeros in data prior to its importation into R. Some zeros for non-compositional data may be legitimate; a zero for applied geochemical compositional variable should be treated as a non-detect and be replaced appropriately. It has proven good practice to process a newly imported data frame with function **ltdl.fix.df**, after which one does not have to worry about non-detects and zeros. Where data contain a high proportion of non-detects, say greater than 10 to 15%, the procedures in 'rgr' may not be appropriate. For non-parametric statistics the limit is far higher and, for example, percentiles displayed as -2 can be replaced by <2 in a publication table. Fortunately, advances in analytical chemistry over the last two decades have led to data sets where non-detects are increasingly infrequent. For parametric statistical methods, i.e. those based on means, variances and covariances, excessive non-detects will lead to inappropriate estimates that can be misleading and worthless. In such cases interested workers are referred to the work of Helsel (2012) for appropriate procedures, and those with compositional data should also consult Palarea-Albaladejo *et al.* (2007).

To avoid the potential problems caused by the presence of NAs in a data set, two R functions, **na.rm** and **na.omit**, are available to process NAs in vectors and data frames, respectively. The 'rgr' functions make explicit calls to 'rgr' function **remove.na**. This function removes any NAs from a vector, or rows containing NAs, from a matrix, and displays the number of NAs or matrix rows removed, and the size of the resulting

vector or matrix. Where NAs represent missing data, say for insufficient sample material for an analytical procedure, it may be possible to impute a value from the multivariate information in the data set. Those requiring such imputations should investigate the robCompositions package (Templ *et al.* 2011*a*, *b*).

It is often informative to work on data subsets defined by a single criterion or several criteria in combination. Function **gx.subset** undertakes such tasks on data frames with the saved object being a new data frame which meets the criterion. The criterion can be any combination of factor (character string), and numeric variables. For example, all the stream sediment samples classified as draining a particular rock type, **RCK3**, and having **Cu >value1** and **Ni <value2**, can readily be placed into a new data frame subset containing only the samples with that unique combination of properties for subsequent inspection or processing.

Three functions are available in 'rgr' to undertake data transformations related to matrices of compositional data. Functions **alr**, **clr** and **ilr** undertake additive, centred and isometric log-ratio transformations, respectively (Aitchison 1984, 1986; Egozcue *et al.* 2003; Filzmoser *et al.* 2009*b*). Functions **clr** and **ilr** are used internally by some 'rgr' functions, but all are available for use in investigations into closure and compositional data sets. A fourth function, **rng**, undertakes a range transformation on the columns of a matrix such that the values range from 0 to 1, corresponding to the lowest and highest values, with the intermediate values dispersed linearly between the extremes. Simple logarithmic, logit and square-root transformations may be achieved with the R functions **log**, **log10, logit** (in 'rgr') and **sqrt**. Other transformations are available in other packages, such as Box-Cox in package MASS required by 'rgr', or may be scripted by the user.

Utility functions

Sometimes it is required to know if a particular data frame is available in a R session. Function **df.test** determines if a data frame is available in the current R session and if it is, the names of the variables are displayed. If **df.test** is queried with both the data frame name and a valid variable name, the number of observations for that variable is displayed.

On occasion one or a few NAs are buried in a long vector or large matrix and it is required to know just where they occur. Function **where.na** identifies the position(s) in a vector, or the row-column locations in a matrix or data frame, containing any NAs. It may also be used in an alternative procedure for removing the NAs from a data set.

Two functions are available for sorting data in a matrix or data frame. Function **gx.sort** operates on both matrices and data frames. By providing the name of the data object, together with the column to sort on, a sorted matrix, optionally in ascending (default) or descending order, is displayed or saved in a new object. Function **gx.sort.df** only operates on data frames, and is capable of complex multi-column sorts into ascending or descending orders, as the user requires. Again, the result of the sort may be displayed or saved as a new object.

Two functions have been implemented to investigate patterns in geochemical data along traverses or transects. The more incisive of these, function **gx.hypergeom**, employs the hypergeometric distribution as described by Stanley (2003). It is effective in determining if a pattern of above- threshold values conforms with a known or hypothesized geochemical model, or could just as easily have occurred by chance. A less powerful technique is the Wald-Wolfowitz Runs test, function **gx.runs**, which is useful for testing pattern coherence with a known or hypothesized geological or geochemical model.

Two tools to assist in regression analysis studies are functions **gx.adjr2** and **gx.lm.vif**. Function **gx.adjr2** calculates adjusted R^2 values for multiple regression models taking account of the number of observations and independent (predictor) variables. Computation is based on data retrieved from the saved object of an R regression exercise. This estimator has been in use for many years and is included so that comparisons may be made with earlier studies. However, any current regression studies should employ Akaike's Information Criteria (AIC) which can be computed from a saved R regression model with R function **AIC** to guide optimal model selection. Function **gx.lm.vif** computes the Variance Inflation Factor (VIF), a measure of collinearity in a regression model. Collinearity is an undesirable property in regression models, and models can usually be improved by removing variables with high VIFs.

Considerable flexibility is available to users in preparing graphics for publications. Three functions display options that are available for line styles and colours (**display.lty)**, plotting symbols (**display.marks)**, and octal codes for including special symbols, such as the Greek μ in plot text (**display.ascii.o)**. These may be printed out and provide useful aide memoires during graphics preparation. A fourth function, **display.rainbow**, displays the various options available for the alternate palettes for use in function **caplot**.

The descriptions and examples above are brief and many options are not fully described. The complete details with examples are in the 209-page 'rgr_1.1.9' Help Manual that is included in the 'rgr' package and is available as hypertext during a 'rgr' session.

DISCUSSION

The specific criteria for GIGS (Crain 1974) were for the system to have: (1) simplicity of operation; (2) interactions; (3) convertability; and (4) low-cost operation. The same criteria lay in the specifications for IDEAS, with the additions of: (1) friendliness; (2) the use of the ISO Standard GKS (Graphics Kernel System) and Fortran 77 standards to ensure interoperability; and (3) the generation of publication-quality graphics and tables (Garrett 1988). The 'rgr' package generally stacks up well against these objectives, better in some areas and less well in others.

The use of R has achieved all the objectives as far as interoperability is concerned: R and 'rgr' run on Windows, Mac and Unix platforms. Low-cost operation has been achieved: R and 'rgr' run on today's entry level desktop and laptop machines, and the software comes at no cost. The software will even run on notebooks and tablets, but the small screens on some of these devices are not suitable for line graphics. The preparation of publication-quality graphics, with the use of colour if desired, is routine, with R saving graphics files in all common formats. Tables can be exported, or 'cut-and-pasted', out of an 'rgr' session and edited for direct use in publications or presentations, etc.

It can be argued that R and 'rgr' have not met all the requirements for simplicity in that neither has a standard Graphics User Interface (GUI) as is available in many commercial software packages. However, the lack of a GUI and the use of a 'command-line' make it easier to exercise the many options available in both R and 'rgr', especially those for preparing graphical output. In some respects, the availability of the hypertext help files (those for 'rgr' include extensive examples), compensates for the lack of a GUI. R and 'rgr' lack some of the interactivity of IDEAS where it was simple to draw a polygon around a group of points on a display and graphically

create a data subset. It is acknowledged that using R and 'rgr' requires a commitment from the user to 'learn a language'. This may be considered a drawback, and a hurdle over which some users may not wish to jump. However, once the 'jargon' is learnt it opens the way to publication-quality graphics and interesting investigations of data sets in a zero-cost Open Source environment.

When development of IDEAS ceased, a number of multivariate analysis procedures had not been implemented, for example, PCA. The decision to base future development on and in R immediately made available all the work and documentation that was in base R, in packages in CRAN, and in shared and published R and S scripts that carried out computations appropriate to 'rgr'. The use of R as a platform facilitated and reduced the time for the development of 'rgr' with proven and reliable software.

Lastly, there is debate concerning the importance of addressing the compositional nature of the chemical data of applied geochemistry, and allowing for closure. Some would argue that failure to completely address closure invalidates the results of any data analysis. It is hard to support this view when many of the procedures in 'rgr' have been used successfully over the years in mineral exploration and other applied geochemistry studies without allowing for closure. The important question that cannot be answered is: "How many mineral exploration failures have there been because closure was ignored?".

The projection pursuit examples in Figures 9, 10, S13 and S14, illustrate the discussion well; from an EDA perspective it is the instance of ignoring the compositional nature of the data that provides the greatest EDA insight (Fig. 10). Templ *et al.* (2008) report similar findings, and state "Informative results are not necessarily obtained by tuning the parameters for cluster analysis in a statistically optimal way".

The EDA procedures in 'rgr' can identify 'unusual' observations; it is then the task of the user to explain and interpret them. When closure has not been considered and the EDA procedures have led to mineral discoveries, it would seem that closure is not a major concern. However, when the task is to understand, hypothesize or identify geochemical processes, it is important to consider the compositional nature of geochemical data. In this respect, one has to balance the realities of mineral stoichiometry with the mathematical implications of compositional data; this provides a continuing challenge. Package 'rgr' provides tools for both approaches to data, and users are encouraged to try different methodologies and develop the experience to guide data analysis for a variety of applied geochemistry applications.

CONCLUSIONS

Package 'rgr' has met many of the graphical and statistical requirements of applied geochemists at the GSC. By basing 'rgr' on the R-Project users have also been introduced to R and are able to take advantage of all the benefits that R and CRAN offer. Some with specific needs are using R in routine and innovative ways to support their projects.

As noted previously, there are many statistical procedures not included in 'rgr' because they are already present in R and other CRAN packages. For example, extensive regression procedures for classical ordinary least-squares, robust and non-linear modelling and various hypothesis testing tools are all available in R. Where 'rgr' has contributed is in generating graphics, tables, and undertaking procedures typically used in applied geochemical data interpretation. Undertaking the work to extensively document 'rgr' so that it is acceptable for CRAN has meant that it is available to the applied geochemical community on a wide variety of computing platforms. The fact that it is all Open Source means that others can build on the software as 'rgr' has been built on previous work of others.

The author expresses his thanks to Graeme Bonham-Carter, Eric Grunsky and Wendy Spirito for their comments and suggestions for improvements to an original typescript, and to two anonymous reviewers for their constructive comments on a draft of this paper. It would be remiss not to mention the collaborations with Eric Grunsky and Graeme Bonham-Carter, both R and S-Plus users, at the GSC over the years, and with Dave Lorenz of the U.S. Geological Survey, Minneapolis, who helped solve a particular 'knotty' problem in 2005. This paper is published as Natural Resources Canada contribution ESS 20110271.

REFERENCES

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

Aitchison, J. 1986. *The Statistical Analysis of Compositional data. Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Becker, R.A. & Chambers, J.M. 1984. *S: An Interactive Environment for Data Analysis and Graphics*. Wadsworth & Brooks/Cole, Pacific Grove, California.

Bivand, R.S., Pebesma, E.J. & Gomez-Rubio, V. 2008. *Applied Spatial Data Analysis with R*. Springer, New York.

Buccianti, A., Mateu-Figueras, G. & Pawlowsky-Glahn, V. (eds) 2006. *Compositional data analysis in the geosciences: From theory to practice*. Geological Society, London, Special Publications, **264**.

Chambers, J.M. 1977. *Computational methods for data analysis*. John Wiley & Sons, Inc., New York.

Cheng, Q. & Agterberg, F.P. 1995. Multifractal modeling and spatial point processes. *Mathematical Geology*, **27**, 831–845.

Cheng, Q., Agterberg, F.P. & Ballantyne, S.B. 1994. The separation of geochemical anomalies from background by fractal methods. *Journal of Geochemical Exploration*, **51**, 109–130.

Cox, T.F. & Cox, M.A.A. 2001. *Multidimensional Scaling*. Chapman & Hall, London.

Crain, I.K. 1974. The Geochemical Interactive Graphics System. *In*: Gordon, T.M. & Hutchison, W.W. (eds) *Computer use in projects of the Geological Survey of Canada*. Geological Survey of Canada. Paper **74-60**, 59–61.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateau-Figueras, G. & Barcelo-Vidal, C. 2003. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, **35**, 279–300.

Filzmoser, P., Hron, K. & Reimann, C. 2009*a*. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. *Science of the Total Environment*, **407**, 1610–6108.

Filzmoser, P., Hron, K., Reimann, C. & Garrett, R. 2009*b*. Robust factor analysis for compositional data. *Computers & Geosciences*, **35**, 1854–1861.

Filzmoser, P., Hron, K. & Reimann, C. 2010. The bivariate statistical analysis of environmental (compositional) data. *Science of the Total Environment*, **409**, 4230–4238.

Filzmoser, P. & Steiger, B. 2012. *Statistical analysis for environmental data*. http://cran.r-project.org/web/packages/StatDA/index.html (Last accessed 27 August, 2013.)

Friedman, J.H. & Rafsky, L.C. 1981. Graphics for the multivariate sample problem. *Journal of the American Statistical Association*, **76**, 277–291.

Friske, P.W.B., Pronk, A.G., McCurdy, M.W., Day, S.J.A., McNeil, R.J. & Boldon, R. 2002. *Regional stream sediment and water geochemical data, Central New Brunswick (NTS 21J/06 and 21J/10)*. Geological Survey of Canada, Open File **4112**, & NB DNRE, Open File **2002–1**.

Friske, P.W.B., Pronk, A.G., Day, S.J.A., McCurdy, M.W., McNeil, R.J. & Boldon, R. 2003. *National Geochemical Reconnaissance: Regional stream sediment and water geochemical data, Central New Brunswick. NTS 21J/11 East and 21J/15 West*. Geological Survey of Canada, Open File **4410** & NB DNRE, Open File **2002–6**.

Friske, P.W.B., Ford, K.L. & McNeil, R.J. 2011. *Soil geochemical, mineralogical, radon and radiometric data from the 2007 North American Soil Geochemical Landscapes Project in New Brunswick, Nova Scotia and Prince Edward Island*. Geological Survey of Canada, Open File **6433**.

Garrett, R.G. 1973. Regional geochemical study of Cretaceous acidic rocks in the northern Canadian Cordillera as a tool for broad mineral exploration. *In*: Jones, M.J. (ed) *Geochemical Exploration 1972*. Institution of Mining and Metallurgy, London, 203–219

Garrett, R.G. 1974. Computers in exploration geochemistry, a review of EDP usage in the Geochemistry Section of the Resource Geophysics and Geochemistry Division. *In*: Gordon, T.M. & Hutchison, W.W. (eds)

Computer use in projects of the Geological Survey of Canada. Geological Survey of Canada, Paper **74–60**, 63–66.

GARRETT, R.G. 1983. Opportunities for the 80s. *Mathematical Geology*, **15**, 389–402.

GARRETT, R.G. 1988. *IDEAS - An interactive computer graphics tool to assist the exploration geochemist*. Geological Survey of Canada, Paper, **88-1F**, 1–13.

GARRETT, R.G. 1989. The Chi-square plot - A tool for multivariate outlier recognition. *In*: JENNESS, S. (ed.) Geochemical Exploration 1987: Proceedings of the 12th International Geochemical Exploration Symposium, 23–26 April, 1987, Orleans, France. *Journal of Geochemical Exploration*, **32**, 319–341.

GARRETT, R.G. 1990. A robust multivariate allocation procedure with applications to geochemical data. *In*: AGTERBERG, F.P. & BONHAM-CARTER, G.B. (eds) *Proceedings of colloquium on statistical applications in the earth sciences*. Geological Survey of Canada, Paper **89–9**, 309–318.

GARRETT, R.G. 1993. Another Cry from the Heart. *Explore, Newsletter for the Association of Exploration Geochemists*, **81**, 9–14.

GARRETT, R.G. 2013. *rgr: The GSC Applied Geochemistry EDA Package*. http://cran.r-project.org/web/packages/rgr/index.html (Last accessed 27 August, 2013.)

GARRETT, R.G. In press. Assessment of local spatial and analytical variability in regional geochemical surveys with a simple sampling scheme. *Geochemistry: Exploration, Environment, Analysis*, http://dx.doi.org/10.1144/geochem2011-085

GARRETT, R.G. & CHEN, Y. 2007. *rgr: The GSC (Geological Survey of Canada) applied geochemistry EDA package - R tools for determining background ranges and thresholds*. Geological Survey of Canada, Open File **5583**, 1 CD-ROM.

GARRETT, R.G. & GRUNSKY, E.G. 2001. Weighted Sums - Knowledge based empirical indices for use in exploration geochemistry. *Geochemistry: Exploration, Environment, Analysis*, **1**, 135–141.

GARRETT, R.G. & GRUNSKY, E.C. 2003. S and R functions for the display of Thompson–Howarth plots. *Computers & Geosciences*, **29**, 239–242.

GARRETT, R.G., KANE, V.E. & ZEIGLER, R.K. 1980. The management and analysis of regional geochemical data. *Journal of Geochemical Exploration*, **13**, 115–152.

GNANADESIKAN, R. 1977. *Methods for Statistical Data Analysis of Multivariate Observations*. John Wiley & Sons, Inc., New York.

GRASS. 2011. Geographic Resources Analysis Spatial System. http://grass.fbk.eu (Last accessed 30 January, 2012.)

GRUNSKY, E.C. 2001. A program for computing RQ-mode principal components analysis for S-PLUS and R. *Computers & Geosciences*, **27**, 229–235.

GRUNSKY, E.C. & BACON-SHONE, J. 2011. The stoichiometry of mineral compositions. *Proceedings of CODAWORK 2011: 4th International Workshop on Compositional Data Analysis, May 9–13, 2011, Girona, Spain*. http://congress.cimne.com/codawork11/Admin/Files/FilePaper/p33.pdf (Last Accessed 27 August, 2013).

GRUNSKY, E.C., KJARSGAARD, B.A., EGOZCUE, J.J., PAWLOWSKY-GLAHN, V. & THIÓ I FERNÁNDEZ DE HENESTROSA. 2008. Studies in stoichiometry with compositional data. *Proceedings of CODAWORK'08: 3rd International Workshop on Compositional Data Analysis, May 27–30, 2008*, Girona, Spain. http://dugi.doc.udg.edu/bitstream/10256/730/1/Grunsky_etal.pdf (Accessed 27 August, 2013)

HELSEL, D.R. 2012. *Statistics for Censored Environmental Data using Minitab and R*. 2nd Edition. John Wiley & Sons, Inc., Hoboken.

HENGL, T. 2009. *A practical guide to geostatistical mapping*. http://spatial-analyst.net/book/system/files/Hengl_2009_GEOSTATe2c1w.pdf (Accessed January 30, 2012).

HOWARTH, R.J. 1973. The pattern recognition problem in applied geochemistry. *In*: JONES, M.J. (ed.) *Geochemical Exploration 1972*. Institution of Mining and Metallurgy, London, 259–273.

HYVARINEN, A. & OJA, E. 2000. Independent Component Analysis: Algorithms and applications. *Neural Networks*, **13**, 411–430.

IHAKA, R. & GENTLEMAN, R. 1996. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

JANOUŠEK, V., FARROW, C.M. & ERBAN, V. 2006. Interpretation of whole-rock geochemical data in igneous geochemistry: Introducing Geochemical Data Toolkit (GCDkit). *Journal of Petrology*, **47**, 1255–1259.

JANOUŠEK, V., FARROW, C.M., ERBAN, V. & ŠMÍD, J. 2011. GeoChemical Data toolkit (GCDkit) for Windows. http://www.gcdkit.org (Last accessed 27 August, 2013.)

KAISER, H.F. 1958. The Varimax criterion for analytic rotation in factor analysis. *Psychometrika*, **23**, 187–200.

KRUMBEIN, W.C. & IMBRIE, J. 1963. Stratigraphic factor maps. *American Association of Petroleum Geologists, Bulletin*, **47**, 698–701.

MILLER, R.L. & KAHN, J.S. 1962. *Statistical Analysis in the Geological Sciences*. John Wiley & Sons, Inc., New York.

PALAREA-ALBALADEJO, J., MARTIN-FERNANDEZ, J.A. & GOMEZ-GARCIA, J. 2007. A parametric approach for dealing with compositional rounded zeros. *Mathematical Geology*, **39**, 625–645.

PEARCE, T.H. 1968. A contribution to the theory of variation diagrams. *Contributions to Mineralogy and Petrology*, **19**, 142–157.

QGIS. 2011. *Quantum GIS*. http://www.qgis.org (Last accessed 30 January, 2012.)

R-PROJECT. 2013. *The R Project for Statistical Computing*. http://www.r-project.org (Last accessed 27 August, 2013.)

REIMANN, C., FILZMOSER, P. & GARRETT, R.G. 2005. Background and threshold: Critical comparison of methods of determination. *Science of the Total Environment*, **346**, 1–16.

REIMANN, C., FILZMOSER, P., GARRETT, R.G. & DUTTER, R. 2008. *Statistical Data Analysis Explained: Applied Environmental Statistics with R*. John Wiley & Sons, Ltd, Chichester.

RENCZ, A.N. & KETTLES, I.M. (eds) 2011. *Workshop on the role of geochemical data in environmental and human health risk assessment. March 17-18, 2010, Halifax, Canada*. Geological Survey of Canada, Open File 6645.

RENCZ, A.N. GARRETT, R.G., ADCOCK, S.W. & BONHAM-CARTER, G.F. 2006. *Geochemical background in soil and till*. Geological Survey of Canada, Open File **5084**.

ROUSSEEUW, P.J. 1986. Multivariate estimation with high breakdown properties. *In*: GROSSMAN, W., PFLUG, G., VINCZE, I. & WERTZ, W. (eds) *Mathematical Statistics and Applications*. Reidel, Dordrecht, 283–297.

ROUSSEEUW, P.J. & VAN DRIESSEN, K. 1999. A fast algorithm for the Minimum Covariance Determinant estimator. *Technometrics*, **41**, 212–223.

SAMMON, J.W., Jr. 1969. A non-linear mapping for data structure analysis. *Institute of Electrical and Electronics Engineers Transactions*, C-18, 401–409.

SINCLAIR, A.J. 1974. Selection of threshold values in geochemical data using probability graphs. *Journal of Geochemical Exploration*, **3**, 129–149.

SINCLAIR, A.J. 1976. *Application of probability graphs in mineral exploration*. Association of Exploration Geochemists, Special Volume, **4**.

STANLEY, C.R. 1986. *PROBLOT – An interactive computer program to fit mixtures of normal (or log-normal) distributions using maximum-likelihood optimization procedures*. Association of Exploration Geochemists, Special Volume, **14**.

STANLEY, C.R. 2003. Statistical evaluation of anomaly recognition performance. *Geochemistry: Exploration, Environment, Analysis*, **3**, 3–12.

TEMPL, M. 2010. *clustTool: GUI for clustering data with spatial information*. http://cran.r-project.org/web/src/Archive/clustTool (Last accessed 5 March, 2012.)

TEMPL, M., FILZMOSER, P. & REIMANN, C. 2008. Cluster analysis applied to regional geochemical data: Problems and possibilities. *Applied Geochemistry*, **23**, 2198–2213.

TEMPL, M., FILZMOSER, P. & HRON, K. 2011*a*. Analysis of compositional data using robust methods. The R package robCompositions. *Proceedings of CODAWORK 2011: 4th International Workshop on Compositional Data Analysis, May 9–13, 2011, Girona, Spain*. http://congress.cimne.com/codawork11/Admin/Files/FilePaper/p16.pdf (Accessed 27 August, 2013.)

TEMPL, M., HRON, K. & FILZMOSER, P. 2011*b*. *robCompositions: Robust Estimation for Compositional data*. http://cran.r-project.org/web/packages/robCompositions/index.html (Accessed 27 August, 2013.)

TENNANT, C.B. & WHITE, M.L. 1959. Study of the distribution of geochemical data. *Economic Geology*, **54**, 1281–1290.

THOMPSON, M.J. & HOWARTH, R.J. 1973. The rapid estimation and control of precision by duplicate analyses. *The Analyst*, **98**, 153–160.

THOMPSON, M.J. & HOWARTH, R.J. 1978. A new approach to the estimation of analytical precision. *Journal of Geochemical Exploration*, **9**, 23–30.

TUKEY, J.W. 1977. *Exploratory Data Analysis*. Wesley, Reading.

U.S. ENVIRONMENTAL PROTECTION AGENCY. 1996. *Method 3050b – Acid digestion of sediments, sludges, and soils*. http://www.epa.gov/wastes/hazard/testmethods/sw846/pdfs/3050b.pdf (Accessed 27 August, 2013.)

VAN DEN BOOGAART, K.G. & TOLOSANA-DELGADO, R. 2008. "compositions": A unified R package to analyze compositional data. *Computers & Geosciences*, **34**, 320–338.

VAN DEN BOOGAART, K.G., TOLOSANA-DELGADO, R. & BREN, R. 2011. *Package "compositions" v_1.10.2*. http://cran.r-project.org/web/packages/compositions/index.html (Last accessed 27 August, 2013.)

VELLEMAN, P.F. & HOAGLIN, D.C. 1981. *Applications, Basics and Computing of Exploratory Data Analysis*. Duxbury Press, Boston.

VENABLES, W.N. & RIPLEY, B.D. 2001. *Modern Applied Statistics with S-PLUS*. 3rd Edition. Springer-Verlag, New York.

WEISBERG, S. 1985. *Applied Regression Analysis*. 2nd Edition. John Wiley & Sons, New York.

Received 29 November 2011; revised typescript accepted 13 April 2012.