What's in 'rgr 1.1.1' ?

1. Statistical graphics functions:

gx.hist	Plots histograms using a variety of bin width selection methods.
cnpplt	Plots a normal cumulative percent probability (CPP) plot.
gx.cnpplts	Plots up to nine CPP plots in a single display, these may be either data subsets or variables.
gx.cnpplts.setup	Permits the user to define the symbols and their colours for the up to nine data sets to be plotted with gx.cnpplts.
gx.ecdf	Plots an empirical cumulative distribution function (ECDF).
gx.ks.test	Plots two ECDFs in a single display and carries out a Kolmogorov-Smirnov test for the two populations being drawn from the same underlying population.
bxplot	Plots a horizontal Tukey boxplot or a box-and-whisker plot.
shape	Plots a combination of histogram, boxplot, ECDF and CPP on a single screen/page.
inset	Plots a combination of a histogram and a CPP plot, together with some summary statistics, for use as an inset on a geochemical map.
inset.exporter	A version of 'inset' for use in a production environment where the graphics file is saved as defined by the user for later map production.
bwplots	Plots vertical box-and-whisker plots for a single variable subdivided into various subsets (groups or factors).
bwplots.by.var	Uses 'bwplot' to plot different variables (elements) side-by-side.
tbplots	Plots vertical Tukey boxplots for a single variable subdivided into various subsets (groups or factors).
tbplots.by.var	Uses 'tbplot' to plot different variables (elements) side-by-side.

All the above functions permit both normal arithmetic and logarithmic scaling, and user-defined axis labelling and titling. The 'bwplot' and 'tbplot' functions permit the groups (factors) to be ordered (left-to-right) and labelled as defined by the user.

2. Mapping and XY Plotting functions:

Mapping:

map.eda7	Displays a map using symbols that correspond to a Tukey boxplot, i.e. lower near and far outliers, in the lower whisker, in the mid 50%, in the upper whisker, etc.
map.eda8	Displays a map using symbols to indicate the magnitude of a variable (element) subdivided by the 2^{nd}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th} and 98^{th} percentiles.
map.tags	Displays a map of posted values.
map.z	Displays a map using circles that increase in diameter with magnitude of the variable (element) being plotted. The rate of increase of symbol size may be user-defined.

caplot — Displays a concentration-area (C-A) plot to assess whether the data are spatially multifractal. The data may be optionally log-transformed, and the interpolated estimates may be accumulated in either direction.

The above functions require that the R library packages 'MASS' and 'akima' (caplot only) be available at run-time. All the above functions require that rectangular coordinates are available for the data points, and permit user-defined axis labelling, titling, and symbol colour and scaling.

Note: the EDA mapping functions are not provided to replace a full mapping or GIS package, but to provide a 'quick-look' in order to appreciate the spatial distribution of the data and to support threshold (upper limit of geochemical background) selection.

Plotting:

xyplot.eda7 — Displays a XY plot using symbols for the third that correspond to a Tukey boxplot, i.e. lower near and far outliers, in the lower whisker, in the mid 50%, in the upper whisker, etc.

xyplot.eda8 — Displays a XY plot using symbols for the third variable to indicate the magnitude of a variable (element) subdivided by the 2^{nd}, 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th} and 98^{th} percentiles.

xyplot.tags — Displays a XY plot of the posted values of a third variable.

xyplot.z — Displays a XY plot using circles that increase in diameter with magnitude of the third variable (element) being plotted. The rate of increase of symbol size may be user-defined.

3. Summary statistics functions:

gx.stats — Computes and displays summary statistics as displayed with 'inset'.

gx.summary1 — Displays a concise one-line summary statistics report.

gx.summary.mat — Displays a concise one-line summary statistics report for selected columns of a dataframe or matrix.

gx.summary.groups — Displays a concise one-line summary statistics report for data subsets grouped by factor name in a dataframe or matrix.

gx.summary2 — Displays a five-line summary statistics report.

fences — Computes and displays the various estimates of background range discussed in Reimann, Filzmoser & Garrett, 2005.

fences.summary — A version of 'fences' for when it is required to estimate background ranges for various subsets (groups or factors) of a variable (element) and to save them in a user-defined 'txt' file for later inspection or other use.

framework.summary — Computes summary statistics for various data subsets (groups or factors), e.g., EcoProvinces, Great Soil Groups, Lithological units, etc., of a variable

(element) and saves them in a user-defined 'csv' file for later inspection with a spreadsheet program, e.g., Excel™.

gx.fractile — Estimate the fractile for a specified quantile of a distribution.

gx.fractile — Estimate the quantile for a specified fractile of a distribution.

4. Bivariate and Multivariate functions:

gx.pearson — Estimate the Pearson product moment correlation correlations for a matrix or columns of a dataframe. The coefficients are displayed in the upper triangle and the significance of them not being due to chance (Ho: coefficient = 0) is displayed in the lower triangle.

gx.spearman — Estimate the Spearman rank correlation correlations for a matrix or columns of a dataframe. The coefficients are displayed in the upper triangle and the significance of them not being due to chance (Ho: coefficient = 0) is displayed in the lower triangle.

gx.rma — Estimates the coefficients of the Reduced Major Axis (RMA) for quantifying the relationship between two independent variables, such as analyses of the same samples by two independent methods. Confidence bounds are estimated for the coefficients to assist in determining if they are significantly different from (0,1).

wtd.sums — Computes weighted sums (see Garrett and Grunsky, 2001) for a set of user defined variables and their 'relative importances'.

gx.2dproj — Computes 2-d projections of p-space data via the Sammon non-linear mapping, metric or non-metric multidimensional scaling, or projection pursuit procedures. Closed compositional, geochemical, data should be passed to this function with an ilr transformation. A log transformation and a variety of data normalization procedures are available as pre-processing options for open, non-compositional, data sets.

gx.2dproj.plot — Displays the results saved from function gx.2dproj.

gx.mva — Undertakes a R-Q Principal Components Analysis (PCA) and estimates Mahalanobis distances using classical covariance estimation procedures.

gx.robmva — Undertakes a R-Q Principal Components Analysis (PCA) and estimates Mahalanobis distances using robust covariance estimation procedures. Options include the minimum covariance determinant (MCD), minimum volume ellipsoid (MVE) procedures, and the use of user-defined weights – not necessarily just zero or one.

gx.robmva.closed — As above but for closed, compositional data. Estimates robust statistics following an ilr transform, then back-transforms to the clr form for the PCA.

gx.rotate — Undertakes a Kaiser Varimax rotation on the PCs saved from either gx.mva, gx.robmva, or gx.robmva.closed.

gx.rqpca.screeplot — Displays a scree plot for the results of a PCA saved from either gx.mva, gx.robmva or gx.robmva.closed.

gx.rqpca.print Displays tables of the loadings of each variable on each PC, and scores of the individuals on the PCs. Optionally the percentage contribution of each variable to the variability of each PC may be displayed.

gx.rqpca.loadplot
Displays a graphic where the loadings above some critical value, 'crit', are plotted for each PC in a space proportional to the variability that each PC contributes to the total data variability.

gx.rqpca.plot Displays R-Q bi-plots for the results of a PCA saved from either gx.mva, or gx.robmva, or gx.robmva.closed.

gx.md.gait Undertakes a graphical adaptive interactive trimming (GAIT) procedure based on Mahalanobis distance estimation for multivariate outlier detection and the selection of clean 'reference' data subsets for use by gx.mv.alloc.

gx.md.gait.closed
As for gx.md.gait, but specifically for closed compositional, geochemical, data. The data are isometrically log-ratio (ilr) transformed for computational purposes, and then back transformed to the centred log-ratio (clr) basis for subsequent processing by gx.mvalloc.closed.

gx.add.chisq Adds probability based fences to Chi-square plots of Mahalanobis distances to facilitate the recognition of outliers and extreme values.

gx.md.plot Displays Chi-square plots for the Mahalanobis distances saved from either gx.mva, gx.robmva or gx.md.gait.

gx.md.print Displays lists of all individuals and/or saves the list as a 'csv' file, or lists individuals whose predicted probabilities of group membership fall below a user-defined cut-off value, i.e. multivariate outliers.

gx.mvalloc A typicality, Mahalanobis distance, based allocation procedure where an individual may be classified into one of up to nine 'reference', geochemical 'background', groups. If typicality, group membership, in all reference groups falls below a used defined cut-off, probability, those individuals are identified as outliers, 'anomalies', and allocated to an 'outlier' group. The allocations are made on the assumption that the reference group covariances are heteroscedastic, i.e. are of different sizes, shapes and orientations

gx.mvalloc.closed
As for gx.mvalloc, but specifically for closed compositional, geochemical data. The reference groups have to have been generated by gx.md.gait.closed or gx.robmva.closed, both of which return an inverse of the robust clr covariance matrix required by the allocation procedure. The input data are clr transformed prior to use in typicality estimations.

gx.mvalloc.print
Displays lists of all individuals and/or saves the list as a 'csv' file, or lists only individuals whose predicted probabilities of group membership fall below a user-defined cut-off value, i.e. multivariate outliers.

gx.lm.vif Estimates the variance inflation factor as a measure of collinearity in the independent (predictor) variables of a linear model.

gx.adj2 Calculates adjusted R^2 values for multiple regression (linear) models taking into account the number of cases (individuals) and independent (predictor) variables.

5. QA/QC support functions:

anova1	Computes a random effects model ANOVA (Analysis of Variance) on a set of duplicate measurements to determine if the analytical, or combined sampling and analytical (within) variability is significantly smaller than the variability between the duplicates. For use where the n duplicates are stored as x1 and x2 in n rows.
anova2	Similar to 'anova1' but for use where the duplicates are stored as 1 to n values of x1 followed by 1 to n values of x2, or as alternating rows of x1 and x2 values.
gx.triples.aov	Computes a random effects model ANOVA and estimates the variance components for a staggered 3-level design of field and analytical triplicates to simultaneously evaluate the significance and relative magnitude of 'regional', local sampling and analytical variability.
gx.triples.fgx	Computes two random effects model ANOVAs to estimate the regional representivity of the 'triples' in the context of the total regional survey variability and the equivalence of the variability of the two field duplicates.
thplot1	Displays a Thompson-Howarth plot for duplicate measurements to visually inspect them as a part of the QA/QC process. A target precision may be entered to aid visual data inspection. For use where the n duplicates are stored as x1 and x2 in n rows.
thplot2	Similar to 'thplot1' but for use where the duplicates are stored as 1 to n values of x1 followed by 1 to n values of x2, or as alternating rows of x1 and x2 values.

All of 'anova1', 'anova2', 'gx.triples.aov' and 'gx.triple.fgx' provide for an optional log-transformation of the data in order to meet homogeneity of variance and normality requirements.

6. Data conditioning functions:

ltdl.fix	Replaces less-than-detection values recorded as –x with x/2. Optionally zero values and/or coded values, e.g., -9999, may be set to a NA, a code used in the S-Language to represent no information, i.e. blank.
ltdl.fix.df	Performs a 'ltdl.fix' on a dataframe, any factor variables are transferred to the new dataframe.
remove.na	Removes any NAs from a vector or matrix, reporting on the number of NA values, or NA containing matrix rows, removed and the number of remaining rows and columns for a matrix.
gx.subset	Extracts a subset of rows from a dataframe on the basis of a criterion supplied by the user, returning a new dataframe.
alr	Computes arithmetic log-ratios for a matrix in order to remove the effects of data closure.
clr	Computes centred log-ratios for a matrix in order to remove the effects of data closure.

ilr	Computes isometric log-ratios for a matrix in order to remove the effects of data closure.
rng	Computes range transformations on the columns of a matrix.

7. Utility functions:

df.test	Determines if a specific dataframe is available (attached) or exists in the working directory. If it does, the names of the variables are displayed, and additionally if a specific legitimate variable name is entered the number of values, length of the vector, is displayed.
where.na	Identifies any positions in a vector or matrix containing NAs, and can be used to remove any NAs from a data vector or matrix.
gx.sort	Displays a sorted, or reverse sorted, dataframe or matrix on the value of a specified column.
gx.sort.df	Displays a sorted dataframe on the basis of any combination of numeric or factor variables in any combination of ascending or descending orders.
gx.hypergeom	Estimate the probability that anomaly (above threshold sites) locations are informative, i.e. coincide with an expected model along transects or traverses (see Stanley, 2003).
gx.runs	Carry out a Wald-Wolfowitz, Runs, test for pattern coherence along transects and traverses.
display.lty	Displays the available line types and colours.
display.marks	Displays the available plotting symbols.
display.ascii.o	Displays the octal numbers corresponding to the Windows Latin 1 font, these are required when inserting symbols such as µ or ° into an axis label or title.
display.rainbow	Displays the 36 colours of the "rainbow" palette.
syms.pfunc	Displays the effect of changing the parameter 'p', which controls the rate of change of circular symbol size, in the 'edamap' function.

Notes:

Dataframes are a data management feature of the S language, they accommodate row and column names, real numbers, factor variables and NAs. NAs are a S and R language feature for identifying data items for which there is no information, as such it is a 'special code' for a 'blank' in a data file.

The boxes of Tukey boxplots, box-and-whisker plots and histograms are infilled in grey (8) from the palette displayed in 'display.lty', alternate colours may be selected from that palette. The "rainbow" palette is used for symbol colours in 'map.eda7', 'map.eda8', 'xyplot.eda7' and 'xyplot.eda8' the user may select alternate colours from this palette if required.

The above list of 80 functions only includes those directly accessible by a user, it does not include some functions that 'lurk in the basement' and are used as 'engines' to achieve the desired graphical and tabular displays.

Changes since 'rgr 1.0.3'

'rgr 1.0.4' was a maintenance release built with R 2.12.0, no new functions were added. Minor changes were made to functions: caplot, edamap and framework.stats. Two utility functions were removed: display.alts and display.ascii.d.

The release was required to bring the help, Rd, files into conformity with the new parsing rules for R 2.11. Some other text changes were made to the help files for shape and fences.summary.

Changes since 'rgr 1.0.4'

'rgr 1.0.5' was built with R 2.12.0 and includes a number of additional functions for the display of univariate data, QA/QC and utility functions. The names of the 'map' functions and dftest were changed to achieve better internal consistency in naming. Some multivariate or pseudo-multivariate functions were added, these carry out various log-ratio transformations to remove the closure effect from compositional (constant sum) data, to compute Pearson and Spearman correlation coefficients and their significance, and to compute weighted sums.

Changes since 'rgr 1.0.5'

'rgr 1.1.0' was built with R 2.12.0. At the maintenance level it includes some corrections of 'typos' in the help file, and a correction to Reduced Major Axis slope estimation (gx.rma). A function is now present to identify the locations of any NAs in a data vector or matrix. The major additions to version 1.1.0 are functions for multivariate exploratory data analysis. These include:

1. 2-d projections of p-space data using Sammon's non-linear mapping and multidimensional scaling. Unfortunately the Friedman and Rafsky implementation of minimum spanning tree 'planing' is not available in the R version of the Venables and Ripley MASS Library;
2. Both classical and robust estimation implementations of R-Q Principal Components Analysis (PCA) and the estimation of Mahalanobis distances. Options for robust estimation include the minimum covariance determinant and minimum volume ellipsoid procedures, and user-supplied weight – not necessarily 0 or 1. A function is provided for the Kaiser Varimax rotation, as are functions for the display of the various results;
3. A robust estimation procedure for undertaking a PCA and the computation of Mahalanobis distances specifically for closed, compositional, data;
4. A graphical adaptive interactive trimming (GAIT) procedure based on Mahalanobis distance estimation for multivariate outlier detection and the selection of clean 'reference', in the applied geochemical context 'background', data subsets. Two functions are provided for the display of the GAIT results, one permits saving the results as a 'csv' file;
5. A Mahalanobis distance based multivariate allocation procedure employing predicted probabilities of reference, geochemical 'background', group membership, i.e. typicalities. Atypical individuals whose membership falls below a user-defined cut-off in any of the reference groups are grouped together as 'unallocated' for further inspection. The allocations are made on the assumption that the reference group covariances are heteroscedastic, i.e. are of different sizes, shapes and orientations. The results may be displayed and/or saved as a 'csv' file; and

6. Two utility functions for linear modelling exercises are included, for estimating variance inflation factors and computing adjusted R^2 values. However, in the latter context, the use of Akaike's Information Criterion, available in the R {stats} library, is recommended.

Changes since 'rgr 1.1.0'

'rgr 1.1.1' was built with R 2.12.0. At the maintenance level it includes corrections of 'typos' in the help files, and changes in the names of functions bwplot, bwplot.by.var, tbplot and tbplot.by.var to bwplots, bwplots.by.var, tbplots and tbplots.by.var. These changes were made to avoid a conflict with function bwplot in the display package 'lattice' with the function of the same name in 'rgr', the 's' was added to the related function names for the sake of consistency. Other changes include:

1. The additions of Kruskal's non-metric multidimensional scaling procedure and a projection pursuit procedure based on the fastICA function; and
2. The addition of two functions, gx.md.gait.closed and gx.mvalloc.closed, together with a modification to gx.rob.mva closed. These additions/changes facilitate the investigation of closed data sets using Mahalanobis distance based procedures. Function gx,mvalloc.closed now requires the inverse of the reference data covariance matrix. This is achieved by back-transforming the inverse of the ilr transformed covariance matrix used for robust estimation in functions gx.md.gait.closed and gx.robmva.closed to the clr basis.
3. The addition of two functions to display the results of Principal Components Analyses, functions gx.rqpca.print and gx.rqpca.loadplot. The former displays tables of the PC loadings and scores of the individuals on the PCs, in addition, the percentage contribution of each variable to the variability of each PC may be displayed. The latter function displays a graphic where the loadings above some critical value, 'crit', set by default to an absolute loading of 0.3, are plotted for each PC in a space proportional to the variability that each PC contributes to the total data variability.

Issues that remain to be addressed in future releases of 'rgr'are:

1. The provision of the Friedman and Rafsky 'planing' tool for visualizing p-space data in 2-d for function gx.2dproj;
2. The correction of any errors in the scripts or help files (manual) as they are identified; and
3. The addition of 'useful' display functions as they are identified and/or developed.

Robert G. Garrett
2011/02/26