ESS / SST

Establishing Background Values in Geochemical Data: Open Source R language tools

Robert G. Garrett *Geological Survey of Canada, Ottawa, Ontario*

Environment and Health

North American Soil Geochemical Landscapes Project

Abstract

Data from geochemical mapping exercises, whether the scale is regional or local, are characterized by two properties - level and relief. Level is a measure of the absolute values in quantification units, and relief is a measure of data heterogeneity, or lack thereof. Areas of complex geology and geochemistry, due to the presence of multiple lithologies and/or the effects of multiple geochemical processes in the primary bedrock and/or secondary surficial environments, are characterized by high relief. The selection of spatial units for which background and relief are to be estimated is an important consideration, to the greatest possible extent they should represent a single entity in the context of the study being undertaken.

Average level is best estimated by the median, a robust measure of central location that is uninfluenced by up to 50% extreme values. However, the average level tells only part of the story. If a question such as, "is a value background?" is to be answered it immediately follows that the range of background has to be known. The estimation of background range is a task that requires the integration of graphical data inspection, the calculation of statistically based estimates, and geochemical knowledge. To support this activity a software package, 'rgr' has been prepared in the Open Source R language that displays the data in a variety of graphical formats familiar to applied geochemists, and computes a variety of estimates of background range based on non-parametric and traditional and robust parametric estimation procedures. A key display in helping visualize background ranges is the Tukey boxplot, see below. The poster illustrates the various tools available within 'rgr' and discusses their relative merits.

Some History

The statistical and graphical tools in the package 'rgr' were developed in the S-Plus™ software (Insightful Corporation, Seattle, WA) over a period of some 10 years to meet internal Geological Survey of Canada (GSC) needs. In 2005 the Contaminated Sites Division of Health Canada requested reports on background levels for a wide range of trace elements in Canadian soils and glacial sediments. Following submission of these reports, a further request was received from Health Canada for the software used to support these reports. A decision was made to translate the S-Plus™ software, a proprietary commercial package, into the Open Source R language. The preparation of 'rgr' for formal submission to the Comprehensive R Archival Network (CRAN) required not only the language conversion but the preparation of extensive help and documentation files that are an integral part of the package. The help of Yiwen Chen in that work is gratefully acknowledged. On submission to CRAN versions that can be loaded into MacOS X and Linux environments are prepared and available.

What's required to run R and 'rgr'?

The 'rgr' package runs comfortably on a PC with an 800 MHz processor and 0.5 GB of memory (my laptop). Access to a colour printer is advantageous, but by changing symbol and palette colours suitable gray-scale images can be prepared for publication. R contains drivers to save graphics in Windows metafile, PNG, JPEG, BMP, PostScript and PDF formats.

At the user level, core R does not have a GUI (Graphical User Interface). However, extensive on-line help files and the command line provide great flexibility in selecting appropriate options and manipulating the graphics to meet presentation needs. The 'rgr' documentation provides a number of tips to aid use, and the help files are replete with examples of how to take advantage of options to manipulate the graphics, labelling, and titling, etc.

Within core R there are functions such as 'line', 'abline' and 'text' that facilitate adding annotation, etc., to images. Finally, the great advantage of learning to use R is the wide range of functionality available in core R and the 1566 (2008.09.12) contributed packages available. All Open Source, as scientists help scientists.

'rgr' contains several data conditioning and utility functions not demonstrated in this poster. These provide tools for handling <DL data recorded as negatives, and setting zeros and coded values to "NAs", no information, that may be removed prior to numerical calculations. Data subsets may be created on the basis of criteria supplied by the user, e.g., the data for a particular Great Soil Group or EcoProvince. Utility functions are provided to display codes for different plotting symbols, line styles, colours, and the effect of changing a parameter, *p*, for mapping.

Two Example Datasets

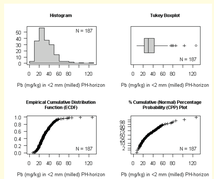
To demonstrate the facilities of 'rgr' the Maritimes (2007) data from the Canadian portion of the Tri-National North American Geochemical Landscapes project are used. The field and analytical data are received from GSC field crews and laboratory contractors, respectively, in a number of different formats and compiled into an Access™ database from which Excel™ files are extracted. These can be directly imported in to R using the 'RODBC' package, or read into R from a .csv file derived from the corresponding Excel™ file using the core R 'read.table' or 'read.csv' functions.

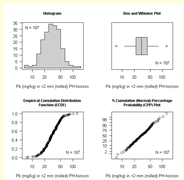
The following examples use As, Ni and Pb data for the Public Health (PH) horizon, the 0-5 cm interval, or the C-horizon of the soil profile to demonstrate the capabilities of 'rgr'. The former are particular interesting in the context of anthropogenic contamination of the surface environment to which people are most often exposed, and the C-horizon that is closest to the soil's geological parent material. For both horizons the <2 mm fraction was milled to finer than approximately 100 µm, and following a 4-acid (HF-$HClO_4$-HNO_3-HCl) digestion the elements determined by ICP-MS.

Statistical Graphics Functions:

Ten functions are available to plot histograms, empirical cumulative distribution and normal cumulative probability plots, box-and-whisker plots and Tukey boxplots. An additional function facilitates preparing and saving plots for latter use as insets on maps. The box-and-whisker and Tukey boxplot functions permit the data to be subdivided into groups (factors) that can be ordered (left-to-right) and labelled as defined by the user. Box-and-whisker plots indicate the data maxima and minima, some upper and lower percentile for the ends of the whiskers, user defined, but by default the 95th and 5th percentiles, and the boxes span the three quartiles, optionally the median line can be replaced by a notch indicating the 95% confidence interval on the median. Tukey boxplots are similar, but the whiskers only extend to actual values and beyond the whisker ends individual high and low, near and far, outliers as defined by the Tukey procedure, are individually plotted in with different symbols.

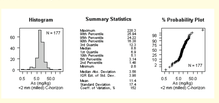
shape – displays a combination histogram (**gx.hist**), box-and-whisker plot or Tukey boxplt (**bxplot**), empirical cumulative distribution function (**gx.ecdf**), and a cumulative probability plot (**cnpplt**). Each of these may be displayed individually if required. An option to log transform the data is available, in which case the histogram bins are equal-log sizes, the axes are logarithmically scaled and the Tukey boxplot calculations are undertaken appropriately. In **shape** the box-and-whisker plots and Tukey boxplots are not notched.

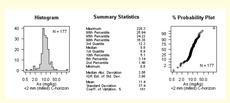




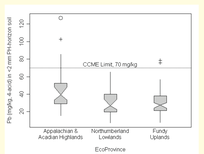
The left display presents the default with no log-transformation and a Tukey boxplot. The box-and-whisker plot option is presented on the right with the whisker ends extending to the 2nd and 98th percentiles, log-scaling of the axes, the default plus sign plotting symbol changed to an open diamond, and for cosmetic reasons the sample size, N, has been moved to the upper left side of the histogram. These two displays scale for an interesting comparison. The display with no log-transformation clearly demonstrates the positively skewed nature of the data. The far outlier is related to emissions from the Belledune, NB, Pb-smelter, and the near outliers to the smelter and other anthropogenic and natural sources (see following map). In contrast the use of a log-transformation obscures this interesting data structure.

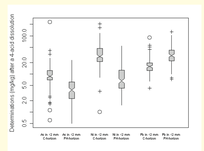
inset and **inset.exporter** – were designed to support map production, where a statistical summary is a useful adjunct to the spatial display. The **inset.exporter** version presents no display, but saves the **inset** graphics as a uniquely named file employing the variable name in the current working directory. Options are available for different histogram bin width selection procedures, logarithmic scaling and moving the sample size indication to the top left corner of the histogram display. The default display is shown left. The display to the right uses the "Freedman-Diaconis" rule for histogram bin width selection rather than the default "Scott" rule, this results in a greater number of bins, and an allocation that is more resistant to the effects of outliers, which exist in this example.





bwplot and **tbplot** – the only difference between these two displays is style and the options specific to the box-and-whisker or Tukey boxplot formats. For the purpose of this example the Tukey boxplot will be used, left. A large number of options concerning individual plot positioning and spacing, labelling, and additional text are available to improve the readability of the graphical output. It is immediately apparent that the distribution of Pb in the Public Health (0-5 cm) horizon is quite different in the Appalachian and Acadian Highlands than it is in the Northumberland Lowlands and Fundy Uplands. The Canadian Council of Ministers of the Environment (CCME) limit for Pb in soil has been added to demonstrate how location, here defined by EcoProvince, is an important consideration.
bwplot.by.var and **tbplot.by.var** – whereas **bwplot** and **tbplot** facilate the plotting of data for subsets of a single element, **bwplot.by.var** and **tbplot.by.var** facilitate the plotting of different measurements (elements), right. The same options for labelling, etc., as are in **bwplot** and **tbplot** are also available.





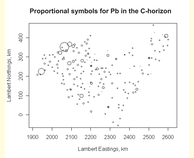
Mapping Functions:

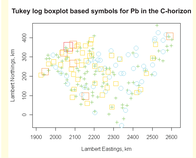
Three functions permit simple spatial data presentations (maps) to be displayed when rectangular sample site coordinates are available. These EDA mapping functions are not provided to replace a full mapping or GIS package, but to provide a quick-look in order to appreciate the spatial distribution of the data and to support threshold (upper limit of geochemical background) selection. A fourth function displays a concentration-area (C-A) plot (Cheng et al., 1994) to determine if the data are spatially multifractal and assist in selecting background ranges. The data may be optionally log-transformed, and the interpolated estimates may be accumulated in either direction, upwards or downwards. This latter function is not demonstrated here, as to yield useful results the data need to be distributed relatively evenly across an area with few 'gaps'.

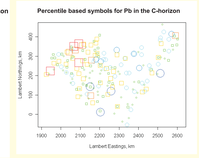
edamap – plots a map, left, using circles centred on the sample site location that increase in diameter with magnitude of the variable (element) being plotted. The rate of increase of symbol size may be user-defined, parameter p, as can the absolute symbol size. The manner in which parameter p influences the rate of change of symbol size is demonstrated in the utility function **syms.pfunc**. Additionally, upper and lower limits of the range over which the scaling function operates can be set so as to limit the effect of gross outliers in the data. In the example, parameter p was set to 2, the scaling factor to 3.

edamap7 –plots a 7-symbol map, centre, where the symbols correspond to the ranges of the Tukey boxplot for the same data. The mid 50% of the data are plotted as a small cross, and three upper and lower classes corresponding to the whiskers, near and far outliers plotted as increasingly sized squares or circles, respectively. A default set of colours is provided, green for the mid 50%, blues for lower classes and browns for upper classes. These may be changed to others from the rainbow palette, **display.rainbow**, or a grey-scale selected. Finally, the data may be optionally log-transformed prior to the Tukey boxplot calculations, as was done for the example below that used the default colours.

edamap8 – plots a map, right, similar to edamap7 but the symbol ranges are based on the percentiles of the data being displayed, specifically the 2nd, 5th, 25th, 50th, 75th, 95th and 98th percentiles. The same symbol shape and colour allocation follows that for **edamap7**, however the middle 50% of the data are divided into two groups median to 75th percentile a small green square, and median to 25th percentile a small blue circle.





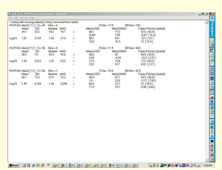


Summary Statistics Functions:

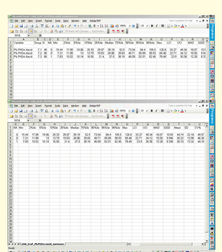
Three functions compute summary statistics. **gx.stats** provides a screen display, a subset of these estimates is presented in the statistical graphics function **inset** for preparing plots for use as insets on maps. **fences** displays the various estimates to support background range selection discussed in Reimann, Filzmoser & Garrett, 2005. **fences.summary** is provided to compute these estimates for various subsets (groups or factors) of a variable (element) and to save them in a user-defined file for later inspection. A third function, **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.stats(Pb.PH2m.4acid,"Pb (mg/kg) in <2 mm (milled) PH-horizon")
 1 row(s) with missing value(s), NA(s), removed from vector

Summary Statistics Display for: Pb (mg/kg) in <2 mm (milled) PH-horizon

Data Set N = 187
Minimum = 7.21              Maximum = 126.8
Median = 29.95              MAD Est = 14.71
                            IQR Est = 15
Mean = 33.67                S.D. = 17.15
Variance = 294.2            C.V. % = 50.94

Table of Percentiles

Maximum Value    126.8
99th Percentile   87.8
98th Percentile   79.02
95th Percentile   61.36
90th Percentile   52.79
80th Percentile   45.48
3rd  Quartile (75th) 42.37
70th Percentile   39.21
60th Percentile   34.08
Median (50th)     29.95
40th Percentile   26.97
30th Percentile   23.04
1st Quartile (25th) 21.99
20th Percentile   19.93
10th Percentile   16.39
5th Percentile    13.75
2nd Percentile    11.52
1st Percentile     9.559
Minimum Value      7.21
```

```
> fences.summary(EcoP.Pb.PH2m.4acid,file="GSA")
 Variable Pb.PH2m.4acid subset by EcoP - output will be in
 D:/R/WD/GSA_EcoP_Pb.PH2m.4acid_fences.txt
```



```
> framework.summary(EcoP.Pb.PH2m.4acid,file="GSA")
 Variable Pb.PH2m.4acid subset by EcoP - output will be in
 D:/R/WD/GSA_EcoP_Pb.PH2m.4acid_summary.csv
```



QA/QC Support Functions:

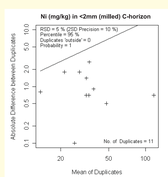
Checking the adequacy of data for the task in hand is an essential part of applied geochemistry. Duplicate analyses are frequently used, and two functions are available to support this activity. For each there are two versions so that three different data formats can be accommodated, duplicates in the same record, interleaved, or sequentially.

anova1 and **anova2** - compute random effects model Analyses of Variance (ANOVAs) 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. Provision is made for an optional log-transformation of the data in order to meet homogeneity of variance and normality requirements.

```
> anova2(Ni.C2.4acid,"Ni (mg/kg) in <2 mm (milled) C-horizon",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 Ni (mg/kg) in <2 mm (…) C-horizon
Source      SS     df      MS        F        Prob
Between    15141   10    1514.1    4744.61   0.0113
Within      6.2637  1     6.2637    19.63    0.0013
Residual    3.1913  10    0.31913
Total      15151   21    721.47

One-Way Random Effects Model for Ni (mg/kg) in <2 mm (…) C-horizon
Source      SS     df      MS        F        Prob
Between    15141   10    1514.1    1761.55   0
Within      9.455  11    0.85955
Total      5151    21    721.47

Source      MS        Var Comp     %age
Between     1514.1    756.64       99.9
Within      0.85955   0.85955      0.1
            757.5

Summary Statistics for Ni (mg/kg) in <2 mm (milled) C-horizon
Grand Mean = 38.705          Variance = 721.47
'Error' S^2 = 0.85955        Std. Dev. = 0.92712
'Error' RSD% = 2.4
Miesch's V = 880.28          Vm = 1760.55
```

From the above one can conclude that almost all the variability in the duplicates pairs is between them, and only 0.1% is within them. Furthermore the RSD for the analyses is 2.4%. Both of these indicate the data are suitable for purpose, i.e. geochemical mapping.

tbplot1 and **tbplot2** - display Thompson-Howarth plots 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.

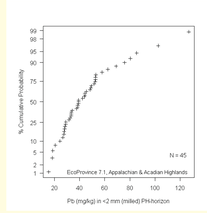


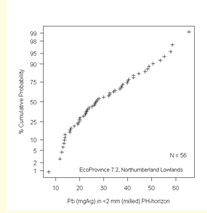
As none of the duplicates plot above the diagonal line one can be confident, at the 95% level, that the target of a 5% RSD for the analyses has been met.

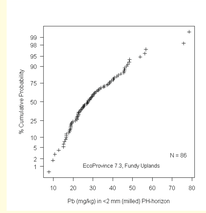
Background Range Selection

How the tools help background range selection is demonstrated with the data for Pb in the Public Health (PH) horizon.

The Tukey boxplot (**tbplot**) for Pb in the PH-horizon of the different EcoProvinces indicates that its distribution in the Northumberland Lowlands and Fundy Uplands is quite similar, except for two outliers in the Fundy Uplands. These outliers are associated with anthropogenic emissions from the Sydney, NS, and Fredericton, NB, areas (see centre **edamap7** display). The Appalachian and Acadian Highlands Pb data clearly reflect different processes. Calculations in **fences** to derive estimates of background range provide conflicting estimates (see summary statistics functions). The best insight is provided through inspection of the cumulative normal probability plots (**cnpplt**) for the Pb data subdivided by EcoProvince.







In combination with the estimates from the fences estimates the following proposal for background ranges can be made.

7.1 Appalachian and Acadian Highlands: Calculated upper limits of background lie in the range of 71 to 127 mg/kg. Which is most appropriate? Here graphics can help inform the decision. The data fall into three groups, 10 to 55 mg/kg, six sites fall in the range 55 to 90 mg/kg, and two above 90 mg/kg. Of the two sites above 90 kg/mg, one is influenced by Belledune, NB, Pb-smelter emissions, the other is in a remote area of northwestern New Brunswick. Similarly, half the 55 to 90 mg/kg sites are likely related to Belledune, and others are elswhere. Northern New Brunswick includes the Bathurst base-metal mining camp, and is characterized by numerous known mineral occurrences that have not been subject to mining. These mineral occurrences can lead to naturally elevated levels of Pb in the surface layer of the soil. Thus for this EcoProvince it is realistic to have two background ranges, 10 to 55 mg/kg for areas without known mineral occurrences, and 10 to 90 mg/kg for areas containing base-metal mineral occurrences. Only study of the specific sites in this second group will determine if high Pb levels are due to natural or anthropogenic processes.

7.2 Northumberland Lowlands and 7.3 Fundy Uplands: Calculated upper limits of background lie in the 52 to 113 mg/kg range. Apart from the two anthropogenically related outliers in the Fundy Uplands, reasonable estimates of background range on the basis of the cumulative normal probability plots for these EcoProvinces would be 5 to 55 and 5 to 80 mg/kg, respectively. A total of 4 sites have values in the low 60s, if further investigation indicated these levels to be due to natural geological and pedological processes the upper limit of background variation could be raised to 65 mg/kg. However, until that can be shown it would be prudent to keep the upper limit of background variation in the 55 to 60 mg/kg range.

The above exemplifies the importance of viewing the diverse calculated estimates of background variation graphically. It is the insight that statistical graphics provide, together with geological and other relevant knowledge, that lends confidence to the proposed ranges of backbround variation.

For the Maritime Provinces as a whole, a background range of 5 to 55 mg/kg of Pb in the <2 mm fraction of the Public Health (PH) horizon can be proposed. An extended high background range of 55 to 90 mg/kg includes sites influences by both natural and anthropogenic processes. Unless a sample site occurred close to an undiscovered Pb-bearing mineral occurrence, natural levels greater than 90 mg/kg are extremely unlikely.

Conclusions

The rgr package puts into the public domain Open Source software that can be used to support decision making surrounding the activities of establishing natural geochemical background levels. The package will continue to be developed and updated as new tools become available.

Acknowledgment

Sincere thanks go to Susan Davis of the Geological Survey of Canada for her assistance in preparing this poster.

Selected Bibliography

Concerning geochemical background, etc.

Cheng, Q., Agterberg, F.P. and Ballantyne, S.B., 1994. The separation of geochemical anomalies from background by fractal methods. Journal of Geochemical Exploration, vol. 51, no. 2, pp. 109-130.

Reimann, C. and Garrett, R.G., 2005. Geochemical background - Concept and reality. Science of the Total Environment, vol. 350, no. 1/3, pp.12-27.

Reimann, C., Filzmoser, P. and Garrett, R.G., 2005. Background and threshold: critical comparison of methods of determination. Science of the Total Environment, vol. 346, no. 1/3, pp. 1-16.

Thompson, M. and Howarth, R.J., 1978. A new approach to the estimation of analytical precision. Journal of Geochemical Exploration, vol. 9, no. 1, pp. 23-30.

Concerning the S language, R and applied statistics

Crawley, M.J., 2007. The R Book. John Wiley & Sons, Ltd., Chichester, England. ISBN: 978-0-470-51024-7, viii + 942 p.

Garrett, R.G. and 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.

Reimann, C., Filzmoser, P., Garrett, R.G. and Dutter, R., 2008. Statistical Data Analysis Explained: Applied Environmental Statistics with R. John Wiley & Sons, Ltd., Chichester, England. ISBN: 978-0-478-98581-6, 343 p.

Venables, W.N. and Ripley, D., 2002. Modern Applied Statistics with S (4th Edition). Springer-Verlag, Dordrecht, Germany. ISBN: 0-387-95457-0, xi + 495 p.