Utilizing open-source programming languages to statistically and spatially analyze regional-scale geoenvironmental datasets.

Shane Carey
Geological Survey of Ireland,
Beggars Bush,
Haddington Road,
Dublin 4, Ireland
shane.carey@gsi.ie

Mairead Glennon
Geological Survey of Ireland,
Beggars Bush,
Haddington Road,
Dublin 4, Ireland
mairead.glennon@gsi.ie.ie

Ray Scanlon
Geological Survey of Ireland,
Beggars Bush,
Haddington Road,
Dublin 4, Ireland
Ray.Scanlon@gsi

Kate Knights
Geological Survey of Ireland,
Beggars Bush,
Haddington Road,
Dublin 4, Ireland
kate.knights@gsi.ie

Abstract

Geoenvironmental surveying and mapping provide a wealth of data for the assessment of natural resources. Efficiently managing and analysing large geoenvironmental datasets can be a challenging task. However, making best use of open-source programming languages such as *R* and *Python*, allows for such datasets to be statistically and spatially analyzed in a more robust fashion. Using such tools allows for better day-to-day management of these geoenvironmental datasets and as a result, it allows for speedier and better decision making on how best to portray and map geoenvironmental datasets.

Keywords: GIS, Python, R, Geoenvironmental Datasets.

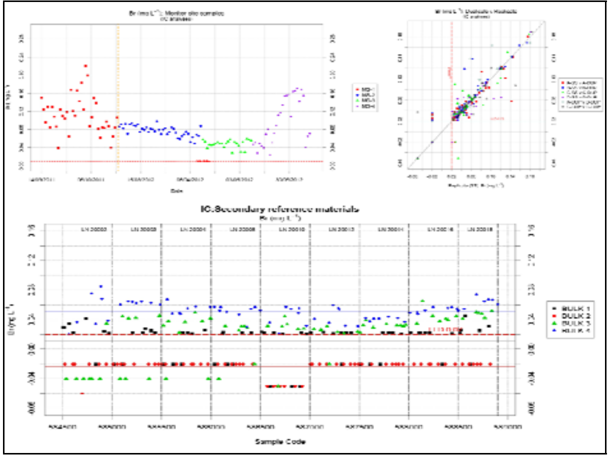
1 Introduction

The Tellus Border project is a European Union INTERREG IVA-funded mapping project that will collect and analyse geoenvironmental data on soils, waters and rocks across the border region of Ireland and integrate these with existing data in Northern Ireland. This cross-border collaboration between the Geological Survey of Ireland, the Geological Survey of Northern Ireland and research partners provides a wealth of data for the assessment of natural resources, sustainable environmental management and improvement of geological mapping on regional, national and cross-border scales. A multi-media geochemical survey was conducted in 2011 and 2012 resulting in *c.* 750,000 analytical results. Statistical analyses and mapping of these data require a high-level, interoperable programming environment to facilitate data management, manipulation and scientific interpretation. This paper and poster describes some statistical and mapping programming techniques developed for handling and interrogating large-scale multi-media geochemical datasets using the *R* and *Python* scripting languages along with GIS and geostatistical packages.

2 R – Statistical analyses

R [5] is a powerful open-source scripting language, widely used for data analysis, statistical computing and graphics. A range of statistical plots and tests (by media and by analyte) are carried out to determine if data are fit-for-purpose with regard to analytical specifications, mapping and multivariate analysis. *R* is employed (along with selected *R* published statistical packages[2,3]) initially to output a range of graphical plots for quality assurance and quality control assessment of analytical data with respect to laboratory reference materials (Figure 1).

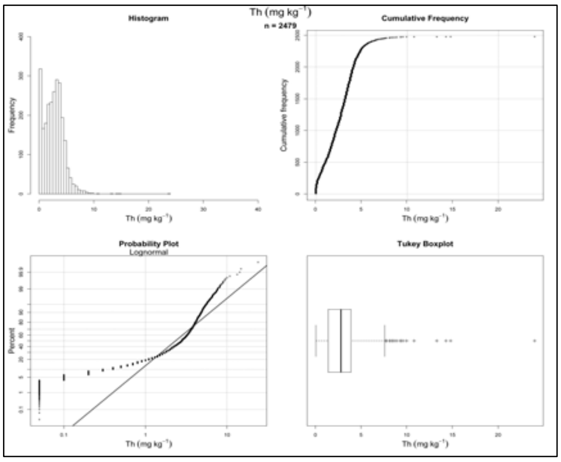
Figure 1 Graphical plots produced in R quality assurance and quality control assessment of analytical data



Source: Tellus Border project (In preparation). Topsoil Quality Control report.

Exploratory data analyses are carried out on the data to assess the data distribution, perform ranking calculations based on percentile classifications, and allow for descriptions of typical and anomalous values within the dataset. Examples of the graphical outputs from these analyses in R include box-and-whisker diagrams, histograms and cumulative frequency and probability distribution plots (Figure 2).

Figure 2 Example of an Exploratory Data Analysis graph produced in R (clockwise from top left: histogram, cumulative frequency plot, cumulative probability plot and Tukey box plot).



Source: Tellus Border project (In preparation). Topsoil Quality Control report

Further analyses on the basis of geological parent material and other variables such as land-use and sample media properties are carried out to provide insights for controls on element distributions in the environment. A random nested analysis of variance [4] is conducted within *R* to compute the proportions of analytical and sampling variance, and to satisfy the analysts of a sound and representative sampling strategy. The authors have adopted this approach from first principles, after [4] and validated against other mainstream statistical packages (e.g. Minitab).

2.1 Statistical and Spatial Correlations

Multivariate analytical techniques such as robust factor analyses and hierarchical cluster analyses are used to investigate statistical and spatial correlations between chemical elements, which might indicate a source or common process governing the spatial distribution of an anomaly. Furthermore, the StatDA package [3] usefully allows for mapping of factor scores, an approach to spatially investigate a limited number of factors which may relate to a source, geochemical process, or combination of these.

3 Mapping

Mapping geochemical data has traditionally been done in GIS software. For this project, an inverse distance weighted (IDW) spatial interpolation algorithm is used in ArcGIS [1] (ArcGIS Spatial Analyst extension) to interpolate between known point measurements, as a means of predicting a spatial data value between sampled points. However, the parameters for this, and the breaks for the colour ramps for a clear visual representation of the data, need to be robustly calculated outwith of the GIS program. A novel approach is to link the *R* and ArcGIS software by means of the *Python* programming environment [6] and using the arcPy mapping module developed by Esri. *R* and *Python* code have been developed to automate the process of exploratory data analysis, spatial data analysis, and data interpolation and map production. Given the large number of variables and results for a wide variety of sample media, the process of automating and live updating within a GIS is greatly beneficial to those working with and interpreting the spatial datasets.

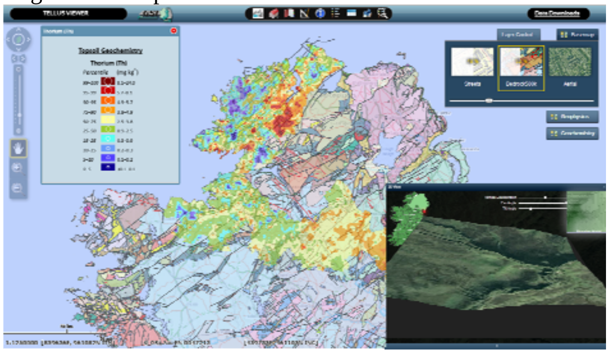
3.1 Large spatial datasets

Handling large spatial datasets comprising many variables is potentially cumbersome and complex. *R* provides the statistical tools to effectively carry out analysis of these large spatial datasets and to generate graphical summaries of these univariate analyses, while *Python* allows for the data to be spatially interrogated, managed and mapped. Coupling of *R* and *Python* is a relatively new technique; however, linking these two programming environments together has allowed the authors to greatly improve the overall data management and interoperability of these datasets, both statistically and geographically.

4 Online Viewer

Finally, ongoing programming is developing a web mapping service and online viewer for these mapped datasets, with live links to a managed database. It is envisaged that user-friendly tools will include Python and R-supported features, such as the ability to select and statistically summarise bespoke local data selections, and to generate a user-orientated report including georeferenced .png and layered .pdf documents. It is intended that this geoenvironmental information and data will underpin a broad range of research and practical applications, and be readily and easily accessible by end users.

Figure 3 Example of Tellus Border Online viewer



Source: http://spatial.dcenr.gov.ie/GeologicalSurvey/TellusBorder/index.html

5 Conclusion

Making best use of open-source programming languages such as *R* and *Python* allows for regional-scale geoenvironmental datasets to be analyzed in a more robust environment, both spatially and statistically. Ultimately, conducting analysis using such programming languages allows for better decisions to be made about the data and allows for datasets to be more easily managed.

References

[1] ESRI (Environmental Systems Resource Institute). ArcGIS 10.0. ESRI, Redlands, California, 2012.

[2] HP. Suter. xlsReadWrite: Read and write Excel files (.xls), 2011. URL: http://CRAN.R-project.org/package=xlsReadWrite, R package version 1.5.4.

[3] P. Filzmoser, B. Steiger. StatDA: Statistical Analysis for Environmental Data. 2012. URL: *http://CRAN.R-project.org/package=StatDA*. R package version 1.6.2.

[4] R.J. Howarth, Vol.2: Statistics and Data Analysis in Geochemical Prospectin, *In Handbook of Exploration Geochemistry*, pages 69-73, Elsevier, Amsterdam, 1983.

[5] R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL: http://www.R-project.org/ ISBN 3-900051-07-0.

[6] XQ. Xia, M. McClelland, Y. Wang. PypeR, A Python Package for Using R in Python. *Journal of Statistical Software, Code Snippets,* 35(2):1-8, 2010