Using the R language for graphical presentation and interpretation of compositional data in mineralogy – introducing the package *GCDkit-Mineral*

V. JANOUŠEK[1, 2], V. ERBAN[2], C. M. FARROW[3]

[1] *Institute of Petrology and Structural Geology, Charles University, Albertov 6, 128 43 Prague 2, Czech Republic; janousek@cgu.cz*

[2] *Czech Geological Survey, Klárov 3, 118 21 Prague 1, Czech Republic; erban@cgu.cz*

[3] *Computing Service, University of Glasgow, Glasgow G12 8QQ, Scotland; c.farrow@compserv.gla.ac.uk*

One of the problems we are facing in mineralogy is the dearth of universal, flexible and inexpensive software for recalculation of large compositional data sets acquired by microbeam techniques. The aim is to recast the chemical analyses from individual grains of various minerals into numbers of atoms per structural formula unit, and classify the individual data points according to rather complex rules based mostly on binary or ternary diagrams. The problem is that the recalculation and classification schemes differ strikingly for each of the main mineral species (IMA 2006).

We have decided to tackle the problem using the R language, as it provides powerful tools for data import/export, handling data matrices and production of publication quality graphics. An elegant solution to the computational problem uses S4 classes (Chambers 1998) to define algorithms as methods for each of the mineral species separately.

The raw data are imported into the system from the clipboard, text files or using the *RODBC* package (Ripley 2005). Individual analyses are split into classes according to the mineral species they belong to which enables the recalculation schemes to be defined, some of them rather complex, as independent S4 methods. There is a set of several generic functions that load the recalculation options for the given mineral class from two small and lucid external database files (ASCII and XML formats) that can be edited by users without prior R experience. Using these functions, the analyses are recast to structural formulae and the atoms are assigned to appropriate crystallographic sites. The minerals frequently form solid solutions and in these cases the data are transformed into a combination of two or more end-member compositions. Finally user-defined subsets of the numeric results can be copied to the clipboard, or exported via *RODBC* and *R2HTML* packages (Lecoutre 2003; Ripley 2005) to several formats (XLS, MDB, DBF and HTML). Special attention has been paid to provide routines for effortless data management, i.e. searching and generation of subsets, using regular expressions and Boolean logic.

In addition, our package for handling mineral compositions contains flexible high-level graphical functions. The diagrams are defined as internal templates that provide a means to create figure objects. The objects contain both the data and methods to make subsequent changes to the plot (zooming and scaling, adding comments or legend, identifying data points, altering the size or colour of the plotting symbols…). Most importantly, the templates are used as a basis for classification. Taking advantage of the algorithms originally developed for spatial data analysis (package sp of Pebsma & Bivand 2005), our general routine looks for the name of the polygon within the diagram (= graphical template), into which the analysis falls according to its x–y coordinates. The outcome can be either a name of the mineral or a link to another diagram, in the case of the more complex classification schemes. Following the compulsory rules of the International Mineralogical Association (IMA), in some cases the classification is not done graphically, but using external functions.

The package, named *GCDkit-Mineral (GCDkit* standing for *'*Geochemical Data Toolkit'), is a part of a broader family of tools designed for mineralogists and igneous geochemists. The overall philosophy each of the packages is to a large extent similar. All their functions are accessible via graphical user interface, as well as in an interactive regime for R-literate colleagues. The core of the system, e.g. the data input/output, data management and graphical functions is identical in each case. The *GCDkit* family tools are distributed as freeware via the WWW; the current version can be downloaded from *http://www.gla.ac.uk/gcdkit.*

Chambers, J. M. (1998). *Programming with Data*. New York: Springer.
IMA (2006). Accessed on February 28, 2006 at http://www.minsocam.org/MSA/IMA/.
Lecoutre, E. (2003). The R2HTML package. *R-news* **3 ,** 33–36.
Pebsma, E. J. & Bivand, R. S. (2005). Classes and methods for spatial data in R. *R-news* **5,** 9–13.
Originally Michael Lapsley and since Oct 2002 B. D. Ripley (2005). RODBC: ODBC database access. R package version 1.1-4.