Implementing spatial data analysis software tools in R[*]

Roger Bivand
Economic Geography Section, Department of Economics,
Norwegian School of Economics and Business Administration, Bergen, Norway
Roger.Bivand@nhh.no

15th February 2002

1 Introduction

This contribution has two equal threads: doing spatial data analysis in the R project and environment, and learning from the R project about how an analytic and infrastructural open source community has achieved critical mass to enable mutually beneficial sharing of knowledge and tools. The challenge is to see whether, and if so how far, we can contribute to the next meeting of the community nurturing R (and other projects) at the Distributed Statistical Computing workshop in 2003. It is fair to say that the statistical and data analytic interests of the community are catholic, rigourous, and enthusiastic, and challenge the perceived barriers between commercial and open source software in the interests of better, more timely, and more professional analysis in the proper sense of the word.

R is an implementation of the S language, as is S-Plus, and often able to execute the same interpreted code; it was initially written by Ross Ihaka and Robert Gentleman (1996). R follows most of the Brown and Blue Books (Becker, Chambers and Wilks, 1988, Chambers and Hastie 1992), and also implements parts of the Green Book (Chambers, 1998). R is associated with the Omegahat project: it is here that much progress on inter-operation is being made, for instance embedding R in Perl, Python, Java, PostgreSQL or Gnumeric. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs out of the box on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux). It also compiles and runs on Windows 9x/NT/2000 and MacOS.

In this abstract, focus is on the second thread, with the first one accessible through the links from the text and the references to reviews elsewhere.

[*] Abstract of proposed contribution to CSISS specialist meeting on spatial data analysis software tools, Santa Barbara CA, 10-11 May 2002.

2 Spatial data analysis in R: status

One of the most effective and conscientious contributors to the R project is Brian Ripley, who is not only very familiar with spatial statistics as an academic statistician (Ripley, 1981 among other publications), but also contributed an early package to R for point pattern analysis and continuous surface analysis, associated with Venables and Ripley (1999 - third edition). Descriptions of some of the packages available are given in notes in R News (Ripley, 2001, Bivand, 2001b), while a more dated survey was made by Bivand and Gebhardt (2000).

At the time of writing, searching the R site for "spatial" yielded 381 hits. As Ripley (2001) comments, some of the hesitancy that was previously observable in contributions of new packages coming forward has been due to the existence of the S-Plus SpatialStats module: duplicating existing work (including GIS integration) has not seemed fruitful. Over recent months, a number of packages have been released on CRAN in all three areas of spatial data analysis (point patterns, continuous surfaces, and lattice data).

3 What R offers as a programming environment and a project

Above, CRAN (Comprehensive R Archive Network) and packages were mentioned. While R provides a rich language and environment for data analysis and visualization, it is also extendible, not just because the user can write new or customised interpreted functions, and dynamically load compiled C, Fortran or C++ code, but because the project provides tools for checking, building, archiving and distributed user-contributed modules known as packages (like CTAN and CPAN for instance). Each such package is required to document functions, to provide examples which should run without error if the package is correctly installed, and optionally supply test data sets (now including remote online test data sets).

This effectively reduces or removes the barrier between users (with a certain insight into the language and their own problem areas) and core developers, and seems to be a good example of the beneficial consequences of an open source development model. It has been important to maintain a certain conservatism, meaning that hard-won experience (and legacy C and Fortran code) is central, while experimentation continues in parallel, and in part in the Omegahat project. It is also worth stressing that the R project is an open community, with multiple commitments to varying data analysis communities, and a clear willingness to adapt within the possibilities offered by open source development, in particular through inter-operation with other visualization software, databases, languages, and so on (even including R as an Excel Addin).

4 Opportunities for advancing spatial data analysis in R

While a good deal is already going on, there are some clear gaps that need to be filled, over and above making more modern spatial data analysis tools and knowledge available. One is the wish that Ross Ihaka made after the last DSC meeting (at which I had talked about GIS integration, Bivand, 2001a) for mapping capability in R. There is some code around, including topology code, other libraries are available (particularly from Frank Warnerdam's work), and all the current spatial data analysis packages try to solve visualization problems in their own ways. GRASS is also moving to positions from which the use of vector libraries is likely to be possible, also GPL and written in C.

A further area is that of inter-operation, using XML and/or Green Book connections methods, or simple programs writing programs. This could also involve plugging R data computation services into other front ends, say R in PostGIS given that R can already be embedded (experimentally) in PostgreSQL. This is more speculative, but Omegahat seems to be progressing vigourously, and highlights inter-system interfaces. It would however build on any pre-existing spatial data analysis functions in R, which would become available in the environment within which R is embedded if so selected.

References

R. A. Becker, J. M. Chambers, and A. R. Wilks. 1998. *The New S Language*. Chapman & Hall, London.

R. S. Bivand. 2001a. R and geographical information systems, especially GRASS, Proceedings of the 2nd International Workshop on Distributed Statistical Computing, Technische Universität Wien, Vienna, Austria.

R. S. Bivand. 2001b. More on Spatial Data Analysis, *R News*, 1 (3) 13-17.

R. S. Bivand and A. Gebhardt. 2000. Implementing functions for spatial statistical analysis using the R language, *Journal of Geographical Systems*, 2 (3) 307-317.

J. M. Chambers. 1998. *Programming with Data*. Springer, New York.

J. M. Chambers and T. J. Hastie. 1992. *Statistical Models in S*. Chapman & Hall, London.

R. Ihaka and R. Gentleman. 1996. R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics*, 5, 299-314.

B. D. Ripley. 1981 *Spatial statistics.* Wiley, New York.

B. D. Ripley. 2001. Spatial Statistics in R, *R News*, 1 (2) 14-15.

W. N. Venables and B. D. Ripley. 1999 *Modern Applied Statistics with S-Plus.* Springer, New York (book website).