USING OPEN SOURCE SOFTWARE IN VITICULTURAL RESEARCH

O. Zecca
Institut Agricole Régional
Région La Rochère 1/A, Aosta, Italy
o.zecca@iaraosta.it

ABSTRACT

Many high quality Open Source scientific applications have been available for a long time.

Some of them have proved to be particularly useful for carrying out the usual activities involved in viticultural research projects, such as statistical analyses (including spatial analyses), GIS work, database management (possibly integrated with statistical and spatial analysis) and even "low-level" often highly time-consuming activities (e.g. repetitive task on text files).

A few essential applications regularly used by the author in agronomic and viticultural research during more than a decade are summarily presented. They have consistently made the successful accomplishment of the projects possible without having to rely on commercial software. The advantages and disadvantages of Open Source applications versus commercial software (with comparable features and quality) are discussed from a more general point of view.

KEYWORDS

FOSS Software – GRASS GIS – Open Source – R language – Scientific applications

INTRODUCTION

Open Source software can be defined as software whose source code is freely available. More precisely, Open Source programs must be distributed with a licence complying with Open Source Definition's criteria, as defined by the Open Source Initiative. Besides free distribution of the source code, this includes the possibility of third party redistribution without having to pay royalties or other fees as well as the possibility of modifying the code and redistributing derived versions under the same terms as the licence of the original software. The complete Open Source Definition can be found at Open Source Initiative website (http://opensource.org). The term Free and Open Source Software (FOSS) is often used to refer in a neutral way both to applications adopting the Open Source Definition and to those adhering to the more strict principles of Free Software, proposed in the mid 80s by the Free Software Foundation (http://www.fsf.org).

Nowadays many hundreds of FOSS applications are available, from whole operating systems to small, simple utilities. A considerable number of high-quality applications are constantly being developed in any scientific discipline, including areas which may interest to viticultural researchers (especially those involved in complex viticultural zoning or other terroir-related projects) as advanced statistical methods, spatial data analysis, geographic information systems, remote sensing, large dataset management.

An analytic description of each of these applications, or a comprehensive comparison of each feature with those of the corresponding commercial alternatives is well beyond the scope of this work. Here, we will limit ourselves to a brief description of the main features of a few

outstanding applications and to point out their interest for agronomical research, mostly on the basis of our own experience in agronomical and viticultural research. Hopefully this could encourage others to get to know them better and directly explore their potential.

FOSS SOFTWARE

All the software described below is available for Windows OS as well as for UNIX systems like (among other) GNU/Linux and Mac OS X.

R

The software of choice for many professional statisticians, R (R Development Core Team, 2009) is also being adopted more and more by researchers involved in any area of biology and natural sciences. As it is described in its own home page, R is a language and environment for statistical computing and graphics. The language is very similar to the S language developed at Bell Laboratories (Becker *et al.*, 1988), of which it can be considered an Open Source implementation. While it is a very powerful application on its own, it deploys its full potential thanks to the available contributed extensions, called packages. At time of this writing (April 2010) there are 2313 contributed extension packages available for download and easy, automatic installation; they cover the whole range of modern statistics. The extensive packages collection can be browsed by topic in the CRAN Task Views (http://cran.r-project.org/web/views). Currently, 26 task views are available; researchers involved in viticultural research may be particularly interested in tasks such as Graphics, Environmetrics, Experimental design, Spatial.

Graphics (Graphic Displays, Dynamic Graphics, Graphic Devices, Visualization): besides the basic graphic system, R offers several packages for visualising data both statically and dynamically. Among them, an implementation of Cleveland's Trellis graphics system (Cleveland, 1993 and Cleveland, 1994) called *lattice* (Sarkar, 2008), and the *ggplot2* (Whickam, 2009), which applies the principles of the grammar of graphics (Wilkinson, Wills, 2005).

Environmetrics (Analysis of Ecological and Environmental Data): several packages are available expanding the modelling functions beyond the linear models: among them, *nlme* (Pinheiro, Bates, 2009) for mixed-effects models, *mgcv* (Wood, 2006) and *gam* (Hastie, Tibshirani, 1999) for Generalised Additive Models. Other packages provide functions for tree-based modelling, ordination and cluster analysis and environmental time series.

Experimental design (Design of Experiments and Analysis of Experimental Data): the *Agricola* package provides several functions for both planning and analysing agricultural and plant breeding experimental designs, as lattice designs, factorial designs, randomized complete block designs, completely randomized designs, (Graeco-)Latin square designs, balanced incomplete block designs and alpha designs.

Spatial (Analysis of Spatial Data): in addition to several functions for visualising and analysing spatial data included in the base R, many packages useful for dealing with geographical data are available (Bivand *et al.*, 2008). Most notable features include: (1) importing and exporting data in any of the several formats supported by the GDAL/OGR libraries; (2) importing and exporting ArcGis/ArcView shapefiles, as well as importing ArcInfo files; (3) interfacing with GRASS GIS (*spgrass6*), SAGA GIS (*RSAGA*) and Generic Mapping Tools (*GMT*) (see below for details on these applications); (4) model-based geostatistics, with *geoR* and *geoRgml* packages (Diggle, Ribeiro, 2007).

Perhaps the most striking evidence of the extraordinary potential of R is that some of the mainstream statistical software companies have started including an interface to R in their products: SPSS now offers access to R functions and even the possibility of adding R functions in the application menus, Statistica can produce output from R scripts in its proprietary formats. Even though Stata does not support directly R, a third party module to run R from inside Stata is available.

GRASS GIS (Geographic Resources Analysis Support System)

GRASS GIS (GRASS Development Team, 2010) is the oldest open source Geographical Information System: his development was originally started by U.S. Army Construction Engineering Research Laboratories (USA-CERL) in 1982. It currently provides a very wide range of functionalities in raster and vector GIS analysis, image processing, 2D and 3D visualisation (Neteler, Mitasova, 2009).

GRASS can be used in conjunction with R (from within the GRASS environment) for carrying out advanced spatial analyses.

QGIS (Quantum GIS)

QGIS can be used for visualising and exploring geographical data (including WMS and WFS data), creating maps, importing and exporting GPS data. It aims to provide an easy to use environment, with a user friendly graphical user interface (GUI). It can also be used as an easy interface to some GRASS functions, like editing and digitisation of GRASS vector maps.

SAGA GIS (System for Automated Geoscientific Analyses)

This is a GIS system originally developed by the Department of Physical Geography, University of Göttingen, Germany (Conrad, 2006). It has two interfaces: a user friendly GUI and a command line interpreter. SAGA functions are accessible from within the R environment thanks to the *RSAGA* package.

GMT (Generic Mapping Tools)

While GMT (Wessel, Smith, 1998) has arguably the most obscure command line interface, it can produce publication-quality maps. A more user friendly application, iGMT, provides a GUI to GMT.

GDAL/OGR libraries

These open source libraries provide the primary data access engine for GRASS, QGIS, GMT, SAGA, R, and many others (including closed source, commercial applications like Google Earth and ArcGIS).

Other GIS applications

Many other high quality FOSS applications are available: from comprehensive GIS systems, to software specialising in remote sensing or in web mapping. See OSGeo site for further information (www.osgeo.org).

Data Management

When statistical and GIS analysis is done on really large data sets, data must be kept in databases. Among the most popular open source database management systems are MySQL and PostgreSQL. R and GRASS provide interfacing capabilities to both, through specific drivers and more generic ODBC drivers (which also allow integration with many commercial RDBMS like MS Access, MS SQL Server, Oracle). PostGIS, an extension of PostgreSQL, adds support for geographic objects allowing it be used as a spatial database.

Managing large datasets coming from text files output by loggers of devices like meteorological stations or the many instruments used in agro-ecological research, can be a

highly time-consuming activity (particularly when data are taken simultaneously in several locations, like in viticultural zoning projects). The use of open source languages like Perl or Python, or even simple tools traditionally included in any UNIX station like the *stream editor* (SED) and the AWK language, has proven to be extremely useful, allowing pre-formatting of text data quickly and efficiently (i.e. minimising the possibilities of errors). R can also serve as a powerful tool for data management (Spector P., 2008).

DISCUSSION

The applications presented above (and several others) are without any doubt perfectly apt to be adopted as standard tools in agronomical research projects; even a completely FOSS-based workflow can be easily implemented. However, this does not necessarily imply that such a workflow would be the best strategy: a thorough comparison between open source software and available commercial counterparts should be carried out, focusing on key items like costs, features, ease of use, support policies.

Costs

Often the first argument in favour of FOSS alternatives is their cost. Mainstream commercial statistical or GIS packages are very expensive: their prices vary from a few to several thousands Euros, depending on which version is bought (only basic versions or also supplementary packages); of course, future costs of upgrades must also be taken into account. On the other hand, all open source applications presented here are available for free, making the choice between commercial and FOSS alternatives apparently obvious, at least from a purely economic point of view. Nevertheless, this does not mean that no explicit or implicit costs are to be expected when an open source strategy is chosen. Most of the FOSS applications considered are difficult to use and lack a graphical interface or only have a rudimentary one; as a consequence, some investment in terms of training, books and, above all, time are likely to be needed in order to achieve adequate proficiency. Thus, a more realistic comparison should include not only the costs of purchasing and maintaining up-to-date systems but also those of learning to wholly exploit the applications' power. Finally, time needed to accomplish the work is also to be taken into account; should the adoption of particularly difficult to use FOSS software result in a significant slowdown of work, purchasing high priced commercial applications may be an economically sound option (actually, in the author's experience, the opposite is much more likely, since, thanks to the high degree of automation allowed by the command line, working times are often dramatically reduced).

An added value of software being free, is that upgrades can be done regularly, as soon as they are made available (often several times per year), which means that it is always possible to work with the most up-to-date tools.

Capabilities

Comparing the overall capabilities of specialised Open Source Applications with their commercial counterparts is not always straightforward, due to differences in covered features (i.e. one or more FOSS applications can provide the same functionalities of one or more commercial counterparts, but the capabilities of each single application do not overlap precisely with any other's). However, the best FOSS applications can certainly suit any need in agronomical research and without any doubt they are, at least, on a par with their commercial counterparts in terms of performance and power. There are, of course, perfectly

adequate commercial alternatives; nevertheless, in order to satisfy all the requirements of a complex research project, often several applications and/or supplementary add-ons would be needed, resulting in a considerable increase of costs.

Ease of use

The learning curve is often considered steeper for FOSS software compared to commercial competitors. While commercial products are usually considered easier, thanks to the availability of comprehensive GUIs, unleashing the entire power of these kinds of applications usually involves the adoption of command line or script alternatives, which make them as difficult to use as their free counterparts. Furthermore, the perceived steep learning curve of the considered FOSS applications may partly be explained by lack of habit in using a command line interface. In the author's experience, learning to use mainstream applications like SPSS or ArcGIS after years of working with R and GRASS was not easy either.

Support

Mainstream software companies offer various kinds of direct support and training, which is often regarded as a strong advantage over the FOSS competition. In fact, while open source projects cannot always offer direct support services, there are usually many opportunities to get expert advice. Comprehensive documentation (tutorials, user guides, wikis) is always available at the project's home page, as well as in sites maintained by contributors or other expert users. Mailing lists are another extraordinary source of support: getting very competent answers to any question it's more than likely, even directly from the project's developers and in a very short time. Finally, there are specialised commercial companies offering support for the most widespread open source products (like R or GRASS).

Main advantages and disadvantages of an open source approach are summarised in Tab. 1.

CONCLUSIONS

A few FOSS applications which can be used to carry out some of the most challenging tasks usually faced in complex terroir-related viticultural research projects, such as large data management, advanced statistical analysis, management of geographical data and spatial data statistical analysis, have been presented. This software can replace commercial alternatives without reducing the quality of the final result.

On the contrary, there are chances that the adoption of an open source strategy may result in an improved workflow, reduced working times and better quality of the research, thanks to the integration between applications, the automation of repetitive tasks, the availability of always up-to-date software. The FOSS software's undeniable advantage of being free must not be underestimated, especially in view of the high costs of the commercial alternatives; still, the free availability of FOSS software should only be regarded as an added value, not the main reason for its adoption. Of course there are a few drawbacks, first of all the steep learning curve; however, after overcoming the initial difficulties, the adoption of a FOSS strategy will probably be greatly rewarding and the previously-used software will never be regretted.

Tab. 1: Advantages and disadvantages of an open source approach in scientific research.

	Capabilities	Costs	Learning curve	Support
Open Source Software	The best FOSS applications can certainly suit any need in scientific research and are (at least) on a par with their commercial counterparts in terms of performance and power.	In order to achieve adequate proficiency, some investment in terms of training, books and time may be needed. The free upgrades always allow working with up-to-date tools.	The learning curve is often considered steeper compared to commercial competitors. Nevertheless this may partly be due to lack of habit in using a command line interface.	Many opportunities to get comprehensive documentation and direct free support are available (tutorials, wikis, user guides, mailing lists). In some cases, commercial consultancy is also possible.
Commercial counterparts	Advanced statistical analysis and data exploration, spatial analysis, GIS analysis often require the purchase of several different applications and/or supplementary add-ons.	Initial investments are usually very high. Furthermore, regular purchases of upgrades must also be taken into account.	While commercial software is usually considered easier, thanks to the availability of complete GUI environments, exploiting the most advanced features usually involves the adoption of command line or script alternatives.	Mainstream commercial software companies offer various kinds of direct support and training.

REFERENCES

Becker R.A., Chambers J.M., Wilks A.R., 1988. The new S language: a programming environment for data analysis and graphics. Chapman & Hall, London.

Bivand R.S., Pebesma E.J., Gómez-Rubio V., 2008. Applied spatial data analysis with R. Springer, New York.

Cleveland W.S., 1993. Visualizing data. At&T Bell Laboratories. Hobart Press, Murray Hill, N.J.

Cleveland W.S., 1994. The elements of graphing data. Hobart Press, Murray Hill, N.J.

Conrad O., 2006. SAGA – program structure and current state of implementation. In: Böhner J., McCloy, K.R., Strobl, J. [Eds.]. SAGA – Analysis and Modelling Applications. Göttinger Geographische Abhandlungen, Vol.115.

Diggle P.J., Ribeiro P.J., 2007. Model-based geostatistics. Springer, New York.

GRASS Development Team, 2010. Geographic Resources Analysis Support System (GRASS) Software, Version 6.4.0. Open Source Geospatial Foundation.

Hastie T., Tibshirani R., 1999. Generalized additive models, Chapman & Hall/CRC, Boca Raton, Fla.

Neteler M., Mitasova H., 2009. Open Source GIS: A GRASS GIS Approach. Springer, New York.

Pinheiro J., Bates D., 2009. Mixed-Effects Models in S and S-PLUS. Springer, New York.

R Development Core Team, 2009. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Sarkar D., 2008. Lattice: multivariate data visualization with R. Springer, New York.

Spector P., 2008. Data manipulation with R. Springer, New York.

Wessel P., Smith W.H.F., 1998. New, improved version of generic mapping tools released. EOS Transactions, 79, p. 579.

Wickham,H., 2009. Ggplot2: elegant graphics for data analysis. Springer, New York.

Wilkinson L., Wills G., 2005. The grammar of graphics. Springer, New York.

Wood S.N., 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall/CRC, London.

USEFUL LINKS

R related links

R Project: http://www.r-project.org

R Graph Gallery: http://addictedtor.free.fr/graphiques/thumbs.php

R Journal: http://journal.r-project.org/current.html

GIS related links

GDAL/OGR libraries: http://www.gdal.org

Generic Mapping Tools: http://gmt.soest.hawaii.edu

GRASS GIS: http://grass.osgeo.org

OSGeo (The Open Source Geospatial Foundation): http://www.osgeo.org

Postgis: http://www.postgis.org

Quantum GIS: http://qgis.org

SAGA GIS: http://www.saga-gis.org