RAPSIM, a statistical toolbox for analysing and visualising large simulation data sets

Mario D'Antuono[1], James Fisher[2] and Doug Abrecht[3]

Department of Agriculture and Food Western Australia, www.agric.wa.gov.au/biometrics
[1] Crop Research Institute, Locked bag 5, Bentley, WA 6151 Email; mdantuono@agric.wa.gov.au
[2] Centre for Cropping Systems, PO Box 483 Northam, WA 6401. Email jsfisher@agric.wa.gov.au
[3] Dryland Research Institute, P0 Box 432 Merredin, WA 6415. Email dabrecht@agric.wa.gov.au

Abstract

Crop simulation models have the capacity to undertake extensive factorial pseudo-experiments over many years, generating large data sets [around 0.5 to 1 Gigabytes] of crop performance such as those available in "National Whopper Cropper" and "WA Wheat" derived from the APSIM crop simulation model.

The analysis of such datasets is beyond the capacity of most commercial spreadsheet or standard statistical programs and generally proceeds by reducing the dataset to fit the capacity of the analytical program.

RAPSIM is a collection of statistical functions which provide an easy approach to the analysis and presentation of large, structured data sets using the language of the R Statistical System. It has also been integrated as a menu within the R Statistical System on a PC computer running Windows XP.

We show that the analysis and visualisation of simulated data from the large "WA Wheat" data set of crop performance using RAPSIM provides a structured approach to summarising and understanding the key sources of variability in the datasets generated by APSIM models.

Key Words

R Statistical System, Windows XP, menu, graphics, analysis

Introduction

The statistical analysis of large datasets poses problems for small computers running Windows XP with a moderate amount of RAM (1-2 Gbytes). The coordination of the data management together with a statistical and graphical analysis of such large datasets has been tackled with the use of the R Statistical System and its libraries of functions. Special functions have been written in the R language to construct a framework called "RAPSIM" for handling large data files generated from APSIM models.

These functions have been made available in an easy to use menu also developed in the R Statistical System using the library called Rcmdr.

Methods

Statistical software

The R statistical system has become very popular as a scripting language used by statisticians worldwide. The system is distributed under the Free Software Foundation's GNU General Public License and is based in Vienna, Austria. There is a wealth of support from many statisticians and computer programmers specialising in particular areas. The system is 'open source' and anyone can view the statistical functions constructed in a library.

Menus

Recently, menus have been added by various users to make the R system more accessible, allowing users to have an easy interface to functions written in the R language. We have chosen to use the Rcmdr library (Fox 2005) as a basis to add or link the "RAPSIM" functions that we have written. Full details can be obtained from our website at www.agric.wa.gov.au/biometrics.

Databases

We originally stored the APSIM outputs as a Microsoft Access database on a Windows XP computer but recently, APSIM outputs have been converted to the netCDF format, a standard format for storing large arrays of data (http://www.image.ucar.edu/GSP/Software/Netcdf/). We examined the Merredin subset of the "WA Wheat" simulation generated from APSIM model with a dimension 725,760 records x 19 variables. There were 9 factors used to generate these data and these data have been described in Abrecht *et al* (2004) and further at this meeting Abrecht *et al* (2006).

Results

Figure 1 shows the development of the Rapsim system and further illustrations of graphical and statistical summaries will be presented in the poster at this meeting.
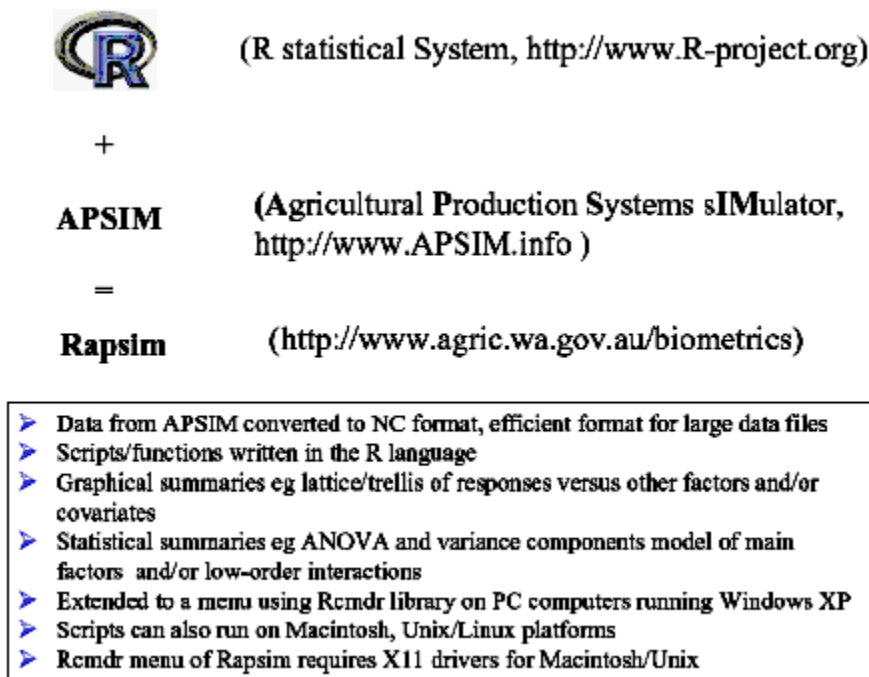

Figure 1. The RAPSIM development

Conclusion

The main features of the RAPSIM are that it is a collection of scripts or programs using the language of the R statistical system. It has been made available in a menu system on computers running Windows XP to assist researchers in developing a structured approach to summarising the large data sets from simulation models such as APSIM. We believe it gives the researcher a sense of purpose by having a structured approach in a proper statistical and graphical framework. RAPSIM scripts can also be run on other environments such as Macintosh and Unix/Linux which also support the R statistical system, but the menu system described above requires the support the X11 drivers to handle Rcmdr for these other environments.

References

Abrecht D, Fisher J and D'Antuono M. (2004) Visualising the yield space for wheat production in the eastern wheatbelt of Western Australia. 4th International Crop Science Congress, September 26 - October 1, Brisbane, Queensland, Australia.

Abrecht D, Fisher J and D'Antuono M. (2006) Assessing the impact of agronomy on wheat yield across soils and locations. 13th Australian Agronomy Conference, September 10-14, Perth, Western Australia.

APSIM Agricultural Production Systems sIMulator (http://www.APSIM.info)

Abrecht D, Fisher J and D'Antuono M. (2006) Assessing the impact of agronomy on wheat yield across soils and locations. Australian Agronomy Conference, September, Perth, Western Australia.

Fox J (2005). The R Commander: A Basic-Statistics Graphical User Interface to R. Journal of Statistical Software, volume 14, issue 9, 42pp. http://www.jstatsoft.org/ Accessed 16 June 2006.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org. Accessed 16 June 2006.