The State of Statistical Software

Jan de Leeuw
Department of Statistics
UCLA

http://gifi.stat.ucla.edu/pub/jsm03.pdf

0: Introduction

Making an inventory and classifying software necessarily reflects my bias, which is

- in academia,
- in research and development.

Thus, I am not a user, and I don't care (much) about sales. If I say "student", I tend to mean "graduate student".

JSS

This is despite the fact that I am also the editor of the Journal of Statistical Software, and that is probably why I am standing here.

http://www.jstatsoft.org

JSS is certainly not limited to academic contributions, but it is obvious that it has the same biases I have.

JSS also has a bias towards open source software, interpreted code, and R. I will argue that this is a good thing.

Classification

In a university there are three main types of activities

- Teaching, both graduate and undergraduate
- Research and development
- Service, including consulting

And, as a consequence, we'll discuss three classes of statistical software. Of course the distinctions are not always perfectly clear, but nevertheless they are useful.

Other criteria

- Web based (DHTML, JS, PHP, Java) - server side or client side.
- Interface: CLI or GUI or spreadsheet or notebook.
- Code Library (in scripting language or compiled language) or application.
- Extendable (with its own little language and/or with compiled code).
- Payware/freeware, open source

1: Research Software

The distinguishing characteristics of research software are

- Build-in interpreter for language (the language can be specific, but also C or Lisp).
- Support for graphics, ideally for dynamic graphics.
- Extensible with object code.
- Interface with most important graphics and database formats.
- Most importantly: used by statisticians for their research.

1.1: General Purpose Research

- There used to be two, now there is one. Although some people refuse to admit it, the XLISP-STAT system is dead. There is also XploRe, which has never really been alive (except locally). And then there are statisticians using Matlab for their research, mostly because they are close to engineering and CS, but this is a fairly small group.
- S/R is now the lingua franca of statistical research, and more specifically of computational statistics, and this is a good thing. We don't want the Tower of Babel, if people working together speak the same language, we gain efficiency. As a consequence we should discourage any development which goes against this grain.

Some nuisances

Nevertheless, not all is 100% well.

- R and S-plus are two dialects of the same language S, and the differences (at the system level) are considerable. This makes books like V&R harder to read than one would like.
- R and S are suffering under the burden of two different class systems. We should get rid of one, but nobody is really in charge.
- The R library (on CRAN) is not reviewed, and there is no coding standard.
- There is no dynamic graphics in R/S (yet).

Is R the way ?

- One of the major advantages of XLISP-STAT was that the language was Lisp. Thus: huge libraries, a byte-compiler, conceivably an object code compiler, a lot of documentation. In the same way, the major advantage of a system like CMAT (http://www.cmat.pair.com/cmat/) is that the language is C. Same for Root or Ch.
- One would have to weight the advantages/disadvantages of having the computational core of R as a Python or Perl module, versus a standalone system with a specialized little language. Now we have to continuously reinvent the wheel.
- But I guess it's too late. Worse-is-better wins again.

Is R the future ?

- R/S must be unified.
- R/S must be extended (byte-compiler, dynamic graphics, efficient handling of large datasets, adapt to modern CPU’s).
- Minimize the geek factor.
- Have a coding style.
- People must abandon other environments.
- A formal link to ASA/IMS/ISA would be helpful.

1.2: Specialized Research Software

Just some examples (here anarchy reigns) :

- http://www.mrc-bsu.cam.ac.uk/bugs/
- http://www.ioe.ac.uk/mlwin/
- http://members.ozemail.com.au/~kjbeath/glmstat.html
- http://www.mmisoftware.co.uk/pages/power.html
- http://www.causaScientia.org/software/Regress_plus.html

2: Teaching Software

There are, by now, may electronic textbooks, which give text together with examples, demos, and calculators.

- http://davidmlane.com/hyperstat/index.html
- http://www.statsoft.com/textbook/stathome.html
- http://cast.massey.ac.nz/
- http://www.stat.berkeley.edu/users/stark/SticiGui/index.htm
- http://statistics.cyberk.com/splash/
- http://www.anu.edu.au/nceph/surfstat/surfstat-home/surfstat.html

2.1: Undergraduate Software

- Clearly, teaching software should be WWW based. Moreover, there already is so much of it that we really do not really need *more*.
- One can make a strong case for Java Applets and Servlets.
- My guess is that the standard still is to use second-generation packages such as JMP, Datadesk, Vista, Stata. If you use those, you need a lab.
- In any case, you need a textbook to go with the software. Or vice versa.

2.2: Graduate Software

- It is obvious, from what we have seen before, that we should teach R to our graduate students, and they should use R in their research. There is no decent alternative.
- We should also teach SAS and/or SPSS and/or Stata to our students, but as something they should know about, certainly not to use it in their research.

3: Service Software

- Consulting software
- Calculators
- Data Analysis Tools
- Data Set Repositories
- Examples and Case Studies
- Topics Pages
- Jumpstations

3.1: Consulting Software

The classical packages (SAS and SPSS) are not really suitable for teaching and for research in statistics. One teaches SAS to students, but one does not use SAS for teaching statistics. At least one should not.

SAS and SPSS packages are intended to do mostly straightforward data analyses on large survey data sets. They do have their own matrix languages, but these languages are quite limited by the way the packages are organized (shades of the mainframe).

SPSS is a discipline for social scientists, SAS (and Excel, unfortunately) is a similar discipline for business. The discipline is related to statistics, but not identical to it.

2nd Generation Packages

- Packages such as Stata and JMP and Datadesk were fortunate enough to be born on desktops, and to have missed the mainframe period that has irreversably mutilated the design of SPSS and SAS.
- These newer packages were designed from the start with user interaction and with graphics output in mind (although obviously some implementations are better than others).
- Nevertheless in all these packages, and even in Vista, the user is still mostly passive and uses the GUI to start up bundled statistical techniques.

Let us be a bit more specific here, because this is an important point. Packages such as Vista and Datadesk have giving a lot of thought to interface design, and to emulating the process of actually doing a statistical analysis. It is unclear what the optimal target group for this approach is. One would think it is probably most suitable for teaching, and for client-level statistics.

These packages have scripting languages, in some cases added as an afterthought (JMP, Datadesk). Fine, but not for our students, and hopefully not for our colleagues. It is also significant, that SAS and SPSS and Datadesk now describe their software as “data-mining software”. So be it.

3.2: Calculator Pages

From an almost infinite universe:

- http://www.webstatsoftware.com/
- http://calculators.stat.ucla.edu
- http://members.aol.com/johnp71/javastat.html
- http://www-sci.lib.uci.edu/HSG/RefCalculators2A.html

It must be emphasized that this is an important service, and that we desperately need some quality control here. We need authorized calculators.

3.3: Data Analysis on the WWW

From what we have discussed before, it follows that two good ways to provide a data analysis service on the WWW are using Java and constructing interfaces to R, with the possibility to submit (sanitized) R code to the host server. Three examples are

- http://www.math.montana.edu/Rweb/
- http://www.webstatsoftware.com/
- http://www.stats.ox.ac.uk/~firth/CGIwithR/index.html

4: Omissions

I have not discussed

- Data sets (teaching)
- Examples and Case studies (teaching)
- Topic pages (research)
- Jumpstations (service)
- Preprint and library pages (service)
- E-journals (service)

because that is stretching the term “software” a bit too much for my taste. Or because I am running out of time. Obviously these areas are important and they often are interfacing with software we have discussed.