The Decision To Use R

A Consulting Business Perspective

by Marc Schwartz

Preface

The use of R has grown markedly during the past few years, which is a testament to the dedication and commitment of R Core, to the contributions of the growing useR base, and to the variety of disciplines in which R is being used. The recent useR! meeting in Vienna was another strong indicator of growth and was an outstanding success.

In addition to its use in academic areas, R is increasingly being used in non-academic environments, such as government agencies and various businesses, ranging from small firms to large multinational corporations, in a variety of industries.

In a sense, this diversification in the user base is a visible demonstration of the greater comfort that business-based decision makers have regarding the use of Open-Source operating systems, such as Linux, and applications, such as R. Whether these tools and associated support programs are purchased (such as commercial Linux distributions) or are obtained through online collaborative communities, businesses are increasingly recognizing the value of Open-Source software.

In this article, I will describe the use of R (and other Open-Source tools) in a healthcare-based consulting firm, and the value realized by us and, importantly, by our clients through the use of these tools in our company.

Some Background on the Company

MedAnalytics is a healthcare consulting firm located in a suburb of Minneapolis, Minnesota. The company was founded and legally incorporated in August of 2000 to provide a variety of analytic and related services to two principle constituencies. The first are healthcare providers (hospitals and physicians) engaged in quality improvement programs. These clients will typically have existing clinical (as opposed to administrative) databases and wish to use these to better understand the interactions between patient characteristics, clinical processes and outcomes. The second are drug and medical device companies engaged in late-phase and post-market clinical studies. For these clients, MedAnalytics provides services related to protocol and case report form design and development, interim and final analyses and independent data safety monitoring board appointments. In these cases, MedAnalytics will typically partner with healthcare technology and service companies to develop web-based electronic data capture, study management, and reporting. We have worked in many medical specialties including cardiac surgery, cardiology, ophthalmology, orthopaedics, gastroenterology and oncology.

Prior to founding MedAnalytics, I had 5 years of clinical experience in cardiac surgery and cardiology and over 12 years with a medical database software company, where I was principally responsible for the management and analysis of several large multi-site national and regional sub-specialty databases that were sponsored by professional medical societies. My duties included contributing to core dataset design, the generation of periodic aggregate and comparative analyses, the development of multi-variable risk models and contributing to peer-reviewed papers. In that position, the majority of my analyses were performed using SAS (Base, Stat and Graph), though S-PLUS (then from StatSci) was evaluated in the mid-90's and used for a period of time with some smaller datasets.

Evaluating The Tools of the Trade - A Value Based Equation

The founder of a small business must determine how he or she is going to finance initial purchases (of which there are many) and the ongoing operating costs of the company until such time as revenues equal (and, we hope, ultimately exceed) those costs. If financial decisions are not made wisely, a company can find itself rapidly "burning cash" and risking its viability. (It is for this reason that most small businesses fail within the first two or three years.)

Because there are market-driven thresholds of what prospective clients are willing to pay for services, the cost of the tools used to provide client services must be carefully considered. If your costs are too high, you will not recover them via client billings and (despite attitudes that prevailed during the dot.com bubble) you can't make up per-item losses simply by increasing business volume.

An important part of my company's infrastructure would be the applications software that I used. However, it would be foolish for me to base this decision solely on cost. I must consider the "value", which takes into account four key characteristics: cost, time, quality and client service. Like any business, mine would strive to minimize cost and time, while maximizing quality and client service. However, there is a constant tension among these factors and the key would be to find a reasonable balance among them while meeting (and, we hope, exceeding) the needs and expectations of our clients.

First and foremost, a business must present itself

and act in a responsible and professional manner and meet legal and regulatory obligations. There would be no value in the software if my clients and I could not trust the results I obtained.

In the case of professional tools, such as software, being used internally, one typically has many alternatives to consider. You must first be sure that the potential choices are validated against a list of required functional and quality benchmarks driven by the needs of your prospective clients. You must, of course, be sure that the choices fit within your operating budget. In the specific case of analytic software applications, the list to consider and the price points were quite varied. At the high end of the price range are applications like SAS and SPSS, which have become prohibitively expensive for small businesses that require commercial licenses. Despite my past lengthy experience with SAS, it would not have been financially practical to consider using it. A single-user desktop license for the three key components (Base, Stat and Graph) would have cost about (US)$5,000 per year, well beyond my budget.

Over a period of several months through late 2000, I evaluated demonstration versions of several statistical applications. At the time I was using Windows XP and thus considered Windows versions of any applications. I narrowed the list to S-PLUS and Stata. Because of my prior experience with both SAS and S-PLUS, I was comfortable with command line applications, as opposed to those with menu-driven "point and click" GUIs. Both S-PLUS and Stata would meet my functional requirements regarding analytic methods and graphics. Because both provided for programming within the language, I could reasonably expect to introduce some productivity enhancements and reduce my costs by more efficient use of my time.

"An Introduction To R"

About then (late 2000), I became aware of R through contacts with other users and through posts to various online forums. I had not yet decided on S-PLUS or Stata so I downloaded and installed R to gain some experience with it. This was circa R version 1.2.x. I began to experiment with R, reading the available online documentation, perusing posts to the r-help list and using some of the books on S-PLUS that I had previously purchased, most notably the first edition (1994) of Venables and Ripley's MASS.

Although I was not yet specifically aware of the R Community and the general culture of the GPL and Open-Source application development, the basic premise and philosophy was not foreign to me. I was quite comfortable with online support mechanisms, having used Usenet, online "knowledge bases" and other online community forums in the past. I had found that, more often that not, these provided more competent and expedient support than did the telephone support from a vendor. I also began to realize that the key factor in the success of my business would be superior client service, and not expensive proprietary technology. My experience and my needs helped me accept the basic notion of Open-Source development and online, community-based support.

From my earliest exposure to R, it was evident to me that this application had evolved substantially in a relatively short period of time (as it has continued to do) under the leadership of some of the best minds in the business. It also became rapidly evident that it would fit my functional requirements and my comfort level with it increased substantially over a period of months. Of course my prior experience with S-PLUS certainly helped me to learn R rapidly, (although there are important differences between S-PLUS and R, as described in the R FAQ).

I therefore postponed any purchase decisions regarding S-PLUS or Stata and made a commitment to using R.

I began to develop and implement various functions that I would need in providing certain analytic and reporting capabilities for clients. Over a period of several weeks, I created a package of functions for internal use that substantially reduced the time it takes to import and structure data from certain clients (typically provided as Excel workbooks or as ASCII files) and then generate a large number of tables and graphs. These types of reports typically use standardized datasets and are created reasonably frequently. Moreover, I could easily enhance and introduce new components to the reports as required by the changing needs of these clients. Over time, I introduced further efficiencies and was able to take a process that initially had taken several days, quite literally down to a few hours. As a result, I was able to spend less time on basic data manipulation and more time on the review, interpretation and description of key findings.

In this way, I could reduce the total time required for a standardized project, reducing my costs and also increasing the value to the client by being able to spend more time on the higher-value services to the client, instead of on internal processes. By enhancing the quality of the product and ultimately passing my cost reductions on to the client, I am able to increase substantially the value of my services.

A Continued Evolution

In my more than three years of using R, I have continued to implement other internal process changes. Most notably, I have moved day-to-day computing operations from Windows XP to Linux. Needless to say, R's support of multiple operating systems made

this transition essentially transparent for the application software. I also started using ESS as my primary working environment for R and enjoyed its higher level of support for R compared to the editor that I had used under Windows, which simply provided syntax highlighting.

My transformation to Linux occurred gradually during the later Red Hat (RH) 8.0 beta releases as I became more comfortable with Linux and with the process of replacing Windows functionality that had become "second nature" over my years of MS operating system use (dating back to the original IBM PC/DOS days of 1981). For example, I needed to replace my practice in R of generating WMF files to be incorporated into Word and Powerpoint documents. At first, I switched to generating EPS files for incorporation into OpenOffice.org's (OO.org) Writer and Impress documents, but now I primarily use LaTeX and the "seminar" slide creation package, frequently in conjunction with Sweave. Rather than using Powerpoint or Impress for presentations, I create landscape oriented PDF files and display them with Adobe Acrobat Reader, which supports a full screen presentation mode.

I also use OO.org's Writer for most documents that I need to exchange with clients. For the final version, I export a PDF file from Writer, knowing that clients will have access to Adobe's Reader. Because OO.org's Writer supports MS Word input formats, it provides for interchange with clients when jointly editable documents (or tracking of changes) are required. When clients send me Excel worksheets I use OO.org's Calc which supports Excel data formats (and, in fact, exports much better CSV files from these worksheets than does Excel). I have replaced Outlook with Ximian's Evolution for my e-mail, contact list and calendar management, including the coordination of electronic meeting requests with business partners and clients and, of course, synching my Sony Clie (which Sony recently announced will soon be deprecated in favor of smarter cell phones).

With respect to Linux, I am now using Fedora Core (FC) 2, having moved from RH 8.0 to RH 9 to FC 1. I have in time, replaced all Windows-based applications with Linux-based alternatives, with one exception and that is some accounting software that I need in order to exchange electronic files with my accountant. In time, I suspect that this final hurdle will be removed and with it, the need to use Windows at all. I should make it clear that this is not a goal for philosophic reasons, but for business reasons. I have found that the stability and functionality of the Linux environment has enabled me to engage in a variety of technical and business process related activities that would either be problematic or prohibitively expensive under Windows. This enables me to realize additional productivity gains that further reduce operating costs.

Giving Back

In closing I would like to make a few comments about the R Community.

I have felt fortunate to be a useR and I also have felt strongly that I should give back to the Community - recognizing that ultimately through MedAnalytics and the use of R, I am making a living using software that has been made available through voluntary efforts and which I did not need to purchase.

To that end, over the past three years, I have committed to contributing and giving back to the Community in several ways. First and foremost, by responding to queries on the r-help list and, later, on the r-devel list. There has been, and will continue to be, a continuous stream of new useRs who need basic assistance. As current userRs progress in experience, each of us can contribute to the Community by assisting new useRs. In this way a cycle develops in which the students evolve into teachers.

Secondly, in keeping with the philosophy that "useRs become developers", as I have found the need for particular features and have developed specific solutions, I have made them available via CRAN or the lists. Two functions in particular have been available in the "gregmisc" package since mid-2002, thanks to Greg Warnes. These are the `barplot2()` and `CrossTable()` functions. As time permits, I plan to enhance, most notably, the `CrossTable()` function with a variety of non-parametric correlation measures and will make those available when ready.

Thirdly, in a different area, I have committed to financially contributing back to the Community by supporting the R Foundation and the recent useR! 2004 meeting. This is distinct from volunteering my time and recognizes that, ultimately, even a voluntary Community such as this one has some operational costs to cover. When, in 2000-2001, I was contemplating the purchase of S-PLUS and Stata, these would have cost approximately (US)$2,500 and (US)$1,200 respectively. In effect, I have committed those dollars to supporting the R Foundation at the Benefactor level over the coming years.

I would challenge and urge other commercial useRs to contribute to the R Foundation at whatever level makes sense for your organization.

Thanks

It has been my pleasure and privilege to be a member of this Community over the past three years. It has been nothing short of phenomenal to experience the growth of this Community during that time.

I would like to thank Doug Bates for his kind invitation to contribute this article, the idea for which arose in a discussion we had during the recent useR! meeting in Vienna. I would also like to extend my thanks to R Core and to the R Community at large

for their continued drive to create and support this wonderful vehicle for international collaboration.

Marc Schwartz
MedAnalytics, Inc., Minneapolis, Minnesota, USA
MSchwartz@MedAnalytics.com

The ade4 package - I : One-table methods

by Daniel Chessel, Anne B Dufour and Jean Thioulouse

Introduction

This paper is a short summary of the main classes defined in the ade4 package for one table analysis methods (e.g., principal component analysis). Other papers will detail the classes defined in ade4 for two-tables coupling methods (such as canonical correspondence analysis, redundancy analysis, and coinertia analysis), for methods dealing with K-tables analysis (i.e., three-ways tables), and for graphical methods.

This package is a complete rewrite of the ADE4 software (Thioulouse et al. (1997), http://pbil.univ-lyon1.fr/ADE-4/) for the R environment. It contains **D**ata **A**nalysis functions to analyse **E**cological and **E**nvironmental data in the framework of **E**uclidean **E**xploratory methods, hence the name **ade4** (i.e., 4 is not a version number but means that there are four E in the acronym).

The ade4 package is available in CRAN, but it can also be used directly online, thanks to the Rweb system (http://pbil.univ-lyon1.fr/Rweb/). This possibility is being used to provide multivariate analysis services in the field of bioinformatics, particularly for sequence and genome structure analysis at the PBIL (http://pbil.univ-lyon1.fr/). An example of these services is the automated analysis of the codon usage of a set of DNA sequences by correspondence analysis (http://pbil.univ-lyon1.fr/mva/coa.php).

The duality diagram class

The basic tool in ade4 is the duality diagram Escoufier (1987). A duality diagram is simply a list that contains a triplet (**X**, **Q**, **D**):

- **X** is a table with n rows and p columns, considered as p points in $\mathbb{R}^n$ (column vectors) or n points in $\mathbb{R}^p$ (row vectors).

- **Q** is a $p \times p$ diagonal matrix containing the weights of the p columns of **X**, and used as a scalar product in $\mathbb{R}^p$ (**Q** is stored under the form of a vector of length p).

- **D** is a $n \times n$ diagonal matrix containing the weights of the n rows of **X**, and used as a scalar product in $\mathbb{R}^n$ (**D** is stored under the form of a vector of length n).

For example, if **X** is a table containing normalized quantitative variables, if **Q** is the identity matrix $\mathbf{I}_p$ and if **D** is equal to $\frac{1}{n}\mathbf{I}_n$, the triplet corresponds to a principal component analysis on correlation matrix (normed PCA). Each basic method corresponds to a particular triplet (see table 1), but more complex methods can also be represented by their duality diagram.

Functions	Analyses	Notes
dudi.pca	principal component	1
dudi.coa	correspondence	2
dudi.acm	multiple correspondence	3
dudi.fca	fuzzy correspondence	4
dudi.mix	analysis of a mixture of numeric and factors	5
dudi.nsc	non symmetric correspondence	6
dudi.dec	decentered correspondence	7

The dudi functions. 1: Principal component analysis, same as prcomp/princomp. 2: Correspondence analysis Greenacre (1984). 3: Multiple correspondence analysis Tenenhaus and Young (1985). 4: Fuzzy correspondence analysis Chevenet et al. (1994). 5: Analysis of a mixture of numeric variables and factors Hill and Smith (1976), Kiers (1994). 6: Non symmetric correspondence analysis Kroonenberg and Lombardo (1999). 7: Decentered correspondence analysis Dolédec et al. (1995).

The singular value decomposition of a triplet gives principal axes, principal components, and row and column coordinates, which are added to the triplet for later use.

We can use for example a well-known dataset from the base package :

```
> data(USArrests)
> pca1 <- dudi.pca(USArrests,scannf=FALSE,nf=3)
```

`scannf = FALSE` means that the number of principal components that will be used to compute row and column coordinates should not be asked interactively to the user, but taken as the value of argument `nf` (by default, `nf = 2`). Other parameters allow to choose between centered, normed or raw PCA (default is centered and normed), and to set arbitrary row and column weights. The `pca1` object is a duality diagram, i.e., a list made of several vectors and dataframes: