FESTSCHRIFT
50 JAHRE ÖSTERREICHISCHE STATISTISCHE GESELLSCHAFT
(SEITEN 61-70)

Vienna and R: Love, Marriage and the Future

Kurt Hornik* and Friedrich Leisch+

*Institut für Statistik,
Wirschaftsuniversität Wien, Austria
+Institut für Statistik und Wahrscheinlichkeitstheorie,
Technische Universität Wien, Austria

Abstract: Vienna is the home of many contributions to the R Project for statistical computing. In this paper we describe the current status of the R data analysis environment and give an overview of the historic development of the software with special emphasis on contributions by Austrian researchers. Finally we discuss the impact on teaching and research in Vienna and plans for the future of R.

1 Introduction

Since 1996, we have been involved with R, an environment for data analysis based on the award-winning S language. We have become members of the R Core development team, created and still maintain the Comprehensive R Archive Network, have made Vienna the venue of the leading conference related to R, and will soon assume key functions in the R Foundation, the new Vienna-based corporate entity which bundles the activities of the R Project for Statistical Computing. In this article, we describe how we got so deeply involved with R.

Any attempt to write down history is probably bound to be subjective, and this paper is no exception. It describes the development of R as seen by us, and others may have different views. R is the result of contributions by many volunteering individuals, a good share of which is based in Vienna.

2 What is R?

R is an interactive programming environment for data analysis, computational inference and visualization. It is highly extensible through user-defined functions and a fast-growing list of add-on packages. It is based on the award-winning S language which *"has forever altered the way how people analyze, visualize and manipulate data ..."* (Association of Computer Machinery Software System Award 1998 to John Chambers). It is open source released under the GPL ("GNU S") and freely available from `http://www.R-project.org`.

R is being developed for the Unix, Windows and Macintosh families of operating systems by an international development team ("R Core"), currently consisting of 16 individuals. New versions of R are made available twice per year (spring/fall).

An R distribution provides a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files, and contains functionality for a large number of statistical procedures. Among these are: linear and generalized linear models, nonlinear regression models, time series analysis, classical parametric and non-parametric tests, clustering and smoothing. The data analytic techniques described in such popular books as Venables and Ripley (2002), Pinheiro and Bates (2000), or Therneau and Grambsch (2000) have corresponding R packages (**MASS**, **nlme**, and **survival**). In addition, there are packages for bootstrapping, various state-of-the-art machine learning techniques, and spatial statistics including interactions with GIS. Other packages facilitate interaction with most commonly used relational databases, importing data from other statistical software, and dealing with XML.

R provides a rich resource of both low-level graphics functions and high-level visualization tools such as contour, mosaic, and condition plots. Low-level functionality includes drawing of lines, points and polygons with fine control of color and line types. Mathematical annotation in plot text is available via input expressions reminiscent of TEX constructs (Murrell and Ihaka, 2000). Users can create new visualization tools using the existing graphics components as building blocks. Production quality output can be rendered on-screen or in a variety of common formats, e.g., PostScript, PDF, PNG and JPEG.

Recent developments of R's graphics capabilities feature the introduction of two new packages called **grid** and **lattice**, which represent a move away from the traditional pen-and-paper rendering model. They involve a redesign of the graphics layout capabilities plus some support for interaction, and an implementation of Trellis graphics.

The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. Most of the user-visible functions in R are written in R. It is possible for the user to interface to procedures written in the C, C++, or FORTRAN languages for efficiency. Data structures include arrays, lists, and data frames (a construct designed to hold measurements in a cases by variables layout). Linear statistical models can be expressed compactly using formula notation. R supports an object-oriented paradigm which allows users to define their own classes and methods. R is a functional language with an evaluation model which is based on Scheme (Gentleman and Ihaka, 2000). In essence, everything in R is an object and hence users can easily perform computations on the language.

R is based on the S system which has been developed by John Chambers and colleagues at Bell Laboratories over the past 30 years. The books Becker et al. (1988), Chambers and Hastie (1992) and Chambers (1998)— also commonly referred to as the "Blue", "White" and "Green" books—describe key features of the S language. A commercial implementation of the S language called 'S-PLUS' is available from Insightful Corporation (`http://www.insightful.com`). Many programs written in the S language can be used under both S-PLUS and R.

3 A (Subjective) History of R

3.1 Southern Hemisphere Origins

What now is the R project started 1992 in Auckland, New Zealand, as an experiment by Ross Ihaka and Robert Gentleman *"in trying to use the methods of LISP implementors to build a small testbed which could be used to trial some ideas on how a statistical environment might be built"* (Ihaka, 1998). The initial interpreter consisted only of about 1000 lines of C code, on top of which data structures supporting statistical work and the command line interface were added. That Ross and Robert decided to use an S-like syntax—which they were both very familiar with—has driven the direction that R development has taken more than anything else.

First binary copies of R were made available on Statlib (`http://lib.stat.cmu.edu`) in 1993. Martin Mächler of ETH Zürich was one of the early adopters and encouraged Ross and Robert to release R as free software under the terms of the GPL (Free Software Foundation, 1991), which happened in 1995. See Ihaka and Gentleman (1996) for a description of early R development, design considerations and achievements.

Interest in R grew, Ross and Robert got bug reports by e-mail and released updated versions of R from time-to-time. As universities in New Zealand charge their departments for the internet traffic they generate, Martin volunteered to use ETH servers for maintaining a mailing list to discuss R development: the `R-testers` list started operation in March 1996 and was replaced by the three lists `R-announce`, `R-help`, and `R-devel` in 1997. The latter three are actively maintained by Martin until today, and the number of emails per day has seen a steady increase ever since.

3.2 Pre-R Times at CI

The "Center for Computational Intelligence" (CI, `http://www.ci.tuwien.ac.at`) is a research group specializing in statistical computing and machine learning and is located at the Department of Statistics and Probability Theory of the Technische Universität Wien. It started out in the late 1980s as the "Vienna Center for Neural Networks" as a joint effort with colleagues from the Department of Applied Computer Science, also of the TU Wien. Initially, applications of neural networks to pattern recognition tasks such as speech, speaker, or handwritten character recognition were typically implemented in C (or extensions of C), as high-level interpreted environments for data analysis were not available on the hardware platforms employed (NeXTstations), or simply prohibitively expensive.

Soon after the first Linux distributions (starting with the long forgotten SLS, the "Soft Landing System") became available in the early 1990s, we started to systematically convert to a homogeneous Linux-only computing environment. This provided us with very cost-efficient access to computing power, but again meant that standard statistical software packages were unavailable to us (at least for local installations). Thus, we were rather enthusiastic when we first came across Octave (`http://www.octave.org`), already then an implementation of the MATLAB™ 3 de-facto language standard. With Octave, we could suddenly easily perform computational tasks that had previously required considerable

programming effort. When in 1994 we started a cooperation with the research group on Psychiatric Genetics of the University Hospital Vienna, we could easily implement the data analysis tools we needed in Octave. This lead to a rather complete basic statistics toolbox for Octave, which together with various other code, mostly for signal processing, has long been integrated in the Octave distribution. However, this approach had at least two serious limitations: first, as Octave is implemented in C++ and does not provide a standard interface to foreign functions in C or FORTRAN, we found it near to impossible to extend Octave by using already existing, high-quality statistical code from e.g. the APSTAT or TOMS archives. And second, the elegance of the approach to statistical modeling described in the White Book (Chambers and Hastie, 1992) made it clear that in the long run we wanted to do things "the S way". Even if we had added missing functionality (such as S-style lists) to Octave, we would still have been confronted with the limitations of the MATLAB style script language. But miracles *do* happen ...

3.3 CRAN and the R Core Development Team

In 1996, we learned about R, a "language not unlike S". It turned out that we could easily compile R on our Linux systems, and as R already then included linear and generalized linear modeling and the power of the S language, it seemed that heaven had come true. Well, almost. We soon found problems and started sending bug fixes to Ross Ihaka and Robert Gentleman, then exclusively in charge. The first bug report by Fritz Leisch is from 1996-10-23 on a cryptic error message concerning heap size specification in R version 0.12. Version 0.13, released on 1996-11-07, already contained a bug fix by Kurt Hornik. We also started porting code from the Statlib S Archive (`http://lib.stat.cmu.edu/S`), writing our own code, and making suggestions about possible enhancements of R. Eventually, and maybe impressed by the quality of our bug fixes and the depth of our suggestions, Ross and Robert decided that it would be much better to let us (and a few other early enthusiasts) do what was thought necessary, instead of having us tell them what they should or could do: they gave us write access to the R sources. This new development model, with a "core group" instead of just Ross, Robert and Martin developing R, was formally recognized with the release of R 0.60 on 1997-12-05: "R Core" then consisted of Peter Dalgaard, Robert Gentleman, Kurt Hornik, Ross Ihaka, Thomas Lumley, Friedrich Leisch, Martin Mächler, Paul Murrell, Heiner Schwarte and Luke Tierney. (It is interesting to note that this release shows *TU Wein* as our affiliation, although it took 2 more years until Grüner Veltliner and Blaufränkischer started to play a role in R development).

With the growth in the R user community, it soon became clear that a central repository for material related to R was needed. We made a vague first proposal in late 1996, and ended up starting CRAN, the "Comprehensive R Archive Network" (`http://CRAN.R-project.org`), ourselves: the first announcement dates back to 1997-03-26. CRAN is a collection of sites which carry identical material, consisting of the R distribution(s), the contributed extensions, documentation for R, and binaries (including all leading Linux distributions, Windows, MacOS Classic and MacOS X). The CRAN master site at TU Wien currently has 12 official mirrors in all 5 continents. Due to this mirror system and the fact that R is included in all major Linux distributions, it is hard to assess interest in R by the number of hits on CRAN. In May 2002 (following the 1.5.0 release)

we had approximately 8000 completed downloads of the base system from the master server alone, which makes the machine one of the most popular web servers of the TU Wien (according to the university's computing center).

Maintaining CRAN means much more than making files available for download and writing fancy web pages. Perhaps the most important function of CRAN is to make contributed add-on packages (currently, 165 in the main section) available for download and installation. As this process should really be plug-and-play as much as possible, we have developed a packaging system which allows for installation and updates via the web, with minimal user interference. We also verify on a daily basis that the contributed package pass a collection of QA tests (which includes running the packages' own example code), for both the current release and development versions of R. (This system also makes it possible to identify incompatible code changes in R within a day.) We are currently working on providing finer-grained access to the wealth of contributed packages: in the future, it should be possible to query for required functionality and come up with the packages needed for meeting the requirements (plus everything these packages need in turn).

3.4 Vienna Congresses

The creation of R Core had resulted in a group of dedicated developers who went about their share of the work with only little coordination involved. This model has always been working surprisingly well. But it cannot serve for strategic decision making, in particular if there are 10 or 15 "board" members involved. A meeting was needed to discuss the future of R, including the roadmap for the 1.0 release, its relation to S, and possible synergies with other environments for statistical computing. (In particular as in 1998, John Chambers, the primary developer of S, had joined R Core and started the Omegahat Project for Statistical Computing (`http://www.omegahat.org`).) With the availability of funding by the Austrian Science Foundation (FWF) through the center of excellence program in "Adaptive Information Systems and Modelling in Economics and Management Science" (SFB 010, `http://www.wu-wien.ac.at/am/`), we were able to organize the first workshop on "Distributed Statistical Computing" (DSC 1999) in Vienna from March 19 to 23, 1999. For the first time, most of the R Core members actually met. The meeting was seminal to the further development of the R (and Omegahat) projects: without Vienna's wine and beer pubs, cafes and restaurants R and hence statistical computing would not be where it is today. Less than a year later, R 1.0 was released on 2000-02-29.

Motivated by the success of DSC 1999, we organized a second conference on "Distributed Statistical Computing" (DSC 2001) in Vienna from March 15 to 17, 2001, see Hornik and Leisch (2001). Again, this consisted of a regular conference track and meetings focusing on R development issues ("R Core and everyone else interested ..."). On the R Core side, the roadmap for R version 2, the "next generation of R" and most likely also to be considered the next generation of S, was laid out: R-ng is planned for the October 2003 release of R. For the regular conference, it turned out that most of the about 60 participants were primarily interested in R. This may be a consequence of the format, but seems indicative of a general problem in statistical computing: the lack of availability of free/open or generally usable statistical computing environments has led to a multiplicity of communities which typically do not share methods or tools. Given

the constraints on available development resources, it clearly seems imperative to try to locate synergy potentials across applications and realize them wherever possible: DSC 2003, again to be held in Vienna from March 19 to 21, 2003, will again be based on this attitude, even though the strong R aspect of DSC will make it an ideal candidate for yet-to-be-established (European only?) R Users' conferences.

3.5 The R Foundation for Statistical Computing

In August 2002, the "R Foundation for Statistical Computing" will officially be incorporated as a non-profit organization working in the public interest and seated in Vienna. The decision about the location, which makes Vienna the official capital of the R realm, was taken after considerable discussions, and appears a combination of practicalities, recognition of Vienna's impact on the R Project, and the fact that no one minds to come to Austria for a board meeting. The R Foundation will serve a role similar to that of other foundations underlying large open source projects. In particular, it will hold and administer the copyright of the R software and documentation and act as an official voice for the R Project, providing means of communication with the press, commercial and non-commercial organizations interested in R. It is very important to have such a platform, especially to coordinate funding for the further development of R, organize "official" R courses, and publish documentation related to R.

4 Plans for the Future

The forthcoming release of R version 1.6 already contains first steps towards a real name space mechanism, which is both a long-standing wish by many programmers as well as a pre-requisite for byte-compilation of R code. Other issues that will be addressed along the road towards R version 2 are multi-threading, event loops and changes to the graphics system. Another important aspect are S4-style classes and methods, which are currently contained in package **methods** and will be more closely integrated with the base system in the future. The **base** package itself will be reduced to a leaner core package containing only the language itself. Statistical functions currently contained in the base package will be moved to packages of their own and loaded on demand, e.g., linear and generalized linear models come to mind. Hook functions and auto-loading will help to hide implementation details like these from the end user.

The R Core team will collaboratively author a book accompanying the release of R version 2. The current plan is to present our view of modern statistical computing and building data mining tools using R as example and reference implementation, such that the book is of interest to a broader audience than the R user community. Topics will include principles of programming language design, interactive inter-operation with other languages and systems and high performance visualization and computation.

We are currently in the process of writing new tools for the administration of package repositories. These will allow

1. us to present the contributed packages on CRAN in a more structured way, such that users can browse the packages sorted by topic or priority.

2. others to setup their own repositories which are compatible with CRAN, such that the automatic download and updating mechanisms of R work. E.g., the Bioconductor project (`http://www.bioconductor.org`) already operates such a special topic repository of R packages.

In September 2002, gR 2002, a workshop on "graphical models in R", will be held in Vienna. This is the start of an initiative launched by Steffen Lauritzen (Aalborg) and Peter Dalgaard (Copenhagen) together with Kurt Hornik in response to the fact that whereas graphical models have now been around for a long time and shown to have a wide range of potential applications, software for graphical models is currently only available in a large number of specialized packages, such as BUGS, CoCo, MIM, TETRAD and others. The idea is to integrate these facilities in software of a more general nature, such as R, of course. Clearly, the success of such a project, which makes timely research available in an open source computing environment which is available on all major operating system platforms, will further add to make R a basic building block for computations in statistical research, and will significantly draw from available competencies in Vienna.

5 Impact on Research and Teaching

The involvement in the R Project of ourselves and many of our colleagues have certainly made Vienna one of the world's centers in modern statistical computing. Of course, there is more to this than being the CRAN master site, the venue of the leading conferences related to R, and the seat of the R Foundation. For one thing, a variety of important R add-on packages are Vienna-based. Many of these were written by graduate students at the Center for Computational Intelligence (CI), e.g.,

- Package **cclust** and the fuzzy clustering routines of package **e1071**, both by Evgenia Dimitriadou, provide newer clustering algorithms from the machine learning community in addition to classical statistical methods.

- Package **strucchange**, maintained by Achim Zeileis, provides a comprehensive toolbox for visualizing, testing and monitoring structural changes in linear regression models (Zeileis et al., 2002).

- Package **tseries** by Adrian Trapletti provides time series analysis methods used in computational finance.

- Thomas Baier's DCOM server allows R to be embedded in other Windows applications like Word or Excel (a client is currently under development).

The above list is not comprehensive, of course, and and only contains a few highlights. Other examples include David Maier's interface to `libsvm` in package **e1071**, created as the answer to a FAQ on the `R-help` mailing list: "Can R do support vector machines?". Andreas Weingessel has written considerable amounts of code used in the SFB on Adaptive Modelling. But not only CI members are in the long list of Viennese contributors to the R project, most notable is probably Erich Neuwirth's R-Excel interface.

The large base of high-quality add-on packages, together with the ease of writing extensions, be it in the R language itself or through integration of C or FORTRAN code or simply interfacing external programs, make R an ideal general-purpose tool for statistical applications and consulting. Our long-term research partners in psychiatry and business science have certainly profited from this.

In teaching, the S language has long had the image of being "for specialists (if at all) only" and to be deprecated in favor of GUI based systems (which could of course also be based on S). But on top of these considerations R has a unique advantage: it is *free and available on all systems*. Students can, at no extra cost and with virtually no effort, have the same environment for analyzing their data at home and at their university's computing labs. Those in need for flexible tools supporting quantitative decision making will certainly gain from using R.

Based on these considerations, several educational programs in Vienna are in the process of increasing their R content. At the University for Economics and Business Administration, we are starting to expose students to R as early as in their first-year compulsory introductory statistics course, either directly or via spreadsheet front-ends. Students interested in data analysis will use R in their additional undergraduate and graduate courses. This should give them a competitive edge in the international job market, both by being able to look at quantitative information in more flexible and intelligent ways, and by having timely access to state-of-the-art methodology developed or made available for R (such as e.g. tools for market segmentation).

At the University of Technology a new bachelors program on *Data Engineering and Statistics* has started in 2001. The program is part of the computer science programs and organized jointly with colleagues from the University of Vienna. R is the computational environment for almost all statistics classes of the curriculum, which is located at the interface of computer science and statistics. Graduates will be able to handle all aspects of data management and analysis, from collection and efficient storage in databases up to proper interpretation and presentation of results from a statistical analysis. Of course these students will also be programmers, forming a natural basis for the next generation of Viennese statistical computing researchers.

6 Summary

The strong involvement in the R Project makes Vienna one of the world's centers of modern statistical computing, which has impact for both research and teaching: educational and research programs integrating R gain a competitive advantage. In particular, students have easy access to state-of-the-art data analytic technologies, and researchers to the worldwide network of statistical computing experts within the R developers community both through long-term active cooperations and the workshops and conferences taking place in Vienna.

The city of Vienna itself provides the R project with something it as been good at for centuries: a place people love to come to for meetings. Without Vienna's wine and beer pubs, cafes, restaurants, and last but not least high-quality classical music performances, R would be nowhere near where it is today. As the master site of CRAN and the seat of

the R Foundation, Vienna could be seen as the "capital" of R, and as Doug Bates once noted: *"The sun never sets on the R empire . . . "*

Acknowledgements

Section 2 is based on material written by the authors in cooperation with Robert Gentleman for the CSDA statistical software newsletter. R would not be what it is today without the R development core team and the invaluable help of all those people from the R user community, who contributed by donating code, bug fixes and documentation. Run function `contributors()` in R for details.

References

Richard A. Becker, John M. Chambers, and Allan R. Wilks. *The New S Language*. Chapman & Hall, London, 1988.

John M. Chambers. *Programming with Data*. Springer, New York, 1998. URL `http://cm.bell-labs.com/cm/ms/departments/sia/Sbook/`. ISBN 0-387-98503-4.

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

Free Software Foundation. *Gnu general public license*. 675 Mass Ave, Cambridge, MA 02139, USA, 2 edition, 1991.

Robert Gentleman and Ross Ihaka. Lexical scope and statistical computing. *Journal of Computational and Graphical Statistics*, 9:491–508, 2000. URL `http://www.amstat.org/publications/jcgs/`.

Kurt Hornik and Friedrich Leisch, editors. *Proceedings of the 2nd International Workshop on Distributed Statistical Computing (DSC 2001)*, Technische Universität Wien, Vienna, Austria, 2001. URL `http://www.ci.tuwien.ac.at/Conferences/DSC.html`. ISSN 1609-395X.

Ross Ihaka. R: Past and future history. In S. Weisberg, editor, *Proceedings of the 30th Symposium on the Interface*, pages 392–396. The Interface Foundation of North America, 1998.

Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

Paul Murrell and Ross Ihaka. An approach to providing mathematical annotation in plots. *Journal of Computational and Graphical Statistics*, 9:582–599, 2000. URL `http://www.amstat.org/publications/jcgs/`.

Jose C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-Plus*. Springer, 2000. URL `http://nlme.stat.wisc.edu/MEMSS/`. ISBN 0-387-98957-0.

Terry M. Therneau and Patricia Grambsch. *Modeling Survival Data: Extending the Cox Model'*. Springer, 2000.

William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S. Fourth Edition*. Springer, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4/`. ISBN 0-387-95457-0.

Achim Zeileis, Friedrich Leisch, Kurt Hornik, and Christian Kleiber. `strucchange`: An R package for testing for structural change in linear regression models. *Journal of Statistical Software*, 7(2):1–38, 2002. URL `http://www.jstatsoft.org/v07/i02/`.

Authors' addresses:

Ao. Univ.-Prof. Dipl.-Ing. Dr. Kurt Hornik
Department of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstr. 7
A-1040 Vienna
Austria
Tel. +43 1 58801 / 10712
Fax +43 1 58801 / 10798
Email: Kurt.Hornik@ci.tuwien.ac.at

Dipl.-Ing. Dr. Friedrich Leisch
Department of Statistics and Probability Theory
Vienna University of Technology
Wiedner Hauptstr. 7
A-1040 Vienna
Austria
Tel. +43 1 58801 / 10715
Fax +43 1 58801 / 10798
Email: Friedrich.Leisch@ci.tuwien.ac.at