available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cosrev

Environment for statistical computing

Jaromír Antoch

Charles University, Department of Statistics, Sokolovská 83, CZ–186 75 Prague, Czech Republic

ARTICLE INFO

Article history:
Received 19 May 2008
Received in revised form
22 May 2008
Accepted 22 May 2008

ABSTRACT

This paper is a short exposition on the current state of art as far as statistical software is concerned. The main aims are to take a look at current tendencies in information technologies for statistics and data analysis, especially for describing selected programs and systems.

We start with statistical packages, i.e. a suite of computer programs that are specialized in statistical analysis, to enable people to obtain the results of standard statistical procedures without requiring low-level numerical programming, and to provide facilities of data management. A big surprise for many statisticians is that the most typical representative in this domain is Microsoft Excel. Aside from that, we touch upon a few commercial packages, a few general public license packages, and a few analysis packages with statistics add-ons.

An integrated environment for statistical computing and graphics is essential for developing and understanding new techniques in statistics. Such an environment must essentially be a programming language. Therefore, we take a closer look at several typical representatives of these types of programmes, and on a few general purpose languages with statistics libraries.

However, there exists quite a clear distinction between practical and theoretical approaches to most statistical work. The majority of software products for statistics are on the practical side, using numerical and graphical methods to provide the user access to existing methods. On the other hand, software packages specifically designed just for pure statistical–mathematical modelling do not exist. Nevertheless, all available computer algebra and/or mathematical systems offer tools for theoretical statistical work. Therefore, we take a look at some possibilities in this area.

Finally, we summarize several major driving forces that will influence, according to our strong belief, the statistical software development process in the near future. Due to limited space, these discussions are cursory in nature for the most part. This paper is based on the personal experience of the author as described in [J. Antoch, Series of papers on statistical software and environments for statistical computing (in Czech for the Czech Statistical Society Newsletter and other publications). [1]] and on the information available on Internet. Very good and interesting source of information is especially Google search machine [Google search machine. [12]], Wikipedia [Wikipedia, a multilingual web-based, free content encyclopedia project. [25]] and the journal Scientific Computing World [Scientific Computing World Journal. [22]].

E-mail address: jaromir.antoch@mff.cuni.cz.

1574-0137/$ - see front matter © 2008 Published by Elsevier Ltd
doi:10.1016/j.cosrev.2008.05.002

1. General thoughts

Computational statistics, or statistical computing, is the interface between statistics, computer science, and numerical analysis. It is the area of computational science (or scientific computing) specific to the mathematical science of statistics. The terms computational statistics and statistical computing are often used interchangeably, although Carlo Lauro, a former president of the International Association for Statistical Computing,[1] proposed in [17] during his Presidential statement during the opening of the Symposium COMPSTAT'96, to make a distinction between them. More precisely, he defined statistical computing as the application of computer science to statistics, and computational statistics as aiming at the design of algorithms for implementing statistical methods on computers, including the ones unthinkable before the computer age (e.g. bootstrap or simulations), as well as to cope with analytically intractable problems.

When speaking about an environment for statistical computing, it is necessary always keep in mind following points.

(1) Most important improvements in the statistical computing always profited much more from progress in "general computing" and hardware development than from progress in mathematical statistics.
(2) The environment for statistical computing should offer not only its statistical part but also to facilitate other tasks as storing and/or retrieval of data from internal and external sources, report writing as well as other publication activities, easy communication with colleagues and clients, etc.
(3) Systems should be open both as regards the changing interests of users and the growing spectrum of tasks to be solved.
(4) Last but not least, computers and operation systems are changing in regular cycles and it takes for a system more than one cycle to reach maturity. Unfortunately, after several cycles almost every program or system is swept away by progress in technical and knowledge levels.

Concerning the last point, the basic technical tendencies these days seem to be:

- Nanotechnology allowing an increase in the speed of both processors and memories and decrease of their size and power consumption.
- Architecture of the processors is typically 64 bit and processors with more than one kernel became a standard.
- RAM memory size starts at 4 GB.
- Prices of hardware "decrease" or, it is better to say, for the same amount of money end-users get more powerful computing facilities.

Most specialists predict two key trends for the computing, namely, the use of powerful ultramobile notebooks for "standard calculations" and the use of specialized grids intended for computationally intensive calculations, simulations etc. On the other hand, thanks to technical progress it appears that for more and more users of statistics, possession of a good scientific calculator will be enough. In these days these can be small computer specialized in scientific calculations, teaching and mobile calculations, or a PDA (personal digital assistant) equipped with Excel and basic statistical programmes, e.g. Statgraphics.[2]

2. Selected statistical packages

According to Wikipedia[3], a statistical package is a suite of computer programmes that are specialized for statistical analysis, enabling people to obtain the results of standard statistical procedures without requiring low-level numerical programming, and providing facilities for data management. In this section we will take a look at typical representatives in this domain. We will start with Microsoft Excel as the most typical representative of spreadsheets, selected commercial statistical packages and free and open statistical software. An interesting and fairly extensive list of statistical software can be found at[4], for a concise comparison of statistical software see[5]. Notice, however, that in our survey we do not touch either data mining nor artificial intelligence systems, being convinced that this type of the software deserves a separate paper.

2.1. *Microsoft Excel*

Despite the antipathy of many statisticians, it is without any doubt that Microsoft Excel is a much-used tool for statistical analysis of real data. Excel is available to many people as part of Microsoft Office and contains some statistical functions in its basic installation. It also comes with statistical routines in the Data Analysis Toolpak, an add-in found separately in the Office CD. Yet Excel is not in its essence a statistical tool. Among its limitations belong:

(1) Many statistical methods are not available.
(2) Several procedures are misleading.
(3) Distributions are not computed with enough precision.
(4) Routines for handling missing data were incorrect.
(5) Regression routines are incorrect for multicollinear data.
(6) Ranks of tied data are computed incorrectly.
(7) Many of Excel's charts violate standards of good graphics etc.

Therefore, third-party add-ins to Excel attempt to compensate for these limitations, adding new functionality to the programme. Comprehensive information can be found at[6].

Among the most reliable add-ins belong NAG's Statistical Add-Ins for Excel,[7] providing the functions covering one sample, matched pairs samples, two samples, K samples,

[1] http://www.stat.unipg.it/iasc/

[2] http://www.statgraphics.com/statgraphics_mobile.htm

[3] http://en.wikipedia.org/wiki/Statistical_software

[4] http://ourworld.compuserve.com/homepages/Rainer_Wuerlaender/statsoft.htm

[5] http://en.wikipedia.org/wiki/Comparison_of_statistical_packages

[6] http://dmoz.org/Computers/Software/Spreadsheets/EXCEL/Add-Ins/

[7] http://www.nag.com/statistical_software.asp

statistics, regression, correlation, time series, experimental design, and generalized linear models, No programming is required to use these statistical algorithms since they can be accessed via the function wizard and, as with standard Excel functions, the results are immediately updated whenever the input cells are altered.

Aside from that, NAG offers a School Excel Add-In developed to support instruction in statistics. In the UK, it became an important tool to aid the teaching of mathematics, statistics, information, and communication, and received approval from the UK government agency Curriculum Online[8] for use with eLearning Credits. In North America, it assists instructors of high school advanced placement and in introductory college-level to support instruction in statistics and decision analysis.

2.2. *Selected commercial statistical packages*

In this subsection we will touch upon a few commercial packages related to statistics and data analysis.

BMDP

BMDP[9] is a statistical package developed in 1961 at UCLA. Based on the older BIMED program for biomedical applications, it used keyword parameters in the input instead of fixed-format cards, so the letter P was added to the letters BMD, although the name was later defined as being an abbreviation for Biomedical Package. BMDP was originally distributed for free, in fact it was the first public software developed for statistics. It is now offered by Statistical Solutions together with nQuery Advisor, the industry standard for sample size and power calculations, and other products.

GenStat

GenStat[10] is a comprehensive statistics package that has been used in the most demanding real-life applications for over 30 years. GenStat was developed at the famous Rothamsted Experimental Station,[11] where many of the statistical techniques still in use today were first discovered. It contains a broad range of leading-edge statistical tools in an easy-to-use package, supported by its powerful and flexible high-level programming language. The current version is available for the Windows environment. Of interest for the developing world is GenStat Discovery Edition, a free version available to non-commercial users throughout Africa and a range of developing countries outside Africa. Examples of users who qualify are students and lecturers of universities and staff of government research organizations or non government organizations.

Genstat changed a lot recently both in what concerns interfaces and in diversification of data analysis. A range of procedures adds to GenStats generic and specialized statistical tools repertoire. One group, interesting not only in its own right but as an instance of the recent trend across this market to extend functioning by accepting and running external material, relates to Markov Chain Monte Carlo simulations. More precisely, GenStat runs WinBugs[12] (a standalone Bayesian program designed to make MCMC available for applied use), accepts its output in CODA[13] form (Convergence Diagnostic and Output Analysis generated by S-Plus routines), and produces plots from either source.

Generalized models, one of the cores of GenStat, gained recently two new procedures: hierarchical GLMs to nonlinear cases by inclusion of calculated variates, and generalized linear modelling of survey data. Surveys also gained a procedure forming a new bootstrap sample on each call from one or two stage stratified data. On top of the additions, there are modifications to a significant number of existing directives and procedures, most of which provide some level of added value to the function concerned. Graphic output is an area on which GenStat has in the past paid less emphasis than other systems, the priority being analysis, but the signs are that it is now a focus of development attention. New directives starting 'LP' are a clear sign of that.

JMP

JMP[14] (John Sall's Macintosh Project) is a computer program enabling simple statistical analyses to be performed. It dynamically links statistics with graphics to interactively explore, understand, and visualize data. This allows clicking on to any point in a graph, and see the corresponding data point highlighted in the data table, and other graphs. JMP provides a comprehensive set of statistical tools as well as the design of experiments and statistical quality control in a single package. It can work with a variety of data formats, such as text files, Microsoft Excel files, SAS data sets, and ODBC-compliant databases. Written originally in 1989 for the Mac, it was later released for Microsoft Windows in 1993 and Linux in 2005. JMP is distributed by SAS Inc.

NCSS

NCSS,[15] originally Number Cruncher Statistical System, is first and foremost an interactive exploratory environment. The most obvious manifestation is the Quick Launch Window, an optional pallette of buttons giving instant one-click access to anything the system can do and to the documentation section, too. The buttons are grouped into logical panels and carry a well-designed and rapidly-internalized system of icons.

Point and click is ideal for exploration, but becomes wearisome in repetition, so the tokenized macro system was a welcome extension for batch operations. A recorder provides the ease of use provided by third party Window macro utilities with none of the delays or debugging, and the result is a concise, editable command script that can be run directly from a command line, shortcut, Intranet html page, or assigned to buttons. Direct scripting offers greater control, for instance, in looping and assignable variables. Beneath the surface, there is a wide range of analytic capabilities comprising more than two hundred commands. Indeed, NCSS offers over 230 documented statistical and

[8] http://www.curriculumonline.gov.uk/

[9] http://www.statsol.ie/

[10] http://www.genstat.com/

[11] http://www.rothamsted.ac.uk/

[12] http://www.mrc-bsu.cam.ac.uk/bugs/

[13] http://www-fis.iarc.fr/coda/

[14] http://www.jmp.com/

[15] http://www.ncss.com/

plot procedures, imports and exports all major spreadsheet, database, and statistical file formats. NCSS is accompanied by PASS (Power Analysis and Sample Size) and GESS (Gene Expression Statistical System for Microarray data), to be purchased separately. Moreover, the data simulator is a useful module with a variety of distributions available.

NCSS is a well designed and worthwhile programme. Nonusers seeking a powerful and accessible interactive analysis environment have strong reasons to consider it. We are using NCSS for teaching non statistical students oriented to natural sciences. Based on this experience we can confirm that they analyze the data and accomplish complicated tasks with the help of NCSS quite easily and quickly.

OMNITAB 80

OMNITAB 80[16] is a high-level spreadsheet for statistical analysis, developed and maintained in the Statistical Engineering Division of NIST.[17] It performs many different statistical analysis, arithmetic and trigonometric calculations, and matrix and array operations. The software responds to simple instructions and uses reliable computational algorithms. Its descendent, Minitab,[18] is a statistical package often used for teaching.

SAS

The SAS[19] System is an integrated system of software products enabling the programmer to perform data entry, retrieval, management, and mining; report writing and graphics; statistical analysis; business planning, forecasting, and decision support; operations research and project management; quality improvement; applications development; data warehousing and platform independent and remote computing. In addition, the SAS System integrates with many SAS business solutions that enable large scale software solutions for human resource management, financial management, business intelligence, customer relationship management, and much more. Thus, it is not surprising that SAS is viewed as a company playing the role of Microsoft among statistical software companies.

The programme is composed of three major parts, the data step, procedure steps (effectively, everything that is not enclosed in a data step), and a macro language. SAS Library Engines and Remote Library Services allow access to the data stored in external data structures and on remote computer platforms. The data step section of a SAS program, like SQL or Focus, assumes a default file structure, and automates the process of identifying files to the operating system, opening the input file, reading the next record, opening the output file, writing the next record, and closing the files. This allows the user/programmer to concentrate on the details of working with the data within each record, in effect working almost entirely within an implicit program loop that runs for each record. All other tasks are accomplished by procedures that operate on the data set as a whole. There are macro programming extensions, that allow for rationalization of repetitive sections of the program. Proper imperative and procedural programming constructs can be simulated by use of the "open code" macros or the SAS/IML component.

Compared to general-purpose programming languages, this structure allows the user/programmer to be less familiar with technical details of the data and how they are stored, and relatively more familiar with the information contained in the data. The SAS System runs on IBM mainframes, Linux and Unix machines, OpenVMS Alpha, and Microsoft Windows.

SPSS

SPSS,[20] originally Statistical Package for the Social Sciences, was released in its first version in 1968. SPSS is among the most widely used programmes for statistical analysis in social sciences. Aside from that, it is often used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations and others. In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored with the data) are features of the base software. SPSS Programmability Extension allows users to extend the SPSS command syntax language with Python,[21] a remarkably powerful dynamic programming language, that is used in a wide variety of application domains.

Statgraphics

Statgraphics[22] is a line of computer programmes that perform and explain basic and advanced statistical functions. In the times of the MS DOS operating system it was probably the first statistical programme that automatically linked all implemented analysis with the graphical output. Worth of noticing is the recent release of Statgraphics, the first sophisticated statistical program that runs on hand-held computers (Pocket PC, Pocket PC Phone Edition, or compatible device running Windows Mobile 5 or Windows Pocket PC 2003).

Statistica

Statistica[23] belongs no doubt to the leaders in the domain of statistical computing. It is a suite of analytics software products which provides a comprehensive array of data analysis, data management, data visualization, and data mining procedures. It includes a wide selection of predictive modelling, clustering, classification, and exploratory techniques in one software platform. StatSoft, its producer, is also one of the largest manufacturers of enterprise-wide quality-control and improvement software systems in the world. These software products are used in mission critical manufacturing applications, in regulated FDA-controlled industries, and as a foundation of corporate-wide Six Sigma initiatives. StatSofts products are also used to help achieve compliance with CFR Part 11 and Sarbanes–Oxley regulations, among others.

Usability and integration with the larger working and communication environments have for a long time been a strength of Statistica. Integration with Microsoft Office includes opening Excel workbooks directly as a replacement for Statistica's own worksheet; outputting reports to Word

[16] http://www.itl.nist.gov/div898/software/omnitab.html

[17] http://www.nist.gov/

[18] http://www.minitab.com/

[19] http://www.sas.com/

[20] http://www.spss.com/

[21] http://www.python.org/

[22] http://www.statgraphics.com/

[23] http://www.statsoft.com/

format documents, though the ability to drag and drop report content (text, tabular or graphic) to a chosen point within any appropriate output handler (a different word processor, a desktop publisher, a spreadsheet, a graphics package) is a richer and more flexible alternative. For output to bitmap file, graphics can be saved at a user specified resolution. Locking has been introduced to avoid inadvertent changes to graphics or worksheets in a multiuser environment.

Extensive automation and centralization of the data handling are available with this system which, along with simplified dashboard style monitoring of complex multivariate processes, allow highly flexible graded information delivery with subtle command and control. With data analytic products an ever more vital currency in industrial governance, this system provides a means to develop standardized and reliable methods independent of operator expertise. Thanks to that, Statistica has achieved an unprecedented record of recognition from both users and expert reviewers since its first Windows release in 1993, being ranked very highly in every review by an independent third party.

SSI

SSI[24] Scientific Software offers, among others, famous program for the structural modelling Lisrel, the pioneering software for structural equation modelling including statistical methods for complex survey data.

StatPlus 2007

With StatPlus 2007[25] you do not need to carry a mainframe computer to perform a complex statistical analysis when you visit a remote lab. Even a notebook computer is no longer necessary. With StatPlus Portable you only need a compact USB flash drive, plug it into any PC in the lab, and get your customary workplace in a matter of minutes.

Systat

Systat[26] is one of the well established names in the data analytic circles stemming from the pioneering 1970s work of Leland Wilkinson at the University of Illinois. It is also one of the serious language based scientific tools in the market. The same producer also developed a software package SigmaStat[27] that does statistical analysis and performs these tasks using a wizard interface.

Statistical tools were extended considerably in recent releases. There are new additions, expansion of existing provisions, and some useful modifications well executed without significant disturbance to existing practice. There is some particular attention to areas of interest for industrial quality work and design of experiments, response surface methods being a prominent and valuable example.

For a long time Systat could be described as a powerful product where you are expected to know what you are doing, however, since its acquisition by Cranes Software[28] this has been softening. Some commands previously available only from the command line are now available from mouse and menus; already available techniques have been made easier to find and used through new or rearranged interface. Users with an interest in exploratory practice particularly welcome this continued evolution. The same interface continues to move towards collective experience and familiarity with de facto conventions. The tabbed notebook structure, customizable to tiled or cascaded display, the worksheet and editors have become more like what you expect from elsewhere etc. There are still some idiosyncratic aspects, but they are now definitely the surprise exception rather than the rule.

2.3. *Free and open statistical software*

In these days there exist several hundreds of open source projects which are more or less related to statistics and data analysis. We start with two important lists. The first one concentrates on statistics while the second one on informatics and mathematics. Links to the other interesting lists can be found, e.g., via[29].

- Probably most complete list of free software for statistics is the one maintained and regularly updated by John C. Pezzullo. The list is available on[30].
- Aside that, many statisticians can find more than useful the programs initially intended for informatics and mathematics. One of the most important lists created within the FreeBSC[31] project can be found at[32].

The rest of this subsection focuses on several typical representatives of different fields of statistics. The choice is by no means exhaustive and represents rather the personal interests of the author.

- Graph Visualization is a way of representing structural information as diagrams of abstract graphs and networks. Automatic graph drawing has many important applications in software engineering, database and web design, networking, and in visual interfaces for many other domains. Graphviz[33] is an open source graph visualization software. It has several main graph layout programs, web and interactive graphical interfaces, and auxiliary tools, libraries, and language bindings.
- Gretl[34] is an open-source software application for compiling and interpreting data coming from the field of econometrics. It is an acronym for Gnu Regression, Econometrics and Time-Series Library. It has a graphical user interface and can be used together with X-12-ARIMA, TRAMO/SEATS,[35] a program for estimation, forecasting, and interpolation of regression models with missing values and ARIMA errors, in the presence of possibly several types of outliers when no restriction is imposed on the location of the missing observations in the series, and R. TEXadvocates will appreciate possibility of the output redirection to TEX.

[24] http://www.ssicentral.com/index.html
[25] http://www.analystsoft.com/
[26] http://www.systat.com/
[27] http://www.systat.com/products/SigmaStat/
[28] http://www.cranessoftware.com/
[29] http://en.wikipedia.org/wiki/Statistical_software
[30] http://statpages.org/javasta2.html
[31] http://www.freebsd.org/
[32] http://www.freebsdsoftware.org/math/
[33] http://graphviz.org/
[34] http://gretl.sourceforge.net/
[35] http://ideas.repec.org/p/wpa/wuwpem/0410008.html

- Gnumeric[36] is a free spreadsheet program that is part of the GNOME desktop and has Windows installers available. It is intended to be a free replacement for proprietary spreadsheet programs such as Microsoft Excel, which it broadly and openly emulates. Many statistical functions are available.
- OpenEpi[37] is a free, web-based, open source and operating system independent series of programs for use in epidemiology, biostatistics, public health, and medicine, providing a number of epidemiologic and statistical tools for summary data. Despite limited number of implemented routines it is very popular in its domain.
- PAST[38] is a free, easy-to-use data analysis package originally aimed at paleontology but now also popular in ecology and other fields. It includes common statistical, plotting and modelling functions.
- RandomNumbers[39] is a project of the University of Geneva and the company id Quantique[40] offering the possibility to download "true" random numbers generated using a quantum random number generator.
- Resampling Stats[41] is a long term project aimed at teaching and analyzing the data not via standard inference techniques but using bootstrap, permutations and simulations instead. It is available both as the add-in for Excel and a software package for Matlab.
- Tanagra[42] is a free data mining software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area.
- Wessa[43] is an advanced, and very well done, interface powered by R enabling an interactive statistical calculations. It is a typical prototype of the recent trend aimed at preparing a sophisticated interface over a reliable "engine" for all those who are much more interested in the results than in the implementation of the routines.

3. Integrated environments for statistical computing

An integrated environment for statistical calculations and graphics is fundamental for developing an understanding new techniques in statistics. Such an environment must essentially be a programming language. Its basic data types must include types that allow groups of numbers, say data sets, to be manipulated as entire objects. However, in model-based analysis numerical data are only part of the information being used. The non negligible remainder is the model itself. In this section we will take a closer look on several typical representatives of this type of programmes.

R environment

R[44] is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment developed at AT&T Bell Laboratories by John Chambers and colleagues. R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems including FreeBSD and Linux, Windows and MacOS.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display. It includes an effective data handling and storage facility, a suite of operators for calculations on arrays, in particular matrices, a large coherent integrated collection of intermediate tools for data analysis, graphical facilities for data analysis and display either on-screen or on hardcopy, and a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities. The term "environment" is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with some other data analysis software.

R, like S, is designed around a true computer language, and allows users to add additional functionality by defining new functions. The S language was, and still for many statisticians is, often considered the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity. More precisely, R can be considered as a different implementation of S. There are some important differences, but typical code written for S runs unaltered under R. Much of the system itself is written in the R, which makes it easy for users to follow the algorithmic choices made. For computationally intensive tasks, C, C + + and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and many more) and graphical techniques. Aside that, R can be easily extended via packages; several hundreds are at the moment available through the CRAN[45] family of Internet sites. One of the R's strengths is the ease with which well-designed publication quality plots can be produced, including mathematical symbols and formulæ where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control. R has its own LaTeX-like documentation format, which is used to supply comprehensive documentation, both on-line in a number of formats and in hardcopy.

Many users think of R as a statistics system. We prefer to think of it of an environment within which many statistical techniques are implemented.

[36] http://www.gnome.org/projects/gnumeric/
[37] http://www.openepi.com/
[38] http://folk.uio.no/ohammer/past/
[39] http://www.randomnumbers.info/
[40] http://www.idquantique.com/
[41] http://www.resample.com/
[42] http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html
[43] http://www.wessa.net/
[44] http://www.r-project.org/
[45] http://cran.r-project.org/

S-Plus

S-Plus[46] does not seem to go in the academic direction as it did five-ten years ago, when the S language was for many people the main vehicle of choice for research in statistical methodology. Today, it is rather the "triumvirate" of life sciences (oriented mainly on clinical data analysis), finances (oriented mainly on risk management and investment management), and marketing analytics (oriented mainly on creation more targeted, effective campaigns) which push S-Plus ahead. Nevertheless, its newest version continues to deliver very good graphical possibilities and the advantages of an open architecture, scalability and flexibility to integrate advanced analytics into the everyday business processes.

Hand-in-hand comes access to the latest versions of all the S-Plus modules, being separate products providing significant enhancement of the S-Plus in particular areas of work.

(1) S+ ArrayAnalyzer providing assay methods primarily of interest to the pharmaceutical industry.
(2) S+ FinMetrics aimed at the financial services sector, being an advanced modelling and estimation package designed for market prediction and trading design.
(3) S+ NuOPT, optimization package for very large data sets, well suited to accompany S+ FinMetrics, and positioned as such by Insightful, but the core methods are applicable to many science situations as well.
(4) S+ Wavelets, enabling image, signal and time series analysis using wavelet methods.
(5) FAME S-PLUS connector accelerating the process of bringing accurate financial time series data from FAME[47] into the S-PLUS environment.
(6) S+ EnvironmentalStats aimed at civic or corporate professionals working directly to legislative or regulatory requirements, but core capabilities are of much wider utility.
(7) S+ SeqTrial providing group sequential methods for clinical trials.
(8) S+ SpatialStats enabling the analysis of spatially distributed data of all kinds.

Those who regularly fight with the growing size of their data will acknowledge the Big Data facility applicable to many analyses. Unfortunately, it is one of the two key features available only in the Enterprise Developer version. Data sets too big for available RAM are handled as the bdFrame type via a binary cache on the hard disk which serves as virtual RAM, not part of the usual OS virtual memory, but a separate and dedicated file. Once data are in, they are not analyzed as fast as in RAM, but nevertheless with very respectable speed. The language sees similar incremental changes. All of this is valuable, but seems to be overshadowed by an implementation of the Eclipse IDE[48] customized for the S language and incorporated as the standalone S-Plus Workbench.

Xlisp-Stat/Lisp-Stat

Xlisp-Stat/Lisp-Stat[49] is an extensible statistical computing environment for data analysis, statistical instruction and research, with an emphasis on providing a framework for exploring the use of dynamic graphical methods. Extensibility is achieved by basing Lisp-Stat on the Lisp language, in particular on a subset of Common Lisp. Lisp-Stat extends standard Lisp arithmetic operations to perform element-wise operations on lists and vectors, and adds a variety of basic statistical and linear algebra functions. A portable window system interface forms the basis of a dynamic graphics system that is designed to work identically in a number of different graphical user interface environments, such as the Macintosh operating system, the X window system, and Microsoft Windows. A prototype-based object-oriented programming system is used to implement the graphics system and to allow it to be customized and adapted. The object-oriented programming system is also used as the basis for statistical model representations, such as linear and nonlinear regression models, and generalized linear models.

Many aspects of the system design and many of the functions were motivated by the S language. Main reasons behind the decision to produce this environment were threefold. At first, to provide a vehicle for experimenting with dynamic graphics and for using dynamic graphics in instruction. Second, to allow experimentation with an environment supporting functional data, such as the mean functions in nonlinear regression models and prior density and likelihood functions in Bayesian analysis. Finally, to explore the use of object-oriented programming ideas for building and analyzing statistical models. Despite the fact that the syntax of Lisp is somewhat unfamiliar to most users of statistics, this system attracted many researchers, especially those oriented in informatics.

XploRe

XploRe[50] is a combination of classical and modern statistical procedures, in conjunction with sophisticated, interactive graphics. XploRe is the basis for statistical analysis, research, and teaching. Its purpose lies in the exploration and analysis of data, as well as in the development of new techniques. In addition, XploRe is a high level object-oriented programming language using which users can write procedures or functions, such as in Pascal or C/C++. Variables can be collected in list structures, so it is possible to hold common information of a data set in a single data object. Of course, all the features of a high-level language like recursion, local variables, loops, and conditional execution are available. Dynamic link calls are possible, so you can incorporate your own methods in XploRe, enabling you to easily extend the environment. The statistical methods of XploRe are provided by various quantlets. An automatic HTML converter ensures the smooth integration of quantlets and libraries into the help system.

[46] http://www.insightful.com/
[47] http://www.fame.org/
[48] http://www.eclipse.org/
[49] http://www.stat.uiowa.edu/~luke/xls/xlsinfo/xlsinfo.html
[50] http://www.xplore-stat.de/

4. Computer algebra and/or mathematical systems and statistics

There exists quite a clear distinction between practical and theoretical approaches to most statistical work. The majority of software products for statistics are on the practical side, using numerical and graphical methods to provide the user access to existing methods. On the other hand, there do not exist too many software packages specifically designed for pure statistical and mathematical modelling. However, practically all available computer algebra or mathematical systems offer, among others, tools for both theoretical and practical statistical work. Therefore, take a look on some possibilities in this area.

GAMS

GAMS,[51] the General Algebraic Modeling System, is specifically designed for modelling linear, nonlinear and mixed integer optimization problems. The system is especially useful with large, complex problems. GAMS is available for use on personal computers, workstations, mainframes and supercomputers.

GAMS is especially useful for handling large, complex, one-of-a-kind problems which may require many revisions to establish an accurate model. The system models the problems in a highly compact and natural way. The user can change the formulation quickly and easily, can change from one solver to another, and can even convert from linear to nonlinear with little trouble. Using GAMS, data are entered only once in familiar list and table form. Models are described in concise algebraic statements which are easy both for humans and machines to read. The GAMS language is formally similar to commonly used programming languages, being thus familiar to anyone with programming experience. Models are fully portable from one computer platform to another. Numerical analysts appreciate very much not only the reliable implementation of included solvers but also the fact that GAMS facilitates sensitivity analysis.

GAUSS

GAUSS[52] Mathematical and Statistical System is a fast matrix programming language for mathematics and statistics. It is a complete analysis environment suitable for performing quick calculations, complex analysis of millions of data points, or anything in between. Its primary purpose is the solution of numerical problems in statistics, econometrics, time series, optimization, and 2D- and 3D-visualization. First compiled in 1984 for MS-DOS; the latest compiled version is available for Linux, Sun SPARC Mac OS X and Windows. Designed for computationally intensive tasks, the GAUSS system is ideally suited for the researcher who does not have the time required to develop programs in C or FORTRAN, but finds that most statistical or mathematical packages are not flexible or powerful enough to perform complicated analysis or to work on large problems.

Maple

Maple[53] is a prototype of computer algebra systems, i.e. a software programs that facilitate symbolic mathematics. The core functionality is manipulation of mathematical expressions in symbolic form. Aside that, it incorporates a full high-level programming language and interfaces to other languages (C, Fortran, Java, Matlab, Visual Basic and Excel). Most of the mathematical functionality of Maple is written in the Maple language itself, which is interpreted by the Maple kernel. Unusually for a commercial program, most of the source code is freely viewable.

Core function features are vital, but not exclusively so. It appears that in the last years Maple has been radically developing its interface. What has been considered by many people to be one of the less friendly computer algebra environments is becoming, in many ways, one of the most imaginative. A slide show mode enabling presentations without leaving the original document, documentation of processes including automatic labelling of context menu operations, mathematical handwriting recognition, just to mention some of them. In combination with the enriched environment it now offers a definite view of where computer algebra interfaces go, especially if used with a graphics tablet and pen or a tablet PC.

Moving on from form to function, there appeared recently important enhancements. Multithreaded code is a headline addition, taking advantage of multiple or multi-core CPUs through its own package (Threads) and separate mathematical engine. Linear algebra and numerics are of key interest not only for the statisticians. Thus, it seems to be a big step ahead a release of a toolbox, the Maple-NAG connector, being a front end to Numerical Algorithm Groups numerical C routines. With its help one can combine the pre-eminent modeling, exploration and application development abilities of Maple with the power of NAG numeric routines. Moreover, in Maple the NAG routines have been extended to allow arbitrarily-large precision.

Matlab

Matlab[54] is a numerical computing environment and programming language allowing easy matrix manipulation, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs in other languages. Although it is "numeric only" system, an optional toolbox interfaces with the Maple symbolic engine, allowing access to the computer algebra capabilities.

Thanks to its excellent numerics, matrices as a natural data type for most of the statistical analysis, open and transparent language and hundreds of macros, Matlab became very popular general mathematical system not only among engineers, scientists, financial analysts, but also among statisticians.

Its Statistics Toolbox provides users a comprehensive set of tools to assess and understand their data, analyze historical data, model and simulate them, as well as to develop easily statistical algorithms and prepare simulations. It offers a rich set of statistical plot types and interactive graphics, such as polynomial fitting and response surface modeling. All included functions are written in the Matlab language. This means that you can inspect the algorithms, modify the source code, and create your own custom

[51] http://www.gams.com

[52] http://www.aptech.com/

[53] http://www.maplesoft.com/

[54] http://www.mathworks.com/

functions. Evidently, Matlab offers not only its core and statistical functions. It is worth to mention, among others, its financial, optimization, signal, image, wavelets, and many others, toolboxes.

Among the clones of Matlab we would like to stop for a while with two, SciLab and Octave.

SciLab[55] is a numerical computational package developed by INRIA[56] and ENPC.[57] It is a high level programming language in that most of its functionality is based around the ability to specify many computations with few lines of code. It does this primarily by abstracting primitive data types to functionally equivalent matrices.

In functionality it is similar to Matlab, but is available for download at no cost. The program enables users to compute a wide range of mathematical operations from relatively simple operations such as multiplication to high level operations such as correlation and complex arithmetic. The software is often used for signal processing, statistical analysis, image enhancement or fluid dynamics simulations. The syntax is similar to Matlab, but the two are not completely compatible, though there is a converter included in SciLab for Matlab2Scilab conversions. SciLab has fewer help files than Matlab. SciLab also includes a package Scicos for modelling and simulation of explicit and implicit dynamical systems including both continuous and discrete sub-systems.

GNU Octave[58] is a high-level language, primarily intended for numerical computations. It provides a convenient command line interface for solving linear and nonlinear problems numerically, and for performing other numerical experiments using a language that is mostly compatible with Matlab, and considered by many users as another open source variant of Matlab.

MuPAD

MuPAD[59] is another powerful mathematical problem solving environment for exact symbolic and numeric computing with arbitrary precision. Originally developed by the MuPAD research group at the University of Paderborn, it is developed and maintained since 1997 by the company SciFace in cooperation with the MuPAD research group and partners from some other universities. Until autumn 2005 the version MuPAD Light was offered for free for research and education, but as consequence of closing the home institute of the MuPAD research group there is now only the version MuPAD Pro available with costs.

The MuPAD kernel is bundled with Scientific Notebook and Scientific Workplace,[60] while former versions of MuPAD Pro were bundled with SciLab. Most recently it was adopted as the computer algebra system for the popular MathCAD[61] package in its version 14 release replacing the previous Maple based engine.

MuPAD Pro provides a Pascal-like programming language allowing imperative, functional and object-oriented programming. Its concept of domains and categories corresponds to object-oriented classes and supports overloading of methods and operators, inheritance, and generic algorithms. The comfortable worksheet interface (called notebooks) includes high-quality interactive 2D/3D graphics tools for visualization and animation, an integrated source-level debugger, a profiler, and comprehensive hypertext help with extensive examples.

The range of tools for statistics is quite broad and resembles those mentioned in the case of Maple and Mathematica, comprising numerous distributions with CDFs, PDFs, quantile functions and random generators, data analysis, linear and non-linear regression, mean, standard deviation and statistical tests, among others.

Wolfram *Mathematica*

Mathematica[62] is considered by many people as the de facto standard for symbolic computation with a strong graphics capability. For more than a decade, aside from its symbolical capabilities, is also used as a high-powered statistical system thanks to its growing statistical capabilities. To that purpose Mathematica uses a well-developed symbolic language that makes it easy to specify arbitrarily extensible statistical models and data analysis methods. Another important feature concerns data acquisition, where Mathematica, thanks to its unique core architecture, provides convenient importing of the world's broadest range of data formats.

Most existing statistical packages have gradually evolved from specific libraries of functions and are improved and enlarged with each new release. Aside the packages enabling data manipulation, data smoothing, statistical plots, cluster analysis, descriptive statistics, hypothesis testing and confidence intervals or linear and nonlinear regression, coming directly with Mathematica, there exist many books presenting a unified approach for doing mathematical statistics with Mathematica. The emphasis is typically on problem solving. Moreover, there exist third party packages prepared outside the Wolfram company. One such an interesting system is:

mathStatica

The Mathematica application package mathStatica[63] was designed to solve the algebraic and symbolic problems that are of primary interest in mathematical statistics. It does so by building upon the symbolic computational power of Mathematica to create a sophisticated tool set specially designed for doing mathematical statistics.

The package is loaded like any other and opens to show a Mathematica palette of four self explanatory options, i.e. continuous, discrete and kernel, while the fourth option is a link to the relevant section of the Mathematica help system. Notice that each of the first three options opens further palettes with respective functions. The full power of the package is, however, available through direct calls from the command line or programs.

Functions behind the palette interface range from small but useful generic developments of Mathematicas own

[55] http://www.scilab.org/

[56] http://www.inria.fr

[57] http://www.enpc.fr/

[58] http://www.gnu.org/software/octave/

[59] http://www.sciface.com/

[60] http://www.mackichan.com/

[61] http://www.ptc.com/

[62] http://www.wolfram.com/

[63] http://www.mathstatica.com/

provision through replacements of existing facilities to new specialized functions. Altogether, there are about 100 functions covering a broad interest of mathematical statisticians which are not native to Mathematica itself like specific plots, order statistics, decision theory, copulæ, numerical and symbolic maximum likelihood estimation, an many more.

5. Future of statistical computing

It is very difficult to predict what will be the next major driving forces that will influence the statistical software development process. However, it seems to us that among the numerous challenges belong:

- Non-expert users, i.e. the individuals who wish to apply statistical techniques to analyze datasets coming from financial, biological, physical, astronomical, and other domains. These users usually have only limited statistical training, and it is easy for them to perform erroneous analysis of their data. We guess that this will lead to the renaissance of the “expert system” and/or artificial intelligent packages, so popular ten years ago but almost abandoned nowadays.
- Data mining of the data that was collected under uncontrolled circumstances, which complicates their analysis. This challenge will continue to grow in coming years.
- Text mining and text data analysis that is necessitated by the requirements of computer security, bioinformatics, automated scientific discovery, intelligence research, and analysis of texts available on Internet.
- Analysis of data arriving in a serial manner that change in time their nature, dimensionality, type, etc. One of the biggest challenges is that their analysis and visualization require new techniques “evolving” with the data and their changes.
- Analysis of very large and complicated data sets coming from genomics, astronomy, nuclear physics, chemistry etc. Analysis of functional data.
- Parallelisation of the algorithms enabling “easy” use of computer grids.
- Development of new techniques for graphical representation of large multivariate data.

6. Further reading

It is not surprising that for almost every software described above there exists at least one book closely connected with. Therefore, we have chosen for most of the programmes described above one typical representative as a recommended further reading for interested readers. Nice introduction to GENSTAT still remain [2], to SPSS [3], to SixSigma [5], to XploRe [13], to JMP [14], to OpenEpi [16], to MuPAD [18], to LISP-STAT [23] and to Systat [26]. On the other hand, instead on the language itself many books concentrate rather on examples and case studies realized with given programme as is the case of [8] and SAS, [9] and R, [19] and SPSS, [20] and Matlab, [21] and Mathematica or [24] and S+. Another important source of the information are Handbooks of Computational Statistics [7], [10] and [11], published recently by Springer as well as the monograph about data visualization [6]. Finally those who are more interested in teaching should consult [4] or [15].

Acknowledgement

The work is a part of the research project MSM 0021620839 financed by MŠMT.

REFERENCES

[1] J. Antoch, Series of papers on statistical software and environments for statistical computing (in Czech for the Czech Statistical Society Newsletter and other publications).
[2] Applied Statistics: Handbook of GENSTAT Analysis, 1st edition, Chapman & Hall/CRC, 1991.
[3] G. Argyrous, Statistics for Research: With a Guide to SPSS, Sage Publications, 2005.
[4] S. Blank, Ch. Seiter, P. Bruce, Resampling Stats in Excel, 2nd edition, Resampling Stats Inc., 2001.
[5] Q.S. Brook, Six Sigma and Minitab: A Complete Toolbox Guide for All Six Sigma Practitioners, 2nd edition, QSB Consulting, 2006.
[6] Ch. Chen, Information Visualization: Beyond the Horizon, 2nd edition, Springer, 2006.
[7] Ch. Chen, W. Härdle, A. Unwin, Handbook of Computational Statistics III — Data Visualisation, Springer, Berlin, 2006.
[8] R. Cody, Learning SAS by Example: A Programmer’s Guide, SAS Publishing, 2007.
[9] P. Dalgaard, Introductory Statistics with R, Springer, 2004.
[10] V. Esposito Vinzi, W. Wynne, W.W. Chin, J. Henseler, H. Wang, Handbook of Computational Statistics II — Partial Least Squares, Springer, Berlin, 2005.
[11] J.E. Gentle, W. Härdle, Y. Mori, Handbook of Computational Statistics I — Concepts and Methods, Springer, Berlin, 2004.
[12] Google search machine.
[13] W. Härdle, S. Klinke, B.A. Turlach, XploRe: An Interactive Statistical Computing Environment (Statistics and Computing), Springer, 2007.
[14] JMP Start Statistics: A Guide to Statistics and Data Analysis Using JMP, 4th edition, SAS Publishing, 2007.
[15] Z.A. Karian, E. Tanis, Probability and Statistics Explorations with MAPLE, 2nd edition, Prentice Hall, 1999.
[16] D.G. Kleinbaum, K. Sullivan, N. Barker, A Pocket Guide to Epidemiology, Springer, 2006.
[17] N.C. Lauro, Computational statistics or statistical computing, is that the question? Computational Statistics & Data Analysis 23 (1996) 191–193.
[18] M. Majewski, MuPAD Pro Computing Essentials, 2nd edition, Springer, 2004.
[19] S.J.P. de Marques, Applied Statistics Using SPSS, STATISTICA, MATLAB and R, 2nd edition, Springer, 2007.
[20] W.L. Martinez, A.R. Martinez, Computational Statistics Handbook with MATLAB, Chapman & Hall/CRC, Boca Raton, 2002.
[21] C. Rose, M.D. Smith, Mathematical Statistics with MATHEMATICA, Springer, 2002.
[22] Scientific Computing World Journal.
[23] L. Tierney, LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics, Wiley-Interscience, 1990.
[24] W.N. Venables, D. Ripley, Modern Applied Statistics with S, 4th edition, Springer, 2003.
[25] Wikipedia, a multilingual web-based, free content encyclopedia project.
[26] L. Wilkinson, G. Blank, Ch. Gruber, Desktop Data Analysis With Systat, Prentice Hall, 1996.