Statistical Software for Today and Tomorrow

Edward J. Wegman
Center for Computational Statistics
George Mason University
Fairfax, VA 22030

Jeffrey L. Solka
Code B10
NSWCDD
Dahlgren, VA 22448

January 13, 2005

Abstract. This article is a relatively short exposition on the current state of the art with regards to statistical software along with some prognostications as to the future trends in the area. The article is somewhat biased in its discussions of the state of the art with a necessitated focus on those packages that seem most promising with regards to their current and future role in the academic research communities.

Keywords and Phrases. statistical software

AMS Subject Classification. Primary: 62-04; Secondary: 68-04

1 Introduction

This article is an assessment of the state of the art of the statistical software arena as of 2004. Much like a snapshot is only part of a movie, this article is a snapshot of a currently evolving system which is being driven in part by the needs of the user community and capitalistic forces. We do however make some predictions as to what the next few frames of the "statistical software" movie might be.

We have attempted to touch upon a few commercial packages, a few general public license packages, a few analysis packages with statistics add-ons, and a few general purpose languages with statistics libraries. For the most part, our discussions are cursory in nature although we have attempted to be unbiased in our reviews.

We begin our discussions with SAS®. SAS® , in our opinion, represents the Microsoft of the statistical software companies and hence our discussions in this section are a bit more lengthy. We next proceed to BMDP. BMDP is a package which had its birth in the bio-medical arena. Next, we proceed to SPSS. SPSS is another commercial package like SAS® and in many ways it represents one of the main competitors to SAS®.We then proceed to discuss S-PLUS®. S-PLUS® has quickly evolved from its birth as the S language into a bona fide commercial package with a loyal user base.

We change gears a bit in the next section and discuss a non-commercial implementation of the S language, R. R, like LATEX, is an example of a general public license success story. The R language benefits from the "many eyes" that examine the code and its large user base provides a user with an alternative to the standard "help desk" to turn to for assistance. R is currently one of our languages of choice for the development of research level software.

The next several sections focus on commercially available mathematical analysis software that have statistical add-on packages. The first of these, MATLAB, is also one of our favorites. MATLAB is a language with strong visualization and matrix capabilities. A discussion of Mathematica follows. Mathematica is touted as the de facto standard for symbolic computation with a strong graphics capability. These capabilities make it an asset for calculations associated with mathematical statistics.

Our next section briefly examines several general purpose software development languages that have various statistics libraries. The first one, JAVA, provides the user with the capability to offer students and clients access to statistical software on-line. This along with its ability to seamlessly execute on numerous platforms running numerous operating systems makes JAVA a strong contender as a general purpose implementation language.

We next discuss C++. C++, like JAVA, possesses the strength of being an object oriented language. This allows one to extend pre-existing standard template libraries in a rather straight-forward manner. C++ has been around long enough to have a large numerical analysis user base associated with it.

The final general purpose language that we discuss is PERL, the Practical Extraction and Reporting Language. PERL is a highly flexible language that allows a user to craft very dense code to perform many purposes. Free form text processing is one of the strengths of PERL and this has led to it being embraced by many diverse communities including the bioinformatics and the computer security communities.

In the next section of the paper we attempt to predict what will be the next major driving forces that will influence the statistical software development process. This is a very difficult task given the fluid nature of the commercial and research sectors. We do believe that some of the predictions will come to fruition though.

In the final section of the paper we review some of the ground that has been covered within. We have chosen to make our synopsis somewhat succinct given the rather lengthy discussions that have proceeded this point in the paper. We close this particular section with a some very brief philosophical musings.

2 Current Statistical Packages

SAS ⓡ

SAS® had its birth as a statistical analysis system during the late 1960's. SAS® grew out of a project in the Department of Experimental Statistics at North Carolina State University. This project led to the formation of the SAS Institute in 1976. Since its infancy, SAS® has evolved into an integrated system for data analysis and exploitation.

SAS® is really much more than a simple software system. The SAS® website claims that their software resides at 40,000 sites worldwide including 90 percent of those companies on the Fortune 500 list. We believe that this expansion has come about in part because the SAS management has aligned themselves with the recent "statistical-like" advances within the computer science community such as data mining. In fact if one visits the SAS website at www.sas.com one is hard pressed to ascertain that the site details a statistical software product. An initial perusal of the site leads one to believe that the product that is being offered by the company is a data mining product. This clever integration of mathematical/statistical methodologies, database technology, and business applications has helped propel SAS® to the top of the commercial statistical software arena.

SAS® has been utilized in a myriad of different business areas including the automotive industry, the telecommunications industry, the life sciences area, the home land security area, the economic forecasting area, and the waste and fraud detection area. SAS also continues to remain one of the languages that are typically taught at academic institutions world-wide.

The architecture for the SAS approach is called the SAS® Intelligence Platform. This Platform is really a closely integrated set of hardware/software components that allow a user to fully utilize the business intelligence (BI) that can be extracted from their client base. The SAS Intelligence Platform consists of the following components: the SAS® Enterprise ETL Server, the SAS® Intelligence Storage, the SAS Enterprise BI Server, and the SAS Analytic Technologies. Our real interest within this chapter is the SAS® Analytic Technologies, but we believe that it is beneficial to at least touch upon the nature of the other components.

The SAS® Enterprise ETL Servers are really just a means to provide a unified database environment to handle metadata and to provide multi-threaded platform independent access to the business information within a GUI-based environment. The SAS Intelligence Storage System serves as a repository for the information that is ultimately analyzed using analytic application in order to provide BI insights. The SAS® BI Server provides the user with universal format type access to their collected data structures along with novice user facilitated interfaces to the analytic software that is used to obtain the BI insights. In this manner, a user who is relatively new to the data analysis area can still obtain quality insights into their data and obtain the BI that is needed to allow them to compete in today's highly competitive markets. The SAS® Analytic Technologies contain the advanced mathematical and statistical algorithms along with visualization frameworks which allow even a mathematically/statistically unsophisticated user to interpret the insights obtained by these algorithms.

The analytic technologies are really the areas that we are most interested in within this current tome. The SAS management has been very astute in marketing their analytic capabilities in such a manner that the buyer can easily ascertain their relevance to their particular problem of interest. The SAS analytic technologies include data and text mining, forecasting and econometrics, operations research, quality improvement, and statistical analysis. We would like to take a moment to briefly touch upon each of these analytic capabilities, ending with the statistical analysis one.

The data and text mining capability is focused on the use of text mining technologies to foster business applications, although one can easily formulate alternative application in the areas of homeland security. We believe that the creators of SAS® have demonstrated great foresight in extending their analysis capabilities in order to provide text mining capabilities. It is the authors' contention that text mining will continue to play a more prominent role in the formulation of effective national scientific strategies. In fact, some of the giants in the IT arena, Microsoft and Apple, have recently been gearing up to compete with the stalwart text search company Google. There are certainly fruitful areas for interaction between standard statistical and computational statistical procedures and text mining.

The econometric, time series analysis, and forecasting area is an area often associated with statistics. SAS® has done a good job of closely coupling time series analysis to business applications. They provide the capability to deal with the computational complexities associated with making large quantities of accurate forecasts utilizing large volumes of data. They also provide novice users with an interactive interface that facilitates their ability to make meaningful forecasts.

Operations research and quality improvement are two areas that one might not immediately identify as statistical applications. However, these two areas are very relevant to the current culture within the business community. The astute reader will, after a moments reflection, realize that optimization, the enabling technology in operations research, and quality control/quality inspection have received previous attention from the statistics community. SAS® has employed very clever business practices in bringing these to the forefront as analytic product areas.

Finally, we turn our discussions to the statistical analysis analytic capability. Many of the previous discussions with regard to the capabilities of their computer/database support architectures lend support to the statistical analysis analytic capability. Their statistical analysis capability can be broken down into the sub-areas of statistical analysis, exploratory data analysis, matrix programming language, matrix programming interface, guided data analysis, market research, and other statistical components. Let us take a moment to consider each of these in turn.

The SAS/STAT component contains those capabilities that one normally associates with a data analysis package. These capabilities include functionality to support analysis of variance (ANOVA), regression, categorical data analysis, multivariate analysis, survival analysis, psychometric analysis, cluster analysis, and nonparametric analysis. One of the strengths of the SAS® SAS/STAT package is the fact that the package is constantly being upgraded with each release in order to reflect the latest and greatest algorithmic developments in the statistical arena.

The exploratory data analysis area is supported by the SAS/INSIGHT package. This package is designed to analyze data via linked visualization strategies. These strategies can be linked across multiple windows and they allow an adroit user to uncover trends, spot outliers and readily discern subtle patterns within their datasets that would not be apparent using other analytical methods. The hallmark of this portion of the SAS software is the interactive nature of the data analysis procedure.

SAS® provides the user with a convenient matrix programming language via their SAS/IML software system. This system allows statisticians, mathematicians, and computer scientists to develop their own matrix-based analysis systems. Formulation of many statistical algorithms are facilitated by the use of a matrix based language.

A matrix programming interface to this language is provided within the SAS/IML Workshop GUI. This GUI gives the user an integrated editor/debugger for the development of SAS/IML based programs. The SAS/IML Workshop GUI also provides the user with the ability to call SAS procedures and external C/Fortran/Java functions.

The SAS/LAB system provides a user with a prompt-driven guided data analysis methodology. The system uses graphical methods to represent basic methods of statistical analysis. In this way, the user is guided through the analysis procedure.

The last SAS analytical package that we wish to discuss is the SAS Market Research package. This package provides the user with a point and click interface that allows them to conduct many common market research functions. Some of these functions include discrete choice analysis, cojoint analysis, and multi-dimensional preference scaling.

Many of the methodologies that we previously discussed serve to support the SAS BI model. These technologies provide a user with on-line analytical processing (OLAP) capabilities in order to obtained summarized data views. This user-based analysis is facilitated by providing the user with easy access to his data, with an interactive data exploration environment, and with a means to facilitate web-based reporting in order to achieve flexible content delivery. The user is referred to [1] for a more in-depth look at the impact of SAS® in the BI arena.

The BI process is of course facilitated through the use of data exploration and analysis via visualization frameworks. The user is free to choose from a plethora of different visualization methods including surface, prism, block, and chloropleth maps. As is the case with virtually all modern data analysis packages the user has the capability to output the graphics to any of a number of different graphics devices.

In summary, SAS® has become one of the "biggest players" in the statistical software arena. It has done a great job at marketing itself as a data mining company. It also has aggressively expanded its client base to include numerous companies, government agencies, and academic institutions. It will continue to be an active statistical software company for the foreseeable future and, based on prior company maneuvers, one can expect that SAS will aggressively position itself to meet the future needs of the business communities.

BMDP

BMDP originated during the 1960s as a bio-medical analysis package. In many ways BMDP has remained true to its roots and this is evidenced by its long list of clients which includes such biomedical giants as Astra Zeneca, Bristol-Myers Squibb, Eli Lilly, the FDA, Glaxo Wellcome, and Merck. BMDP is currently produced by Statistical Solutions. In addition BMDP has chosen to align itself with a number of other currently popular statistical products including StatXact 5.0, LogXact 5.0, SOLAS, EquivTest, SigmaPlot, Meta Analysis, and SYSTAT. These alliances have extended the base capabilities of BMDP on many different fronts.

The current release of BMDP is Professional Version 2.0. BMDP is available for a variety of platforms/operating systems including Microsoft Windows XP, MS DOS, HP 9000, Sun Solaris OS, IBM AIX, Dec Alpha, Dec Vax/VMS and more. The base distribution of BDMP includes routines for data manipulation, data description, group comparisons, various plots including histograms, frequency tables, correspondence analysis, regression, maximum likelihood estimation, non-linear regression, analysis of variance, multivariate analysis, nonparametric analysis, cluster analysis, missing value analysis, survival analysis, and time series analysis.

SPSS

The SPSS software system was developed in the late 1960s by SPSS Chairman of the Board Norman H. Nie, C. Hadlai (Tex) Hul, and Dale Brent. The three of them were Stanford University graduate students at the time. Nie and his colleagues founded SPSS in 1968. SPSS incorporated in 1975. It established its headquarters in Chicago, Illinois where it currently resides. SPSS became publicly traded in August of 1993.

SPSS was originally a software package designed for mainframe use. SPSS introduced SPSS/PC+ for computers running MS-DOS in 1984. This was followed by releases of a UNIX software version in 1988 and a Macintosh version in 1990.

SPSS 13.0 for Windows was current as of 2004. The current release of the base distribution for MAC® OS X is version 11.0. There is also a server version of the software and and instructional versions for undergraduate and graduate academic institutions.

SPSS supports numerous add-on modules including one for regression, advanced models, classification trees, table creation, exact tests, categorical analysis, trend analysis, conjoint analysis, missing value analysis, map-based analysis, and complex samples analysis. In addition SPSS supports numerous stand-alone products including Amos™(a structural equation modeling package), SPSS Text Analysis for Surveys™(a survey analysis package that utilize natural language processing methodology (NLP)), SPSS Data Entry™(a web-based data entry package), AnswerTree®(a market segment targeting package), SmartViewer® Web Server™(a report generation and dissemination package), SamplePower®(sample size calculation package), DecisionTime® and What if?™(a scenario analysis package for the non-specialist), SmartViewer® for Windows (a graph/report sharing utility), SPSS WebApp Framework (web-based analytics package), and the Dimensions Development Library (a data capture library).

As can be ascertained from the above discussions, the SPSS Corporation is targeting many of the same application areas as SAS. However, it still seems that the client bases for SAS is much more extensive than that of SPSS. The user is referred to [2] for a recent in-depth review of SPSS as of release 12.

S-PLUS®

S-PLUS® is an extension of the statistical analysis language S. S was originally developed by Trevor Hastie, Richard A. Becker, Alan Wilks, John M. Chambers, and William S. Cleveland while they were working for AT&T. Becker Chamber, and Wilks provided the original description of the S language in 1988 [3]. A more recent tutorial on the use of S and S-PLUS® is contained in Krause and Olsen, 1997 [4].

S-PLUS® is manufactured and supported by the Insightful Corporation. R. Douglas Martin has been one of the visionaries who heavily influenced the development of S-PLUS®. He has continued his association with the Insightful Corporation but he has most recently been associated with the University of Washington and with FinAnalytica Inc. The history of the development of S-PLUS® has been characterized by contributions from such statistical luminaries as Rob Tibshirani, Jerome Friedman, Bill Venables and Brian Ripley.Much of the original S distribution is still available from http://lib.stat.cmu.edu.

S-PLUS® is available for both Windows and UNIX platforms. S-PLUS® provides the user with a fully extensible environment in that there is ready support for the user to develop their own functions using the S-PLUS® language. In addition S-PLUS® provides the user with the capability to call an S-PLUS® functions within C, C++, or JAVA and the capability to call a C, C++, or JAVA function from within the S-PLUS® environment.

Insightful also markets several add-ons to S-PLUS® and a few closely related products. Insightful claims that S-PLUS® contains over 4,200 data analysis functions which implement modern and robust statistical procedures. One of the strong points of S-PLUS® is its graphical capabilities. The graphics within S-PLUS® are "head and shoulders" above many of the the other statistical packages. The S-PLUS® based distribution comes well equipped to perform many of the standard statistical analysis including the ability to generate random data from many of the standard distributions, the ability to perform numerous hypothesis tests including Student's t-test and the Wilcoxon test. The package also allows the user to perform linear, nonlinear, and projection pursuit regression. In addition there are capabilities for multivariate analysis, discriminant analysis, and clustering. Standard time series analysis techniques are also supported within the S-PLUS® environment.

For those customers looking to use S-PLUS® to analyze large datasets and to conduct the analysis using moderately skilled technicians, Insightful offers Insightful Miner 3. This is a highly scalable workbench that allows non-technical users to deploy predictive intelligence systems throughout the business enterprise. Insightful Miner is used by such clients as the Bank of America and Pfizer.

The specialized S-PLUS® add-ons include S+ArrayAnalyzer™(a microarray analysis module), S+NuOPT™(a portfolio and general optimization package), EnvironmentalStats for S-PLUS (a set of functions for performing statistical and graphical analysis on environmental data), S+FinMetrics™(an economics analysis package), S+SeqTrial™(a package for designing, monitoring, and analyzing clinical trial using group sequential methods), S-PLUS® for ArcView GIS (a package for the integration of S-PLUS® and ArcView GIS 3.2) and S+SpatialStats™(a spatial statistics analysis package).

In closing we would like to point out that we expect that S-PLUS® will continue to receive a strong support from the academic and commercial sectors. There are of course subpopulations of users in particular discipline areas that are adamant about the benefits of S-PLUS. The user is referred to [5] for in-depth discussions with regards to S-PLUS®.

R

R originated at the University of Auckland, New Zealand in the early 1990's. Ross Ihaka and Robert Gentleman were looking for a statistical environment to use in their teaching lab. At the time, the labs were populated with Macintosh computers and there was no suitable statistical software available for the Macintosh environment. The two of them had some previous familiarity with S and so they decided to implement a new language that would be based on an S-like syntax. Initial versions of R were provided to statlib at Carnegie Mellon University and the preliminary user feedback indicated a positive reception for the language. This positive response encouraged them to release R under the Open source software initiative. The first version of R was released to the public in June of 1995. About the same time Martin Mächler came on board the R development team.

The Open source software license agreement allows users to modify the R software as long as their changes also fall within the Open source software license agreement. The Open source software paradigm had previously worked well for such software as LINUX and APACHE and this also proved to be the case for

R. A software system that exists under the Open source paradigm benefits from having "many pairs of eyes" to examine the software to help insure quality of the software. Support for R is provided via an on-line mailing list. This list allows users to pose questions to the R "collective consciousness." Often the users receive responses within a matter of minutes.

R developed a huge following and it soon became too difficult for the three of them, Ross, Robert, and Martin to keep up with the submitted software changes. This led them to select a Core group, in 1997, of around 10 members. This group was responsible for changes to the source code. John Chambers joined the Core team in 2000.

The R development paradigm has led to the rapid development of the R software system. R has been endorsed by many academic institutions as an alternative instructional language in several of their undergraduate courses. In addition they have also utilized it in their graduate and research programs, in part because of the ease of disseminating their developed prototype software to the research community at large.

As of this writing, R possesses an extensive statistical capability. R is organized into various packages. The base package contains support for 2-D and 3-D plots, including many of the more esoteric statistical plots such as star plots. The base package also includes sample, evaluation, cumulative distribution,and inverse cumulative distribution functions for many of the standard statistical distributions. There is also support for one and two sample hypothesis testing and categorical data analysis. Besides the base package there are numerous other packages including boot (bootstrap R and S-plus functions [6]), class (functions for classification), cluster (functions for clustering [7]), datasets (R datasets), exactRankTests (exact distribution for permutation and rank tests), foreign (read data from Mini-Tab, S, SAS, SPSS and others), graphics (T graphics package), grDevices (R graphics devices and support for colors and fonts), grid (grid graphics package), ISwR (package to accompany Introductory Statistics with R [8]), KernSmooth (functions functions for kernel smoothing for Wand and Jones [9]), lattice (Lattice Graphics), MASS (main package of Venables and Ripley's Modern Applied Statistics with S [10]), methods (formal methods and classes), mgcv (generalized additive models with generalized cross validation smoothness estimation and generalized additive mixture models by restricted maximum likelihood and penalized quassi-likelihood), nlme (linear and nonlinear mixed effects models), nnet(feed-forward neural networks and multinomial log-linear models), rpart (recursive partitioning), spatial (functions for kriging and point pattern analysis), splines (regression spline functions and classes), stats(the R stats package), stats4 (statistical functions using S4 classes), survival (survival analysis, including penalized likelihood), tcltl (TCL/Tk interface), tools (tools for package development), and utils (the R utils package).

There are also several major projects that are "R spin-offs." "Bioconductor" is an R package for gene expression analysis. "Omega" is an R package that is focused on providing seamless interface between R and a number of other languages including PERL, PYTHON, MATLAB, and numerous others. "gRaphical models" is an R package for graph theoretic modeling. "R GUISs" is a project for the development of a GUI-based R package. R, for the most part, is a command line based language. There is a small contingent that is interested in a fully GUI-based R language. “R spatial projects” is an ongoing effort to extend the spatial statistics capabilities that are resident in the R base distribution. The user is referred to [11] fro a more in-depth discussion of R and its related projects.

3 Current Analysis Packages with Statistical Libraries

MATLAB

MATLAB had its birth as a set of FORTRAN subroutines for solving linear (LINPACK) and eigenvalue (EISPACK) problems. These routines were developed primarily by Cleve Moler in the 1970’s. He, Moler, later went on to teach mathematics courses. During these courses he wanted his students to have ready access to LINPACK and EISPACK without needing to know FORTRAN. Hence he developed MATLAB as an interactive system to provide his students with access to LINPACK and EISPACK. Originally MATLAB gained much popularity primarily through word of mouth since it was not being officially distributed at the time. MATLAB was rewritten in C during the 1980’s. This rewrite provided the user with more functionality including routines for data plotting. The parent MATLAB company, the Mathworks, Inc., was created in 1984. The Mathworks, which is located in Natick Massachutes, is now responsible for the sale, development, and support of MATLAB.

MATLAB has grown to have a very large user-based in government, academia, and the private sector. The base distribution of MATLAB allows a user to read and write data in ASCII, binary, and MATLAB proprietary format. The data is presented to the user as an array, the MATLAB generic term for a matrix. A scaler is simply a 1 x 1 array and a n by m matrix is a n by m array. The base distribution comes with all of the expected mathematical functions including trigonometric, inverse trigonometric, hyperbolic, inverse hyperbolic, exponential ,and logarithmic. In addition, MATLAB provides the user with access to cell arrays, a data structure that can have different types of data types in each i, j entry of the array and structures which allow a user to create a simple data object in a manner that is analogous to a C or C++ data structure. MATLAB also provides the user with various numerical methods including optimization and quadrature functions.

One of the real strengths of MATLAB is its data plotting capabilities. MATLAB supports standard two and three dimensional scatter plots along with surface plots. In addition MATLAB provides the user with a graphics property editor. The graphics property editor allows the user to customize virtually every aspect of the appearance of a graph and it also serves as a framework for GUI development.

MATLAB, like R, is fully user extensible. The user is provided with an integrated text editor that serves as a debugger for code development. One can write either scripts or functions using the MATLAB language. Functions are automatically compiled prior to execution.

MATLAB supports many additional toolboxes including a statistics toolbox. The statistics toolbox supports 20 different probability distributions. There are five associated functions with each distribution including: a probability density function, a cumulative distribution function, an inverse cumulative distribution function, a random number generator, and the mean and variance as a function of the parameters. MATLAB also provides functions for computing confidence intervals and parameter estimates for many of these 20 functions. The toolbox also provides support for descriptive statistics, linear models, nonlinear models, hypothesis tests, multivariate statistics, statistical plots, statistical process control, design of experiments, and hidden Markov models. The toolbox has been through several iterations in its development cycle and it is rapidly evolving into a bona fide package in its own right.

Wendy and Angel Martinez have worked diligently over the past several years to extend the MATLAB statistical capabilities into the computational statistics realm. Their first book [12] helped bring MATLAB into the computational statistics arena and the accompanying software package was written to be mostly independent of the pre-existing MATLAB statistics package. This package provided capabilities for kernel-based probability density estimation, mixture-based density estimation, projection pursuit, along with many other more recent computationally intensive methods. Their second book [13] and associated package focuses on exploratory data analysis. This book provides MATLAB-based code for many of the standard statistical exploratory data analysis methods and graphics procedures. In addition, there is a greatly expanded dimensionality reduction section that includes MATLAB codes and discussions on the nonlinear isometric embedding (ISOMAP) dimensionality reduction procedure.

Mathematica

The driving force behind the development of MATHEMATICA has been Stephen Wolfram. His preeminent role in symbolic computing was foreshadowed by his receiving the MacArthur Prize in 1981. During the early 1980's he developed a language denoted SMP (Symbolic Manipulation Processor). SMP was written in C. In 1985 Wolfram extended SMP and renamed it Mathematica. Mathematica is most well known as a symbolic computation package but over recent years it has developed advanced numerical algorithms and sophisticated GUIs to facilitate the user interaction with the compute engine. Mathematica originally came bundled with the NeXT operating system. This certainly provided the NeXT machines with a high quality application to bundle with there initial product line.

Like MATLAB, Mathematica provides the user with an application programming interface (API) which enables it to call Fortran or C functions, and to also be callable from C and Fortran external programs. Mathematica can function in its own right as a programming language. In addition, Mathematica can function as a notebook document system which allows the user to keep track of the symbolic computations and Mathematica commands that led to a particular figure in a paper. It also provides a system for complex analysis including image and signal processing, a repository of reams of information on numerical methods and mathematics, and a bona fide system for numerical and symbolic computation. The Mathematica graphics techniques have grown over recent years to rival MATLAB and even surpass MATLAB in the area of graphical rendering of surfaces from their intrinsic equations.

Mathematica has recently provided users with statistical analysis tools. Mathematica provides tools for analysis of variance, classical hypothesis testing, confidence interval estimation, data smoothing, univariate and multivariate descriptive statistics, nonlinear and linear regression, and a large suite of powerful optimization techniques. Mathematica may fall a little short in its ability to perform some of the advanced statistical graphics procedures such as multipane graph linking but it makes up for this weakness with its keen capability to aid a user in arduous symbolic calculations that occur in mathematical statistics or in advanced computational statistics methodologies. The user is referred to [14] for an in-depth discussion on the uses of Mathematica in mathematical statistics.

4 Some General Languages with Statistical Libraries

In this next section we will briefly highlight the statistics libraries available with some of the more common software development languages.

JAVA

It is difficult to assess the state of the art with regards to JAVA statistical libraries in that there may be many custom user developed packages that we are unaware of. Given this caveat, we would like to mention three packages. The first package had its inception under the name WebStats. It was developed as an instructional statistical framework by Webster West of the University of South Carolina. WebStats has recently been named StatCrunch. StatCrunch 4.0 provides the user capability to perform interactive exploratory data analysis, logistic regression, nonparametric procedures, regression and regression diagnostics, and others. The user is referred to a recent review of StatCrunch 3.0 that appeared in the Journal of Statistical Software [15].

Another source of JAVA-based statistics functions is the Apache Software Foundation Jakarta math project. The math project seeks to provide common mathematical functionality to the JAVA user community. The initial code base will consist of the following statistical algorithms: simple univariate statistics (mean, standard deviation and confidence intervals), frequency distributions, t-test, chi-square test, random number generation from Gaussian, exponential, and Poisson distributions, random sampling/resampling, along with bivariate regression and correlation. The user is referred to the project's website for additional information [16].

The final source for JAVA-based statistical analysis that we will discuss is the Visual Numerics JSML$^{\mathrm{TM}}$package. The JSML$^{\mathrm{TM}}$package provides the user with an integrated set of statistical, visualization, data mining, neural network, and numerical packages. One would expect that the JSML$^{\mathrm{TM}}$package is of very high quality given Visual Numeric's previous releases, PV-WAVE$^{\mathrm{TM}}$and IMSL$^{\mathrm{TM}}$. The reader is referred to [17] for additional discussions on JMSL$^{\mathrm{TM}}$.

C++

We will only briefly mention two of the many C++ statistical libraries. The first one is the GNU Object-Oriented Statistics Environment (Goose). Goose is a C++ library that is dedicated to statistical computation. The Goose project has several goals including the creation of a useful and complete system that takes advantage of C++'s features to improve the clarity of statistical code and that is also easier for programmers to use. The other two main goals of Goose are to provide its functionality to Guile and Guppi two other ongoing GNU projects. Goose has been developed primarily under Linux but it has been designed to be compatible with other Unix and Win32 systems. Goose currently supports a standard set of statistical functions including cumulative and inverse cumulative distribution functions for many of the most common statistical distributions. In addition Goose possesses a high-quality Mersenne twistor-based random number generator. Goose also provides a RealSet class which is an optimized C++ container to hold statistical data. In addition Goose provides support for t-tests, F-tests, Kruskall-Wallis tests, Spearman tests, McNemar's test, and Cochran's Q test along with an implementation of simple linear regression models. The reader is referred to [18] for more in-depth discussions on the Goose project.

A statistics library product that may be of more interest to Microsoft Windows developers is the Probability and Statistics for .NET v3.3 by WebCab Components. This package actually consist of five packages: statistics, discrete probability, standard probability distributions, hypothesis testing and correlation and regression. Some of the capabilities offered by these components include standard measures of centrality and dispersion, discrete and continuous probability distribution functions including random sampling, cumulative distribution function, and inverse distribution function. Modules to perform various hypothesis tests and to compute confidence intervals are also provided. A strength of these modules are there support for various interfaces including C# and C++.NET. WebCab Component's website provides the user with extensive documentation on their package, see [19].

PERL

The Practical Extraction and Report language, PERL, has been embraced by a large group of users from various communities. Proponents of PERL include individuals in the system administration, computer security, bioinformatics, language processing, and financial communities. One can often write PERL code that it much shorter than equivalent code in other languages. PERL is particularly well known for its parsing and string matching capabilities. All of these reasons led us to include a brief mention of it within the current chapter. The comprehensive PERL archive network, CPAN) serves as a common repository for much of the user community developed PERL code. A simple search at this site reveals that there have been many different PERL statistics packages developed. These packages include some for simple descriptive statistics calculation, linear regression analysis, chi square testing, rank correlation computation, and t-test and dependent t-test. One can peruse the additional PERL functionality at the CPAN site [20].

5 The Future of Statistical Computing

We feel that there are numerous challenges and opportunities that will help shape the future of statistical computing. The first challenge is provided by the non-expert user. There are more and more individuals who wish to apply statistical techniques to analyze their financial, biological, chemical, astronomical or other type of dataset. These users usually have only a modicum of statistical training and it is easy for them to inadvertently perform erroneous analysis on their data. We believe that there remains much work to be done in providing the user with computer software based aids to oversee their application of statistical techniques. We believe that these aids will ultimately take the form of some sort of "expert system" or other artificially intelligent package and that these aids will help the common user avoid pitfalls that are often encountered by the novice user.

Novice users are not the only challenge to the future of statistical computing. The statistical computing community has previously encountered the challenge of "data in the wild" as part of their support of data mining efforts. By "data in the wild", we mean data that was collected under uncontrolled circumstances. This sort of situation greatly complicates the nature of the analysis of this type of data. We believe that this challenge will continue to exist and that it may even grow in coming years. In fact the evolution of this "data in the wild" can help drive the software development process on several fronts.

The first front is characterized by the need for the analysis of text data. Text data analysis can be necessitated by the needs of computer security, bioinformatics, automated scientific discovery, and intelligence research. Statistics has a role to play in the analysis of features that might semantically capture the meaning of a collection of text documents or as a means of analyzing simpler "word count histogram" descriptions of text documents. We have recently completed the first phase of a project to automate the discovery of serendipitous relationships between disparate discipline areas [21], [22]. It is our contention that the analysis of such data will play a more prominent role in today and tomorrow's document rich environment. We think that our position is well-founded based on the recent interest in text mining of several computer giants, Microsoft and Apple, and major statistical software companies, SAS$^{®}$ and SPSS.

The second front is characterized by streaming data. Streaming data can be thought of as a dataset that manifests itself over a period of time in a serial manner. Many sorts of data fall into the streaming category including video, computer security, and voice. Computer security data can take the form of session data, where a human operator interacts with the computer via an ongoing session, internet messenger chat data, simple packet exchange headers, and many other types. The streaming nature of the data would suggest the consideration of recursive algorithms. In fact one of the largest challenges of the streaming data type is that visualization of said data requires graphics techniques that evolve with the data. As the data changes its nature, dimensionality, type (continuous vs. discrete), the visualization software must adapt/evolve appropriately. See for example [23], [24]. for some recent discussions on these matters. Software developers and various government agencies are only now gearing up for concerted efforts in this area.

The third front is based on the need for statistical software in graphical/network based models. Network based models have recently received attention for their use in social network analysis (SNA). SNA and in particular dynamic SNA is an area which is of great interests not only to the sociology community but also to those individuals prosecuting the wars on drugs and terrorism. The application of statistics to SNA in general and dynamic SNA in particular is still in its early stages. It will be interesting to see which of the major statistical software companies come forward with a toolbox/package aimed at graph theoretic modeling and analysis. Right now, R and Mathematica have some preliminary offerings in this area.

6 Concluding Remarks

We have provided an "evaluation" of the current state of statistical software. Our treatment has been tightly focused based on our own experiences and prejudices in this area. We have, however, attempted to be "even handed" in our appraisal of the discussed software.

We have also attempted a limited prognostication with regard to the future of statistical software. We have attempted to identify those issues and those applications that are likely to shape the future of the area. We realize of course that the future is usually full of surprises. These surprises are what makes for an interesting journey as we go through life.

Acknowledgments

The work of the first author, Edward J. Wegman, was support by the Defense Advanced Research Projects Agency under Agreement 8905-48174 with The Johns Hopkins University. The work of the second author, Jeffrey L. Solka, was supported by the Office of Naval Research (ONR) In-house Laboratory Independent Research (ILIR) Program

References

[1] On-line News Published in DMRreview.com (2004). New SAS 9 Software Revolutionizes the BI Industry *www.dmreview.com/article_sub.cfm?articleId=1001037*.

[2] Hilbe, J. (2003). A Review of Concurrent SPSS Products SPSS 12, SigmaPlot 8.02, SigmaStat 3.0, Part 1. *The American Statistician,* Vol. 57, No. 4, 310–315.

[3] Becker, R. A., Chambers, J. M., and Wilks, A. R. (1988). *The New S Language.* Wadsworth and Brooks/Cole, Pacific Grove, CA.

[4] Krause, Andreas and Olson, Melvin (1997). *The Basics of S and S-Plus (Statistics and Computing)* Springer-Verlag.

[5] The Insightful Corporation (2004). The Insightful Webpage. *www.insightful.com*

[6] Davisin, A. C., Hinkley, D. V., and Gill, R. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press.

[7] Kaufman, Leonard, and Rousseeuw, Peter J.(1990). *Finding Groups in Data:An Introduction to Cluster Analysis.* Wiley-Interscience.

[8] Dalgaard, Peter(2002). *Introductory Statistics with R.* Springer-Verlag.

[9] Wand, M. P., Jones, M. C. (1994). *Kernel Smoothing.* Chapman and Hall/CRC.

[10] Venables, W. N., Ripley, Brian D. (2002). *Modern Applied Statistics with S.* Springer-Verlag.

[11] The R software compendium (2004). The Comprehensive R Archive Network. *cran-r-project.org*

[12] Martinez, Wendy L., Martinez, Angel R. (2001). *Computational Statistics Handbook with Matlab.* CRC Press.

[13] Martinez, Wendy L., Martinez, Angel R.. (2004). *Exploratory Data Analysis With Matlab (Computer Science and Data Analysis).* Chapman and Hall/CRC Press.

[14] Rose, Colin, Smith, Murray D. (2002). *Mathematical Statistics with MATHEMATICA.* Springer-Verlag.

[15] West, R. Webster, Wu, Y., and Heydt, D. (2004). An Introduction to StatCrunch 3.0. *Journal of Statistical Software, Volume 9, Issue 6.*

[16] The Apache Foundation (2004). The Jakarta Proposal for math Package. *jakarta.apache.org/commons/math/proposal.html*

[17] Visual Numerics Incorporated (2004). JMSL^{TM}Numerical Library for JAVA Applications. *www.vni.com/products/imsl/jmsl/jmsl.html*

[18] GNU Project of the Free Software Foundation (2004). JMSL^{TM}Goose: The GNU Object-Oriented Statistics Environment. *www.gnu.org/software/goose/goose.html*

[19] WebCab Components Website (2004). Probability and Statistics for .NET v 3.3. *www.webcabcomponents.com/dotNET/dotnet/pss/index.shtml*

[20] CPAN Website (2004). Categories >> Data Type Utilities >> Statistics. *cpan.uwinnipeg.ca/chapter/Data_Type_Utilities/Statistics*

[21] Solka, J. L., Wegman, Edward J., and Bryant Avory C. (2005). Identifying Cross Corpora Document Associations Via Minimal Spanning Trees. *Interface 2004:Computational Biology and Bioinformatics 36th Symposium on the Interface* to appear.

[22] Solka, J. L., Wegman, Edward J., and Bryant, Avory, C. (2005). *Text Data Mining With Minimal Spanning Trees.* in Handbook of Statistics, C. R. Rao, Edward J. Wegman, Jeffrey L. Solka (Eds). to appear.

[23] Wegman, Edward J.(2003). Evolutionary Graphics and Recursive Algorithms for Streaming Data. *Keynote Address, Conference on Applied Statistics in Ireland* Mullingar, Ireland.

[24] Wegman, Edward J.(2003). Evolutionary Graphics for Streaming Data. *Joint Statistical Meeting 2003* San Francisco, CA.

Further Reading

Cti Statistics (2004). Alphabetical list of reviews. *www.stats.gla.ac.uk/cti/activities/reviews/alphabet.html*

Jan de Leeuw (2004). Journal of Statistical Software. *www.jstatsoft.org*

Scientific Computing World (2004). Software Reviews from Scientific Computing World. *www.scientific-computing.com/reviews.html*

Related Entries: ESS2, LaTeXtemplate, Wiley