for submission to *Human and Ecological Risk Assessment*

Software Review of S-Plus, Version 4.0

David E. Burmaster
Alceon Corporation, POBox 382669, Cambridge, MA 02238-2669
tel: 617-864-4300; fax: 617-864-9954; email: deb@Alceon.com

Program:	S-Plus, Version 4.0
Source:	MathSoft, Inc.
	Data Analysis Products Division
	1700 Westlake Avenue North, Suite 500
	Seattle, WA 98109
	Tel: 800-569-0123; fax: 206-283-8691
	Email: mktg@statsci.com
	Web: http://www.mathsoft.com
System:	Windows 3.1x, Windows 95, or Windows NT
	80486 processor running at 66 MHz or higher,
	with 32 MB RAM and math coprocessor,
	80 MB available hard disk and a CD-ROM drive,
	VGA or higher graphics adapter
Price:	Prices available directly from the company

Abstract

S-Plus is a premier high-end tool for exploratory data analysis and statistical analysis of multivariate data. Based on the object-oriented S language first developed at AT&T's Bell Laboratories (now Lucent Technologies) as a flexible environment for data analysis, S-Plus has been one of the most popular tools for research scientists, statisticians, and sophisticated business analysts for a decade. With the release of Version 4.0, MathSoft has made the software much more accessible to a much wider audience by adding an intuitive and powerful graphical user interface (GUI) and many new features.

Features

Evolving as it has from the S language developed under UNIX at AT&T's Bell Laboratories, S-Plus has a clean architecture and a well deserved reputation for power and speed. Over the years, S-Plus's developers have pioneered many new methods in exploratory data analysis (EDA), an approach to graphical techniques and data visualization first championed by John W. Tukey of Princeton University and the original Bell Telephone Laboratories (see, Tukey, 1977).

Beyond the fabulous 2D and 3D plots and "smooths" now accessible by the click of a mouse on floating palettes, S-Plus has more features than most people will need in a career. For example, the "Guide to Statistical and Mathematical Analysis" has long sections devoted to the program's features for these and other topics: statistical inference (both one and two sample problems with robust and nonparametric tools); several types of goodness of fit tests; statistical inference for counts and proportions; cross-classified data and contingency tables; regression and smoothing, including multiple types of robust regression and smoothing; the generalized linear model; local regression models; classification and regression trees (CART); ANOVA and MANOVA; principal components analysis; factor analysis; cluster analysis; time-series analysis; survival analysis; the Cox proportional hazards model; quality control charts; linear algebra; and a object-oriented matrix library.

In Version 4, the developers give the user (i) complete control to customize the "Toolbar" with standard or custom routines/algorithms and (ii) complete control over every element of a graph, including points, lines, axes, labels, and legends. Now, the user may create and export PowerPoint® slides directly from S-Plus.

Performance

S-Plus is a compiled product; it runs fast! For example, it can "spin" a large 3D scatterplot in real time, even when several subsets of points are highlighted in

color and labeled with text. Several of the 2D and 3D "smoothers," e.g., the kernel smoother, the lowess regression, and the different smoothing splines, require heavy computation, but S-Plus does them in a flash. Similarly, S-Plus's speed makes CART methods (classification and regression trees) feasible for routine use.

Documentation

S-Plus comes with 9 volumes of documentation (weighing ~8 pounds): Read Me First; A Gentle Introduction; A Crash Course; User's Manual; Guide to Statistical & Mathematical Analysis; Trellis Graphics User's Manual; Programmer's Manual, with Supplement; and Global Index. Each volume gives clear presentations with sample problems keyed to data files distributed with the software. The style appeals to novice and to veteran users, and the style conveys advanced statistical and computational instructions without intimidating the reader. Throughout, the manuals cite books and articles that explain the methods, techniques, algorithms, or point of view. S-Plus also has context-sensitive help files available on-line at any time with the click of the mouse on an icon in a menu bar.

If you browse in a technical bookstore, you will find a number of books that illustrate or teach S-Plus. For example, Cleveland's two books (1993 and 1994) illustrate the power of S-Plus but do not teach the S-Plus language. People wanting more than the User's Manuals should consult one or more of the many books that teach S-Plus, e.g., Spector (1994), for an easy introduction to S-Plus; Venables and Ripley (1997), for a much longer and more comprehensive introduction to the software; Chambers and Hastie (1993), for more sophisticated techniques; and Bruce and Gao (1996), for an introduction to "wavelet theory" for signal analysis.

Ease of Learning and Ease of Use

A person comfortable with the Windows® operating system can easily work the examples (and data files) distributed with the software. Anyone with a semester or two of statistics can begin productive work in well under an hour. Now that users can access the program's features from menus, floating palettes, and dialog boxes, novices need not learn the S-Plus language to analyze a dataset and plot publication-quality graphs -- but the language is always there (in a separate Commands Window) for power users needing to write custom routines.

Error Handling

For this review, I ran a beta copy of Version 4.0 without a hitch and without a single crash. MathSoft will continue to refine the software for the commercial release.

Support

MathSoft provides technical support for the software and for some statistical questions, but the company cannot provide tutorials for people wanting to learn statistics from scratch. The company assumes, quite reasonably, that people who buy and use this product have had one or more semesters of probability and statistics in college or graduate school.

Value

Excellent. MathSoft charges a fair price for this sophisticated software. Granted, not everyone needs the power of S-Plus to analyze simple problems and small datasets. With the new graphical user interface (GUI), the new features, and the continuing features, the cost/benefit ratio for buying this software is very attractive.

An Example

Cleveland (1993; pp 272 - 292) analyzes air quality data measured on 111 days from May to September 1973 at meteorological stations in the New York City metropolitan region: ; (i) ozone concentration, measured in parts per billion (ppb), the averaged hourly

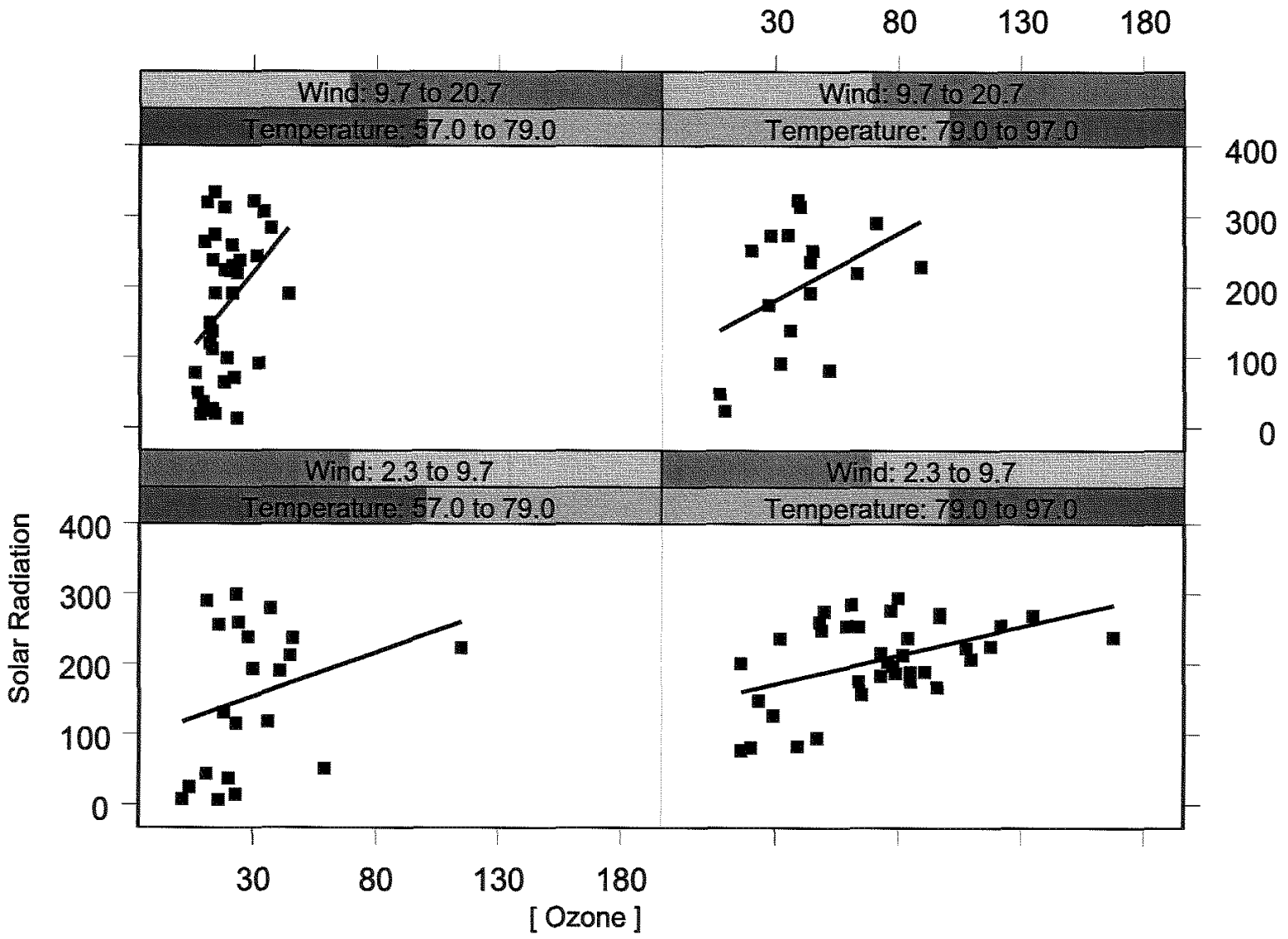

Figure 2: Trellis Graph of Solar Radiation against [Ozone]

values from 1300H to 1500H at LaGuardia Airport; (ii) solar radiation, measured in langleys, from 0800H to 1200H in the frequency band 4,000 - 7,700 Angstroms in Central Park; (iii) wind speed, measured in miles per hour (mph), the average of values at 0700H and 1000H at LaGuardia Airport; and (iv) temperature, measured in degrees Fahrenheit (degF), the daily maximum at LaGuardia Airport. In his book, Cleveland plotted [Ozone]$^{1/3}$ to make the empirical distribution more symmetric.

Without visualization, one cannot understand the highly nonlinear patterns in these four variables. Using S-Plus, one can quickly see the broad patterns in the data by making a bivariate scatterplot matrix and by "spinning the data" in various trivariate scatterplots. Upon first seeing these data plotted, I was struck by the large amount of scatter and asymmetry in each of the four variables and by the nonlinear relationships between pairs of variables.

Using "lowess regression," a standard routine in S-Plus V4.0 (and some other software packages), an analyst can quickly plot Figure 1 to see the highly nonlinear relationship between solar radiation and ozone concentration. The "lowess regression" line, really a spline of many short linear regressions (Cleveland 1993 and 1994), reveals the strong curvature.

Using the "trellis graphics" feature found only in S-Plus (as far as I know), an analyst can plot Figure 2 with a few keystrokes. Each panel shows the data points and best-fit linear regression (OLS regression) for solar radiation against ozone concentration as conditioned on a partition of the other two variables, wind speed and temperature. The lower left panel shows the data points and OLS regression line for slower winds and lower temperatures, while the upper right panel shows the data points for the faster winds and higher temperatures. The other two panels show the combinations as labeled. Now it is easy to see that slow winds and high temperatures have a strong adverse effect on air quality in the City.

Acknowledgments

Alceon Corporation funded this review.

Trademarks

References

Bruce & Gao, 1996
Bruce, A. and H-Y Gao, 1996, Applied Wavelet Analysis with S-Plus, Springer, New York, NY

Chambers & Hastie, 1993
Chambers, J.M. and T.J. Hastie, 1993, Statistical Models in S, Chapman & Hall, New York, NY

Cleveland, 1993
Cleveland, W.S., 1993, Visualizing Data, AT&T Bell Laboratories, Hobart Press, Summit, NJ

Cleveland, 1994
Cleveland, W.S., 1994, The Elements of Graphing Data, AT&T Bell Laboratories, Hobart Press, Summit, NJ

Spector, 1994
Spector, P., 1994, An Introduction to S and S-Plus, Duxbury Press, Belmont, CA

Tukey, 1977
Tukey, J.W., 1977, Exploratory Data Analysis, Addison-Wesley, Reading, MA

Venables & Ripley, 1997
Venables, W.N. and B.D. Ripley, 1997, Modern Applied Statistics with S-Plus, Second Edition, Springer-Verlag, New York, NY

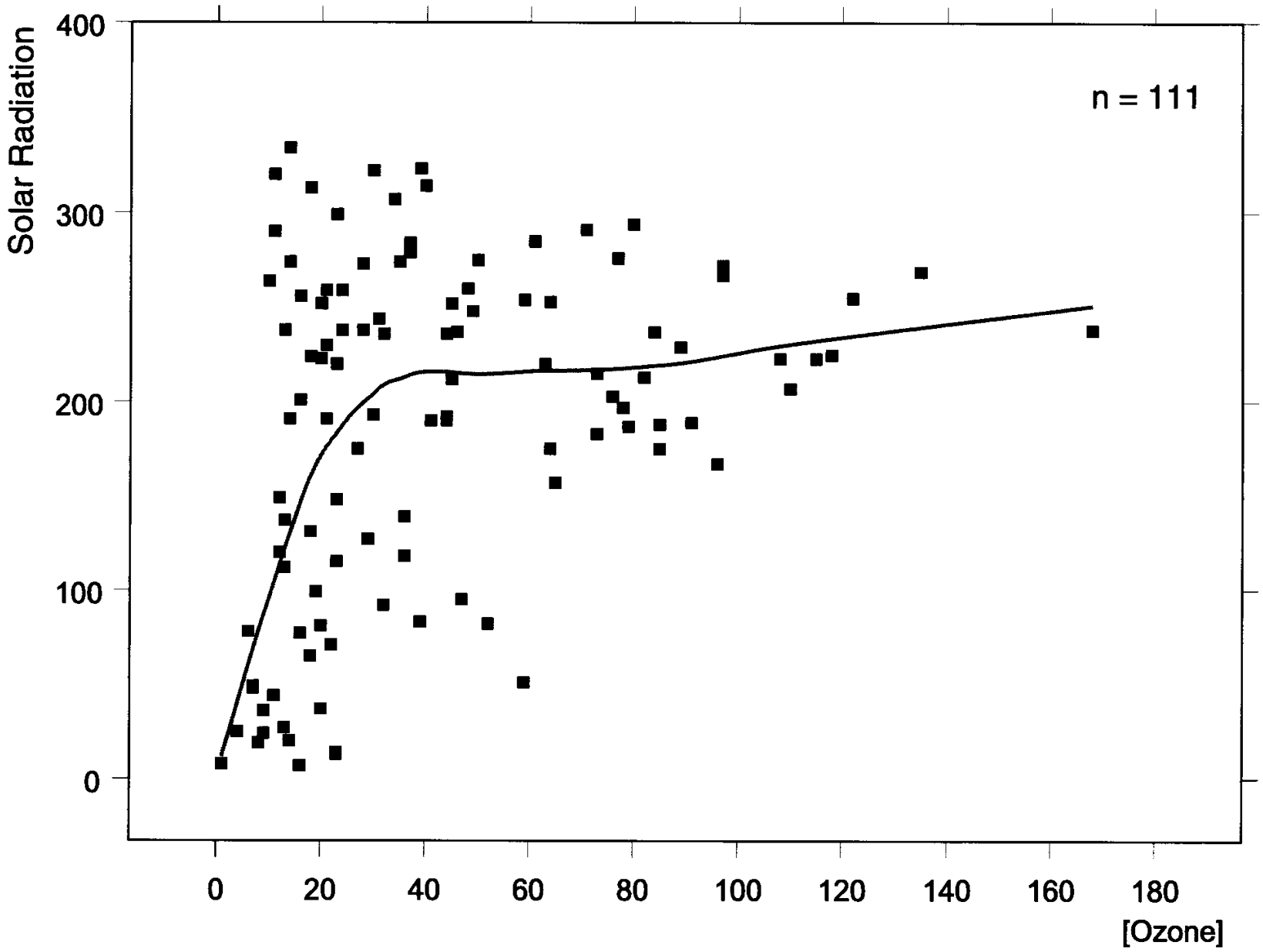

Figure 1: Lowess Regression for Solar Radiation against [Ozone]