A joint newsletter of the
Statistical Computing
& Statistical Graphics
Sections of the American
Statistical Association.

August 94 Vol.5 No.2

Statistical COMPUTING & GRAPHICS

A WORD FROM OUR CHAIRS

Statistical Computing

Trevor Hastie is the 1994 Chair of the Statistical Computing Section. Currently with Bell Laboratories, this fall he moves to Stanford University, Department of Statistics.

The Interface meeting was held in June this year at Research Triangle, NC, and was a big success. Congratulations to John Sall for a great program. For those who don't know, *Interface* refers to the interface of computing and statistics. This conference is a natural for members of our section, and indeed, I saw many of our members there, actively involved in the program. The 1995 Interface meeting is in Pittsburgh, PA, and is being organized by our newsletter editors Mike and Jim. In 1996 we take a dive, down under that is, to Sydney, Australia, where Nick Fisher is arranging the Interface program.

Our section's executive had a board (not bored!) meeting during the Interface, and probably the most exciting outcome to emerge is the inauguration of a special *Statistical Computing Student Paper Session* at the national meetings, starting in 1995. In fact, it will be a competition, with the session featuring the four winning papers amongst those submitted, plus a discussant. The winning student authors will have their conference expenses covered. This is one of several initiatives underway in the section consistent with our goals to promote statistical computing and applied statistics at colleges and universities. See the full announcement on page 3 for more details.

CONTINUED ON PAGE 3

A WORD FROM OUR CHAIRS

Statistical Graphics

Roy Welsch is the 1994 Chair of the Statistical Graphics Section. He and the editors encourage feedback on the issues, directly to the chair, or with letters to the editors.

Everyone is invited to the Statistical Computing Section and Section on Statistical Graphics Business Meeting and Mixer at 7:30 pm on Monday, August 15 at the ASA annual meetings in Toronto. Good food, good friends, and those famous (nearly random) drawings for a wide selection of items donated by friends of the sections. We need your input (and volunteers) so that the section can accomplish the goals you have (or will) set.

The Statistical Graphics Section program, organized by William DuMouchel, was listed in the last Newsletter and contains a wide variety of invited and contributed papers as well as roundtable lunches. Those of you with a special interest in the future directions of the section should attend the roundtable organized by Chair-elect David Scott called "Statistical Graphics Section: Spending Section Funds and New Initiatives" on Wednesday, August 17, 12:30 to 2. Due to the able administration of Secretary/Treasurer, Linda Clark, there will be funds available to spend on new initiatives.

The next Interface Symposium, to be held in Pittsburgh PA in April 1995 will place considerable emphasis on statistical graphics. You can influence the program for that meeting by attending the roundtable organized by James Rosenberger, "Statistical Analysis of Graphical Data: Interface '95" also on Wednesday, August 17, 12:30 to 2:00pm.

CONTINUED ON PAGE 2

FROM OUR CHAIRS (Cont.)...

Statistical Graphics

CONTINUED FROM PAGE 1

The program chair for the 1995 ASA annual meeting in Orlando is Sallie Keller-McNulty and she would welcome your ideas about the program for next year. She may be reached at 913-532-6883 or `sallie@cecil.stat.ksu.edu`. Most of the planning is done before and during this year's ASA meeting in Toronto so do not delay getting your suggestions to her.

One possible use of Section funds is to pay the expenses of speakers who would not normally attend ASA meetings. This would allow us to bring in people working in hardware, software, the "future" of graphics, and, most important of all, speakers from a wide variety of applications areas that have made exciting use of statistical graphics or generated new forms of statistical graphics. Sallie would like your suggestions since we want excellent speakers that one or more of our members have heard before and can recommend highly. The executive committee of the section will be addressing policy for the payment of outside speakers as well as the size of the budget for this purpose. The members of this committee have mixed opinions on both of these issues and the input of the section membership will be important in determining policy. It is your money and your section. Speak up!

A final reminder. The Statistical Graphics Section is co-sponsoring two short courses at the joint annual meetings, which come highly recommended. David W. Scott presents "Multivariate Density Estimation and Visual Clustering", a one day short course on Saturday, August 13. Susan Holmes presents "Exploring Multivariate Data with Modern Software: Examples in S", on Monday August 15.

See you in Toronto.

Roy Welsch
Chair, Statistical Graphics Section
`rwelsch@sloan.mit.edu`

EDITORIAL

This August 94 issue (Vol. 5, No. 2) of the Newsletter should arrive before the Summer Joint Statistical Meetings in Toronto. The third and final issue for 1994 should arrive at year end. The deadline for the final issue is October 15, 1994. Thanks again to all who have contributed.

The Statistical Computing Section of the ASA is proposing a student paper competition for the annual meetings next year, to be held in Orlando. Final details of this competition will be published in the December issue of this Newsletter, and in a Fall issue of Amstat News. Students and their advisors should plan on having their graduate students submit an abstract and paper on a topic of interest to the statistical computing section. The winners will receive registration and travel funds, and the rest will still have the opportunity to enter their paper as a contributed paper or poster session—winners all. See details in our chair's column and on page 3.

The statistical computing section executive has approved the use of color in all issues of the Newsletter. Wherever possible we encourage the submission of material which beneficially utilizes color, especially in the presentation of graphical results. The last issue included color plots in two articles, though we realized too late the compromise resulting from the size of the figures. We hope to improve the quality in future issues.

Our regular columnists have continued their suite of interesting articles. Phil Spector in his Unix Computing column writes about UNIX regular expressions; Mike Conlon describes the issues arising in selecting software in his Departmental Computing column; Dan Carr takes on greyscale after his colorful article in the last issue; and Mike Meyer describes both snooping the Internet with Mosaic and the WWW and provides the tools to contribute your own resources to the world of cyberspace.

We again wish to encourage submissions of articles to this newsletter. Letters to the editors, short informative pieces, information, and longer articles are all welcome. Send submissions and questions about appropriateness via e-mail to the editors. We encourage submissions in LaTeX or TeX, but will gladly accept plain ASCII text files as well.

James L. Rosenberger
Editor, Statistical Computing Section
`JLR@stat.psu.edu`

Mike M. Meyer
Editor, Statistical Graphics Section
`mikem@stat.cmu.edu`

FROM OUR CHAIRS (Cont.)*. . .*

Statistical Computing

CONTINUED FROM PAGE 1

For the remainder of this article, let me tell you a little about the section, our executive, and more importantly, *what we do for you*. Ok, ok, we do spend some of your dues on free lunches when we have our board meetings, but hopefully after you read this you will forgive us this indiscretion. In a nutshell, our main functions are:

- Arranging our section's share of the ASA program,
- creating and sponsoring continuing education events (short courses, workshops, tutorials),
- Finding creative and productive ways to spend our income (from dues and proceedings sales); our student paper competition is an example,
- liaising with the ASA as a whole (i.e. defending our turf), and
- producing this newsletter.

Our current program chair, Sallie Keller-McNulty, has arranged all the invited and contributed paper sessions sponsored by our section for the coming meetings in Toronto. This is truly a mammoth task (I know—I did it once), for which she gets a big round of applause. Sallie had to select six invited paper session organizers, who in turn invited speakers for their sessions. She had to keep track of all the invited speakers and make sure they stuck to deadlines regarding abstracts, registration, and so on. She had to attend a very long program chair's meeting in Virginia, where she successfully fought for an additional invited paper session. Then in a very rushed few days she had to mash all the contributed papers together into a coherent set of sessions. As a warm up for the same job the next year, our program chair-elect, John Rice, arranged the section's roundtable luncheons. So when you see something like "sponsored by statistical computing" or "SC" associated with a session, you know its one of ours (so go to it!)

I am sure you are aware of the short-courses and tutorials that take place before the national meetings and sometimes the Interface meeting. For many years now Tom Devlin has been our representative, soliciting course proposals, arranging "scholarships" for deserving attendees, then coordinating and arranging the courses down to the finest details (e.g. holding dry-runs with the presenters!). Most recently our section cosponsored and Tom managed a short course at the Interface meeting, "Modern Regression and Classification", given by Rob Tibshirani and some sidekick (actually, yours truly!). Tom was there, collecting tickets, handing out course-notes, collecting evaluations, etc. We are very grateful to Tom for this service. He is always keen to hear of new proposals and suggestions for topics and presenters, both from within our membership, and also from other societies.

Jim Rosenberger is our newsletter editor, together with Mike Meyer from the Graphics section. Can you imagine how much work that is, and it's all voluntary! Again, enormous thanks to both of them for the terrific job they are doing. In fact I heard rumblings from them that they are due for parole soon, so we hope to hear from (but more likely of!) qualified volunteers.

Look out for members of the Statistical Computing executive at the ASA in Toronto—our names were listed in the last newsletter. Introduce yourselves, and share any suggestions you may have with us. We hope to see you all at our joint mixer on Monday evening, August 15.

Trevor Hastie
Chair, Statistical Computing Section
`trevor@research.att.com`

Student Paper Competition

The Statistical Computing Section of the ASA will sponsor a Student Paper Session next year at the Orlando Joint Statistical Meetings in 1995. The topic of the session will (naturally) be *Statistical Computing*. Three to four students will be selected to participate in this session, which will include a discussant nominated by the selection committee. Fees associated with registration, accommodation and travel to the conference will be awarded to the participants in this Session.

Call for Papers

Students at all levels (undergraduate, Masters, and Ph.D.) are encouraged to participate. To be eligible, an applicant must be a registered student in the fall of 1994. The applicant must be the first author of the paper.

To be considered for selection in the session, students must submit an abstract, a six page manuscript, a resume, and a letter of recommendation from a mentor familiar with their work. The manuscript should be single-spaced in a 10 point font with one inch margins (this is consistent with ASA's Proceedings guidelines.) In the case of joint authorships, the mentor should indicate what fraction of the contribution is attributable to the applicant.

All application materials MUST BE RECEIVED by January 10, 1995. They will be reviewed by the Statistical Computing Section Student Paper Competition Award committee. The topic of the paper should be in the area of statistical computing, and might be original methodological research, some novel application, or any other suitable contribution (for example, a software related project). Selection will be based on a variety of criteria at the discretion of the selection committee, and will include novelty and significance of contribution, amongst others.

Award announcements will be made January 23, 1995. The selection committee's decision will be final. There will be a discretionary cap of $1000 on any given award, but it is anticipated that this figure should be more than sufficient to cover the expenses.

Students not selected for inclusion in the Session may submit their abstract and a registration fee to ASA by February 1 if they plan to attend the Joint Meetings. Those abstracts must be submitted following the ASA abstract submission instructions described in AMSTAT News. Students selected for inclusion in the session will receive further information about abstract submission and fee waivers from the award committee.

Inquires and materials should be emailed or mailed to either one of the following:

Trevor Hastie, Statistics Department, Sequoia Hall, Stanford CA 94305.
`trevor@playfair.stanford.edu`

Daryl Pregibon, room 2C264, AT&T Bell Laboratories, 600 Mountain Avenue Murray Hill, NJ 07974.
`daryl@research.att.com`

All electronic submissions of papers should be in postscript.

Continuing Education in Statistical Computing

by Thomas F. Devlin
Statistical Computing CE Chair

Short Course Proposals Sought

Proposals for short courses on topics in statistical computing or statistical graphics are invited for presentation at the Continuing Education Programs of the 1995 Joint Statistical Meetings in Orlando, Interface 95 in Pittsburgh, or another professional society meeting. Courses may be either one-day (6 hours of instruction) or half-day (3 hours) in length. They may be text-based, but that is not a requirement.

In addition to a modest honorarium and travel expenses, offering a course provides: professional recognition for yourself and the Section, an opportunity to share new ideas with colleagues, exposure to a wide audience of statisticians, a service to the profession and the Section, and an opportunity to promote statistical education and statistical thinking.

1995 Joint Statistical Meetings

Proposals for 1995 JSM are being accepted through September 30, 1994. Proposals need to follow Guidelines established by the ASA Advisory Committee on CE. The Guidelines are available for anonymous ftp from `mozart.montclair.edu` in the directory `/pub/asascs` and from the sections' CE chairs. Proposals, inquiries and suggestions should be sent to either:

Thomas F. Devlin
CE Chair, Statistical Computing Section
Mathematics & Computer Science Dept.
Montclair State University
U. Montclair, NJ 07043
Voice: 201-655-7244 `devlin@mozart.montclair.edu`

Abbe Herzig
CE Chair, Statistical Graphics Section
Consumers Union
101 Truman Avenue
Yonkers, NY 10703-1057
Voice: 914-378-2308 Fax: 914-378-2908

Interface '95 or other Professional Society Meeting

For more information about Continuing Education at Interface '95 or if you would like to suggest a course for presentation at a meeting of another professional society, contact Tom Devlin.

Thomas F. Devlin
Montclair State College
201-655-7244
`devlin@mozart.montclair.edu`

Current Index to Statistics/Extended Database

by Ron Thisted

The 1994 edition of the Current Index to Statistics/Extended Database (CIS/ED) will be available in early September. This edition incorporates thousands of corrections and greatly expanded content. Once again, the database itself will be available both on CD-ROM and on MS-DOS diskettes. (Additional files too large to include on diskettes, such as inverted indexes and software that uses them, will be provided on the CD-ROM edition.)

1994 Edition Extensions

The new edition covers the literature indexed in the printed volume of CIS from 1975 through 1993. In addition, coverage has expanded to extend the database to major journals prior to 1975. Journals now covered from their initial volumes include the Annals of Probability, the Annals of Statistics, Technometrics, The American Statistician, Advances in Applied Probability, the Australian Journal of Statistics, the Canadian Journal of Statistics, the Journal of Applied Probability, and the Zeitschrift fuer Wahrscheinlichkeitstheorie. Coverage of the Annals of Mathematical Statistics, JASA, Biometrics, Biometrika, Applied Statistics, JRSS (Series A and B), and Sankhya (Series A and B) now extend back to 1965 or earlier. Records that were missing from the earliest years of CIS have also been added.

In response to many requests, author names now appear in reversed order, that is, with family name preceding given names or initials. This feature will make both searching and correct alphabetization of records easier to accomplish. To make life easier for authors of search software, the identifying field (Field 1) for each record has been made unique, and contains additional information. This may result in increased availability of publicly-available software for the database.

A systematic effort has been mounted to improve the use of TeX codes in the mathematical portions of titles and key-words. The User Guide, which has expanded considerably, will contain a section on searching for Russian authors whose transliterated names can cause problems for searchers (such as Tchebysheff=Chebyshev).

Contributed search software for IBM and Macintosh users of the CD-ROM will join the contributed software packages for Unix systems introduced last year. Although not very polished, these programs may make it easier for some users to do CIS searches more routinely. (These freeware or shareware programs are not part of CIS/ED, but are included on the CD-ROM as a convenience for users.)

Five license arrangements are offered in 1994: Commercial, General, Academic, and Four-station licenses are available to organizations, and a Personal License is available at greatly reduced price to individual members of the ASA or IMS (which sponsor the Current Index). For licensing and ordering information, contact Richard Foley at the ASA Office `richard@asa.mhs.compuserve.com`, Fax: (703)-684-2037 or Barbara Lindeman at the IMS Office.

Ronald Thisted
Editor
Current Index to Statistics Extended Database
The University of Chicago
`cised@galton.uchicago.edu`

DEPARTMENTAL COMPUTING

Choosing Software

by Michael Conlon

In the previous column, we explored issues of uniformity—having all users in the department use the same computing platform. The issue was considered primarily from the hardware/operating system side—users of Macs, PCs and UNIX systems are each fervent in the support of their platform.

The issue of uniformity is related to the issue of software choice. If the platforms are uniform but the software is not, then the goals of platform uniformity—reduced training, interoperability—may not be met. A uniform environment may extend to application software as well as hardware.

Choosing for yourself

Let's start by considering the lone user. Is there such a thing? We are considering a user whose machine not on a network and not associated with other machines in another location. If one has a home computer, presumably that computer's files must interoperate with machines at work. I personally have ducked all that by having an Xterminal at home connected via a serial line and Xremote software to my UNIX environment at work. That way, I have "nothing" at home, no files, no software, nothing to move back and forth, nothing to maintain at home. Minimal investment ($700 terminal, $250 modem) and whenever things are upgraded at work, I receive the benefit at home. Most home computer users are somewhat bound by what they have in the office, since files must move back and forth.

But suppose you have no constraints. Perhaps you are a one person consulting outfit working from your home (so you don't have two locations).

You choose software in terms of cost/benefit. How much does the software cost in initial dollars, maintenance, disk space, learning time. What benefit is returned in terms of productivity—professional, correct, salable results.

Of the factors on the cost side, the key may be learning time. You must be productive and you must be productive soon. People dread learning new systems. And most people resist learning. This has always struck me as odd, particularly in the academic community, where people have entered the profession presumably because of a love of learning. It's interesting to see how focused people can remain—I want to learn X, not Y. And Y usually includes their software tools.

> ***Many people confuse the issue of learning, with the issue of productivity. They claim that the most productive tool is the one they know,...***

But many people overestimate learning time. They believe everything is as hard to learn as FORTRAN was back when they were in school. The arcane syntax, the problems with subscripting, formatting output, job control language. Let's face it, that was all pretty tough. Most systems now are much easier to learn than that. We have on-line help systems, graphical front ends, better manuals, actual books written by people who can write. It's all quite a bit more civilized than back in the days of punched cards and vendor manuals.

The most important issue regarding the benefit of using software, perhaps the only true benefit of using software at all is productivity. It's faster for the computer to do calculations. The laser printer does a better job of showing output. Graphics software can enable data analytic insights in moments that would have taken weeks to discover with traditional "output."

Many people confuse the issue of learning, with the issue of productivity. They claim that the most productive tool is the one they know, because they won't "lose time" learning something new. Stop for a minute and think about what that attitude would have meant if applied broadly in the twentieth century. I'm too busy with carbon paper to learn how to use the xerox machine. I'm too busy with my calculator to learn how to use the computer. I'm too busy with software system X to learn software system Y.

How to choose

So, in general, it makes sense to seize opportunities to use more productive software. The real issue is how to know if software Y will be more productive than your current software. If it takes you a long time to learn to use software Y productively, you may have lost something in the short run. How can you know if you will win in the long run? One way is to choose software you won't outgrow. That way, you have a shot at the long run.

Choosing software means judging software. Often people refer to feature lists. Like a car advertisement, a feature list runs down a list of things you might want to have. But like buying a car, you have to drive it. A car may list "air conditioning" but does it work the way you want? A software feature list may list "Postscript output" but how how well does it work? How does one choose software that works well and you won't outgrow?

- Read reviews. Judging software is tough. You can't just look at it, walk around it, flip through the manual. Software can be "opaque"—it's hard to see obvious things that may be missing.
- Try the software. Get a trial version (not a crippled demo). Or better yet, try the software actually installed on someone else's system. Look for opportunities to learn and use the software productively.
- Follow the market. Why are some software systems popular? Often because they are cost effective.
- Judge support. Software should be enabled with support mechanisms both formal (training, manuals) and informal (network discussion groups, email lists).
- Consider features. Feature lists can be revealing. They typically indicate a cultural bias toward one or another type of computing. The two features I typically look for (read on) are rarely on lists.
- Consider interoperability. A software system does not exist in a vacuum. And despite vendors' fantasies, your software must get along with other software on your system and other systems. We

must get data into vendor specific formats and get data from vendor specific formats into other systems. These can be non-trivial, non-productive tasks. Good software anticipates the need for interoperability for input (data and “programs”) and output (graphics, listings). Reading specific other vendors’ formats is often a small step in the right direction. A better approach is general input/output facilities to industry standard formats.

- Consider extensibility. How are the features of the software extended by the user? No feature list is complete. All software must be extended. How does one extend the software in question? An email program must be taught how to sort incoming mail the way you like it. An editor must be taught how to format based on types of documents, a word processor must have templates, a statistical software system must provide a facility for extending its stock analytic tools. Of all the features I look for, this may be the one I spend the most time considering.

Why extensibility is key

It seems paradoxical that extensibility might be the key feature in a software tool. After all, aren’t we trying to find something that will actually do the work, rather than find something that can be taught how to do the work?

The answer is that the work changes. Everything changes.

We need extensible software to respond to local changes. Our data may be collected in a new way and we then need to modify our systems so they adapt. We can’t anticipate the formats that will be used in the future. We need to have software that can be modified and extended to meet future needs.

We need extensible software to respond to global changes. Advances in the field of statistics change the definition of what constitutes a good analysis. Rapid changes in graphics and particularly dynamic graphics enable us to see things we missed previously. Software must be capable of rapid response to new advances. We cannot wait years for vendors to provide us with new releases of software to provide us with last year’s advances.

> ***We need to have software that can be modified and extended to meet future needs.***

We need extensibility to escape the “tyranny of the vendor”. In the old days, we used to call it the “tyranny of the programmer”—we got what the programmer was ready to give us. Many of us learned to program as a result. We shouldn’t be bound by a vendor’s feature list to the choices that have been provided for us, no matter how attractive they may look during a tour. There are many truly wonderful features in modern software systems today. There are many more waiting to be developed.

We should certainly encourage rich feature sets. We should also encourage seamless, thoughtful and productive extensibility. We should choose software with both. We need software we can learn, be productive with, and not easily outgrow.

Michael Conlon
Department of Statistics
Box 100212 HSC
University of Florida
Gainesville, FL 32610
`mconlon@stat.ufl.edu`
`http://www.clas.ufl.edu/~mconlon`

GEOGRAPHIC INFORMATION SYSTEMS

Gap Analysis, Biodiversity, and GIS

by Mark Monmonier

This column describes the use of geographic information systems to protect biodiversity. The application is called *gap analysis* because it attempts to identify "gaps" in the ecological safety net maintained by the National Park Service and the U.S. Fish and Wildlife Service, their state-level counterparts, and private-sector organizations like the Nature Conservancy.

As a GIS application, gap analysis seems to be a straightforward extension of map overlay. An environmental scientist using commercial GIS software begins by compiling “coverages”—electronic layers in a massive cartographic sandwich—representing (1) land in various parks and protected wilderness areas and (2) the ranges (habitats) of the region’s noteworthy flora and fauna. Overlaying these two types of coverage allows the scientist to search for unprotected places rich in endangered species—in other words, “gaps” in biodiversity protection. Identification of gaps is important if biological protection is to focus on habitats rather than a small number of charismatic species.

The algorithm seeks to maximize the number of previously unprotected species by identifying the area with the greatest number of unprotected species.

Because funds for biodiversity protection are limited, the analyst might want to experiment with an operations research strategy for identifying a complementary set of *N* unprotected habitats. One promising optimization strategy is the “greedy-add” algorithm, which can identify an optimally efficient ordered set of gaps. The algorithm, which seeks to maximize the number of previously unprotected species, begins by identifying the area with the greatest number of unprotected species. The process then identifies a second gap that adds more new members than any other area to its list of newly protected species. Similarly, in selecting the third-ranked gap, the algorithm considers only species not in one of the first two gaps or an existing protected area. An so on until *N* areas have been identified.

Gap Analysis

As with other stepwise optimization techniques for finding multi-location solutions, a simple greedy-add algorithm need not find the best, "globally optimum" solution. Consider, for example, a situation in which gaps A, B, and C would constitute the best possible three-gap solution by extending protection, respectively, to 19, 12, and 10 new species, for a total of 41. But this globally optimal three-gap solution would be overlooked if the algorithm found area D with 20 new species in its first round. If three of these species are in areas B and C, the algorithm might then be forced to choose areas E and F, which add only 9 and 7 new species for a total of only 36. If the number of gaps *N* is large, optimization schemes that reevaluate previous selections are very computationally intensive.

Optimality assumes, of course, that all *N* gaps will be protected. A wildlife agency unable to acquire the *i*th gap in an *N*-gap solution should reevaluate its remaining *N* - *i* selections if optimality is essential.

A further complication for gap analysis is the questionable reliability of existing species-range maps—often based on wildly optimistic extrapolations of a limited number of sightings, some of which are not at all recent. Gap analysis and the greedy-add algorithm will yield a solution, but because of uncertainty in the data, the validity of that result is questionable. GIS software must accommodate species-range maps annotated with sighting dates, qualified interpretations, and dubious yet potentially useful information.

Although gap analysts need not employ sophisticated spatial-search techniques, they cannot ignore the issue of data quality. Uncertainty calls for data entries coded to reflect relative reliability as well as highly interactive software with which scientists and planners can assess the stability of trial solutions. Equally important are graphic symbols that make apparent the uncertainty of a range, boundary, or location.

Another challenge is the limitation of regional gap analyses, which ignore protective efforts just beyond a state or administrative region. An optimum set of gaps for southern Idaho, say, need not represent an efficient national investment if there is substantial redundancy in adjacent portions of Oregon, Nevada, and Montana.

Equally important are graphic symbols that make apparent the uncertainty of a range, boundary, or location.

Also important is the issue of relative weighting. Are all species equally worth protecting? If not, how much more clout should a large mammal have in comparison to, for example, a comparatively mundane beetle? Although biodiversity protection must recognize the value of all species, for a variety of ecological and aesthetic reasons some groups would warrant more concern than others.

Despite these challenges, gap analysis is a promising GIS application. If Secretary of Interior Bruce Babbitt honors his commitment to the new National Biological Survey, there is little doubt that new research initiatives will seek improved ways of collecting, coding, analyzing, and displaying data on species and habitats.

References

Machlis, G. E. (1992), “The contribution of sociology to biodiversity research and management,” *Biological Conservation*, 62 (3), 161-170.

Scott, J. M., F. Davis, and B. Csuti. (1993), “Gap analysis: a geographic approach to protection of biological diversity,” *Journal of Wildlife Management*, 57, (supplement no. 123), 1-41.

Pennisi, E. (1993), “Filling in the gaps,” *Science News*, 144 (16), 248-251.

Mark Monmonier
Geography Department
Maxwell School of Citizenship and Public Affairs
Syracuse University
`mon2ier@mailbox.syr.edu`

BITS FROM THE PITS

Uncertainty and Sensitivity Analyses for Deterministic Models

by Jeremy C. York

This column features statistical computing and statistical graphics activities in science and industry. Your comments and suggestions for future columns are requested. Please send comments, inquiries, and suggestions to the editors or to Albert M. Liebetrau, Analytic Sciences Department, Battelle-Northwest, MS K7-34, P.O. Box 999, Richland, WA 99351, AM_Liebetrau@pnlg.pnl.gov*, 509-375-2694.*

Introduction

As part of the Manhattan Project (the U.S. Government World War II program that developed the first atomic weapons), facilities to produce and process nuclear fuels were constructed by the Army Corps of Engineers in an area of Washington state that became known as the Hanford site. For more than 40 years, these facilities were used for producing weapons grade plutonium and related activities.

One result of these activities was the release of various radioactive isotopes into the environment. Concerns about public health led the U.S. Department of Energy (DOE) to initiate the Hanford Environmental Dose Reconstruction (HEDR) project in 1987, under the direction of an independent 18-member Technical Steering Panel. Responsibility for this work was transfered to the U.S. Centers for Disease Control in 1992. The technical work was conducted by Battelle Pacific Northwest Laboratories.

The goal of the project was to estimate the individual dose received by people in the areas surrounding the Hanford site, from 1944 to 1972. A series of deterministic computer models and Monte Carlo simulations were used to combine historical information such as plant operating history, weather patterns and lifestyles of area residents with expert knowledge on transport mechanisms, absorption rates, and so on.

Because of the implications the findings of such a study might have, it was essential to assess the uncertainties associated with the estimated doses. The deterministic modeling, and the uncertainty and sensitivity analyses conducted, are summarized in this article.

The Models

The processes being modeled are quite complex. Radioactive materials from the Hanford site were emitted into the air and the Columbia River. People were exposed to airborne radionuclides by various pathways including inhalation and ingestion. For the sake of brevity, exposure routes associated with the river will not be discussed here.

The HEDR models traced radioactive materials from their source, through a variety of transport mechanisms, to eventual exposure and dosage to humans (Shipler and Napier, 1994). Following is a list of models used for the airborne radionuclides.

- Source Term Release model (STRM)
  This model used the operating histories of the Hanford nuclear reactors and fuel processing plants to estimate amounts of radionuclides that were released into the air and water.
- Atmospheric dispersion (RATCHET)
  This model used data on weather patterns and the outputs of STRM to describe the dispersion of radionuclides in the air and the resulting ground depositions.
- Environmental accumulation (DESCARTES)
  This model used the outputs of the RATCHET model to estimate the resulting concentrations of radionuclides in soil, vegetation, and animal products (milk, beef, poultry, etc.) Other inputs to this model included information on farming practices and the rates at which animals and plants absorb nuclear materials.
- Individual dose calculations (CIDER)
  This model was used to combine all the avenues of exposure with exposure scenarios to estimate the radionuclide dose to typical individuals.

This series of integrated models takes input parameters that describe everything from nuclear facility yields to cows' appetites and translates them into the amounts of airborne radionuclides transmitted through the various exposure pathways (Farris *et al*, 1994). That information then had to be translated into yearly and total dose estimates for persons living in various parts of a large region. For the purpose of estimating these doses, a set of representative individuals (RI's) was selected. The characteristics of these persons (age, home, diet, time spent indoors, etc) were intended to approximate those of selected segments of the population.

Uncertainty and Sensitivity

The models described above are deterministic, but knowledge about many of the input parameters is in-complete. The resulting uncertainty in estimated dose was assessed by Monte Carlo simulations in which the uncertainty about each input was described by a probability density function. Also, the sensitivity of the estimated dose to the uncertain input parameters was investigated by using regression models on the results of the Monte Carlo runs (see Simpson and Ramsdell (1993), Farris *et al* (1994)).

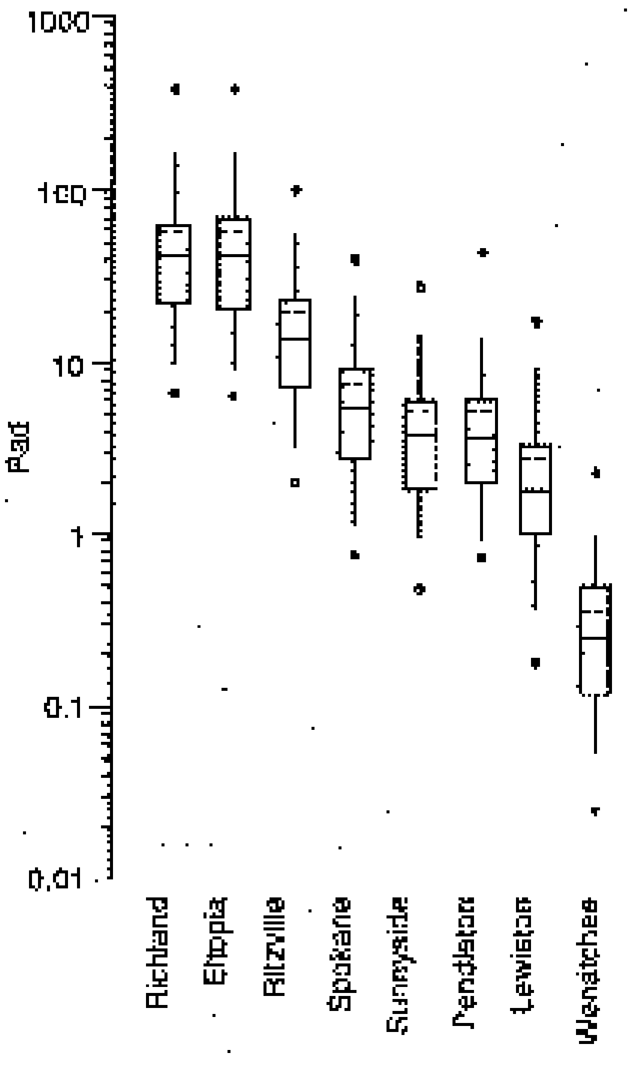


Figure 1: Thyroid doses for child raised on milk from a pasture–fed cow, 1945; adapted from Figure 5.2, Farris *et al* (1994).

Uncertainty Analysis

Selection of the distributions for the input parameters involved first determining the maximum and minimum values possible for that parameter, and then constructing a distribution over the resulting interval. The initial ranges were based on a literature review. The distributions were then selected in much the same way a Bayesian might select informative priors, based on expert opinion and historical data.

Once these distributions were determined, the Monte Carlo simulation was conducted. A complete set of input values was drawn from the distributions described above and fed into the HEDR models, resulting in one realized dose value for an RI (e.g., a child in the town of Eltopia who was fed milk from a meadow grazing family cow). This process was repeated 100 times, resulting in 100 realizations of the dose. The resulting sample was displayed with boxplots, as in Figure 1. In many cases the 95th percentile is an order of magnitude larger than the 5th percentile.

Sensitivity Analysis

Since there are many input parameters, with differing uncertainties, that affect the dose estimates via complex models, determining which parameters have the greatest impact is no easy affair. In the HEDR project, sensitivity was assessed by a hierarchical multiple regression model selection procedure. The response variable is the dose for a particular RI, and the predictor variables are inputs to one of the deterministic models described in section 2. All of these regressions were done using the ranks of the variables instead of their actual values. This approach made the procedure more reliable in the presence of non-linear relationships.

Initially, the doses were regressed on the inputs to the last of the series of models (CIDER), to determine which of those inputs were the most influential. Many of the inputs to CIDER are output from previous models, so similar analyses were performed to determine which inputs to the previous models had a major influence on these CIDER inputs.

This process was continued, tracing influential parameters backward through the layers of models to determine which of the raw inputs had the greatest impact. At this point, one final regression was conducted on all the influential inputs to summarize their impacts.

The regression results were summarized in pie charts that represent the proportion of the variability in an output due to uncertainty in an input. The area of the wedges is proportional to the coefficients of determination calculated in the regression. One such plot is reproduced in Figure 2.

The single most important parameter for nearly all of the RI's is the conversion factor from the amount of radioiodine ingested to individual thyroid radiation dose.

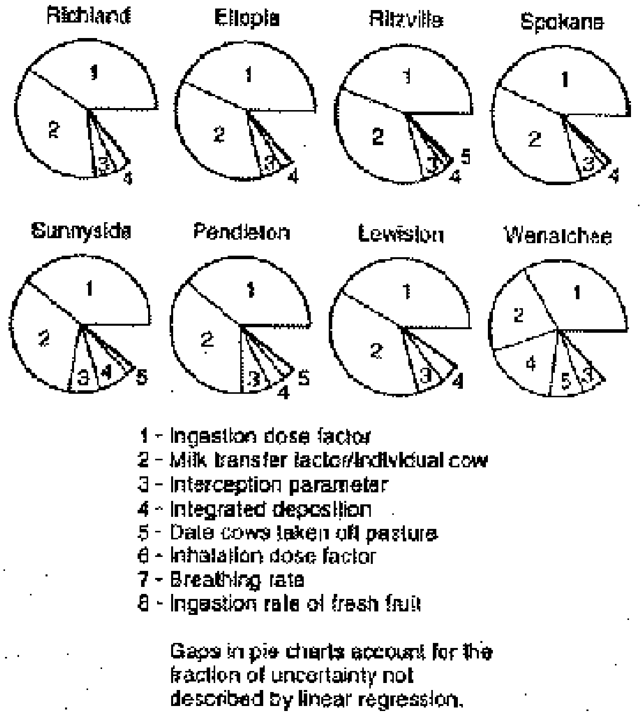


Figure 2: Relative importance of parameters for Iodine-131 dose to the Thyroid of a child raised on milk from a pasture–fed cow, 1945; adapted from Figure 5.5, Farris *et al* (1994).

Other parameters which may be important, depending on the RI, are the rate at which a cow will transfer ingested radionuclides into her milk, the rate at which a human will absorb inhaled radioactive particles and the rate at which airborne particles will settle to the ground.

Conclusions

The modeling of radionuclide transport, exposure and dose in a wide number of scenarios over a span of 50 years across a large geographical area was an ambitious effort. The careful analyses of uncertainty and sensitivity made it possible to produce useful and meaningful results. Models are currently being modified to provide dose estimates and uncertainty assessments for specific individuals whose characteristics differ from those of the RI's defined for the HEDR project.

References

Farris, W.T., Napier, B.A., Ikenberry, T.A., Simpson, J.C. and Shipler, D.B. (1994), "Atmospheric Pathway Dosimetry Report, 1944–1992." PNWD-2228 HEDR, Draft, Battelle Pacific Northwest Laboratory, Richland Washington.

Shipler, D.B. and Napier, B.A. (1992), "HEDR Modeling Approach." PNWD-1983 HEDR, Battelle Pacific Northwest Laboratory, Richland Washington.

Simpson, J.C. and Ramsdell, J.V. Jr. (1993), "Uncertainty and Sensitivity Analyses Plan." PNWD-2124 HEDR, Draft, Battelle Pacific Northwest Laboratory, Richland Washington.

Jeremy C. York
Battelle Pacific Northwest Laboratories
jc_york@pnl.gov

TOPICS IN SCIENTIFIC VISUALIZATION

Using Gray in Plots

by Dan Carr

In the last newsletter I discussed color. A point that I forgot to make was that colors can get hard to discriminate when the colored regions become very small. The newsletter editors reminded me of this fact by reducing the area of my choropleth residual plot by a factor of ten. Perhaps if I keep this article short the key plot in the article will be full sized. *(We too were surprised at and regret the quality of the color plots in the last issue, due to their size. As we gain experience with color plots, from YOUR submissions, we hope to improve the results. Eds.)*

This article continues on the topic of using gray in plots. Figure 1 shows a summary of the EPA's Toxic Release Chemical Inventory (TRI) for 1987 and provides the discussion example. Figure 1 is a plot that reexpresses most of a visually intimidating two-page table in Courteau 1990. In Figure 1, the basic patterns can be seen at a glance. It doesn't take long to find the two states with the greatest totals. A second glance indicates that the dominant releases for the two states go underground rather than being emitted to the air, stored on land, transferred to other sites, etc. Several aspects of the Figure 1 design can be called out: background grids, panel sizing to maintain comparability and preserve resolution, separate scaling of the margin total panel and its placement, sorting of states and panels by margin totals, staggered tic labels, and general labeling. Carr (1994) discusses these design aspects in a extended treatment of row-labeled plots. In this article attention focuses on design features related to gray and on plot interpretation.

Grids are back! Grids used to be common when people plotted on gridded graph paper. Historically Tukey (1977), Tufte (1983) and others have noted that heavy grids interfered with seeing data and hence interfered with geometric pattern perception. Soon grids disappeared entirely as some used tracing paper to improve hand drawings and many began generating plots using computers. This was taking the guidance too far. Recently Cleveland (1993a, 1993b) has demonstrated that background grids improve perceptual accuracy of extraction and help in making comparisons. The basic question now is how best to produce gentle background grids.

Grids are back! Grids used to be common when people plotted on graph paper.

The grid lines in Figure 1 are white lines on a light gray background (assuming decent color reproduction). The Washington Post commonly uses this approach. Cleveland's plots use gray grid lines on a white background. I have a slight preference for white lines on a gray background since the gray background reduces the contrast and gives the plot more of a value-added appearance. With light gray and white both methods provide unobtrusive reference grids.

In terms of production, laser printers that provide halftoning will produce gray regions and gray lines. Unfortunately low resolution half-tone gray appears sloppy and half-tone gray does not reproduce well given commonly accessible copying technology. What one really wants is gray ink. Some elegant gray ink examples can be found in Tufte (1983, 1990) and Grant (1993). While the current emphasis on black and white plotting will diminish over time as did the emphasis on black and white television, in the short term monochrome laser printing will have do for routine graphics. As for a modest number of copies, the best bet is to generate new originals from the electronic version.

Returning to the topic of grid lines, the horizontal grid lines in the plot help the eye match the state names with the bars in the different panels. The spacing after every fifth state creates smaller perceptual units (perceptual grouping is important, see Kosslyn 1994) that help in the matching process. With grid lines and grouping, the error rate for matching names to bars should be low. The vertical grid lines provide a basis for fairly accurate comparison of bar lengths. Note that in the right six panels the grid line spacing has the same units and the same physical size. Thus bar lengths in the panels are directly comparable and the grid lines make comparisons more accurate. The panels have differing widths to cover the differing ranges of the data. The margin total panel on the left has its own is separate scale. The design separates this panel from the other plots (slightly) to call attention to the fact. Also the design uses black bars in the panel to further emphasize that the scale and the vertical grid lines are different.

The symbols chosen for the display are bars. Bars, like most area filling symbols are visually pronounced. Patterns among the bars are easily seen despite separating panel lines. Unfortunately regularly spaced black bars on a white background can introduce the moire effects as described by Tufte (1983). A bit of this may be observed in the "Total" panel. The gray bars in the remaining panels diminishes the contrast with the background. This makes it easier to really look at the bars.

A weakness in the display scale is the inability to distinguish the small values from zero. Shifting the location of zero to the right of the panel edge would allow positive values to show as a thin line but this costs in terms of plot resolution and makes plot look busy. Producing a dot plot on a log scale would provide more resolution for the smaller values but is harder for many people to interpret. The bar plot as it stands focuses attention on the large values and they are the basic message.

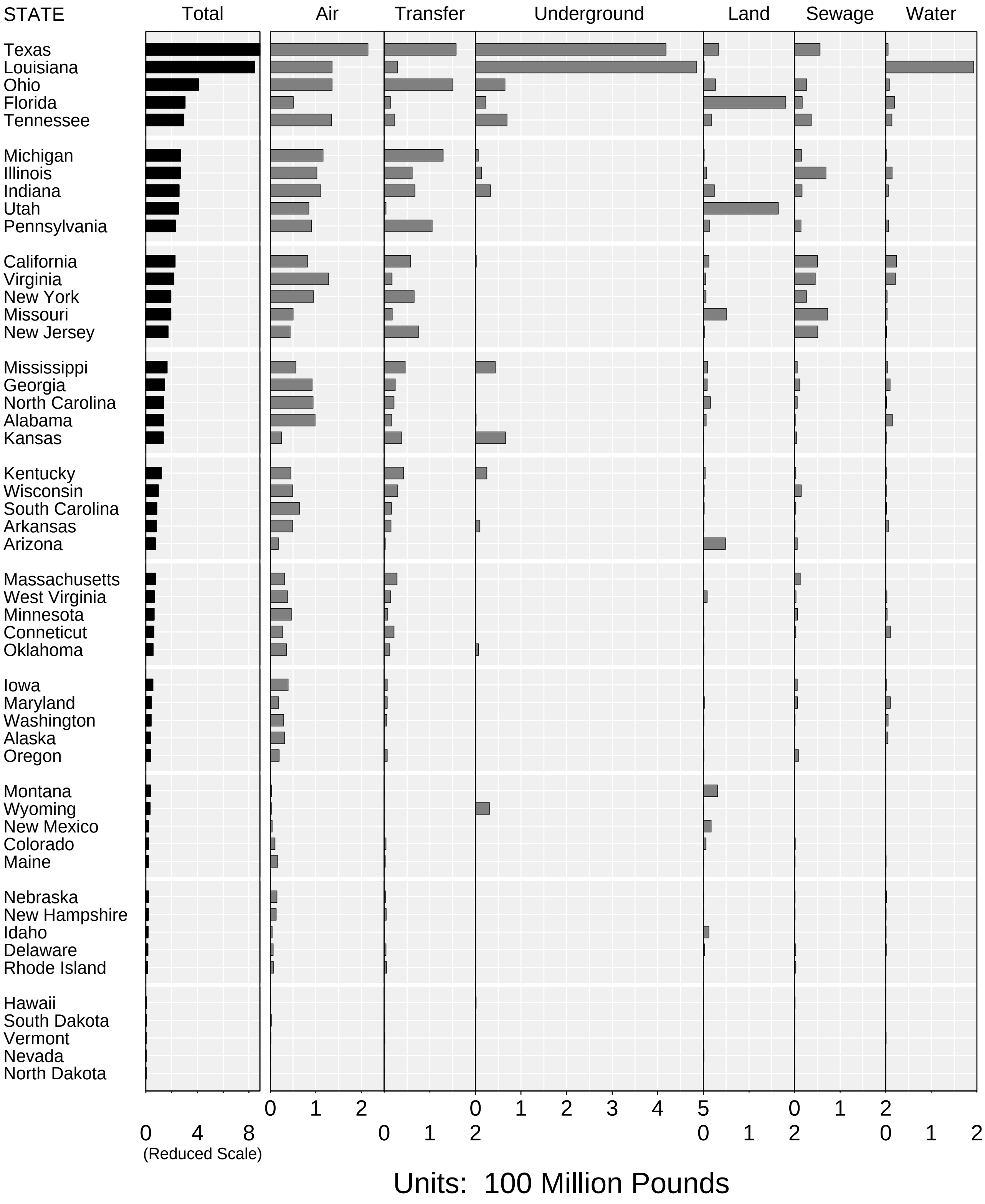
Truth in labeling for plots seems reasonable in an age of information consumption.

Unfortunately the interpretation of this graphic is limited by the data. The total of seven billion pounds of toxic materials being released or transferred may seem more than enough, but actually represents only a fraction of the unknown total. The regulations behind the Toxic Release Chemical Inventory only requires reports from not-small companies with certain SEC codes. In 1987, one-third of the companies who were obligated to report did not do so. The company reports are self-assessment estimates. While larger companies have environmental engineers to fill out the reports, in some cases the task falls to secretaries. Further, summarizing in units of pounds is convenient but hides the differing toxicities of the chemicals released. Figure 1 represents a summary of the information collected. The connection of the summary to the state of the nation is tenuous except as providing guesstimated lower bounds.

Metadata should be attached to plots like this to provide an appropriate context for interpretation. A significant problem is that while government documents often do an excellent job of providing the context in writing, the plots are often at risk of being extracted and used to influence public opinion and public policy. Truth in labeling for plots seems reasonable in an age of information consumption, and that is a good topic for another article.

TRI Releases And Transfers For 1987

Totals By State and Distribution Class
Grand Total = 7 Billion Pounds



As for producing similar plots, the Splus functions, script files and data for all the row-labeled plots described in Carr (1994) are available by anonymous ftp. The computer is `galaxy.gmu.edu` and the directory is `/pub/submissions/rowplot`.

Acknowledgments

Research related to this article was supported by EPA under cooperative agreement No. CR8280820-01-0. This article has not been subjected to the review of the EPA and thus does not necessarily reflect the views of the agency and no official endorsement should be inferred.

References

Carr, D. B. (1994), "Converting Tables to Row-Labeled Plots." Technical Report No. 101, Center for Computational Statistics, George Mason University, Fairfax, VA.

Cleveland, W. S. (1993a), "A Model for Studying Display Methods of Statistical Graphics," *Journal of Computational and Graphical Statistics*, 2 (4), 323-343.

Cleveland, W.S. (1993b), *Visualizing Data*, Summit NJ: Hobart Press.

Courteau, J. B. (Editor) (1988), *Toxics in the Community, 1988 National and Local Perspectives*, EPA 560/4-90-017, Washington, D.C: U.S. Government Printing Office.

Grant, J. P. (1993), *The State of the World's Children 1993*, New York, NY: Oxford University Press.

Kosslyn, S. M. (1994), *Elements of Graphic Design*, New York, NY: W. H. Freeman and Company.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Reading MA: Addison-Wesley.

Tufte, E. R. (1983), *The Visual Display of Quantitative Data*, Cheshire, CT: Graphics Press.

Tufte, E. R. (1990), *Envisioning Information*, Cheshire, CT: Graphics Press.

Daniel B. Carr
George Mason University
dcarr@galaxy.gmu.edu

UNIX COMPUTING

Regular Expressions

by Phil Spector

As you spend more time in a UNIX environment, one concept which you will see again and again is regular expressions. A variety of UNIX tools and commands use regular expressions, unfortunately not always consistently. Nevertheless, regular expressions are a very powerful tool for searching and processing text, and even in their simplest form can make many seemingly complex tasks very easy. In this article, I'll define just what a regular expression is and show some simple ways to use them.

Regular expressions are a very powerful tool for searching and processing text.

A regular expression is a way of describing a pattern of characters to a program so that it can either display or modify the occurrences of that pattern. The UNIX utility `grep` (which is an acronym for "global regular expression print") is the most basic UNIX tool which supports regular expressions; given a regular expression and one or more file names as arguments (in that order), `grep` will display to standard output each line of the file which contains the regular expression. (`grep` can also be used as a filter, by using it on the right hand side of a pipe and omitting the file name.) The simplest form of a regular expression is just a literal string; for example, if we wish to display each line of a file which contains the string `dog`, we can use the command "`grep dog filename`". A number of flags increase the utility of `grep`; they are displayed in the table below.

Flag	Function
`-i`	Ignore case
`-n`	Include line numbers
`-w`	Search for words
`-l`	List filenames only
`-v`	List non-matching lines

Table 1: Options for `grep`.

Of course, it often becomes necessary to search for something more elusive than a literal string. To describe more general patterns, special characters are used. For example, in a regular expression, the period (`.`) represents exactly one occurrence of any character other than a newline. Thus, the regular expression `d.g` will be matched not only by `dog`, but by `prodigy` and `bandage` since each of them contain a `d` and a `g` separated by exactly one letter. The command "`grep d.g` `filename`” would display all those lines which contain such a pattern. To narrow the range of acceptable characters in a regular expression, a character class can be used. This is simply a list of one or more characters surrounded by square brackets `[ ]`, and it is matched by only the characters within the brackets. If the first character in the bracket is a caret `^`, then the character class is matched by anything **except** the characters in the brackets. Ranges of consecutive characters (such as `a-z`, `A-Z` or `0-9`) can also be used in character classes, and have their obvious meanings. An asterisk (`*`) after a character or character class is matched by zero or more occurrences of that character class. To restrict matches based on the location of a pattern within a line, a caret (`^`) can be used to anchor a match to the beginning of a line, and a dollar sign (`$`) can be used to anchor a match to the end of a line.

In addition, escaped angle brackets (`\< \>`) can be used to specify word boundaries, as an alternative to the `-w` flag in `grep`. Thus the regular expression `dog` would be matched by `dog`, `endogenous`, `bulldog`, etc., while `\<dog\>` would be matched by only the word `dog`, and `\<dog` would be matched by `dog` and `dogwood`, but not `bulldog`.

> ***In fact, many of the features of regular expressions are used by the shell when it performs filename substitutions.***

It’s very important to remember that many of the special characters which are used to form regular expressions also have special meaning to the shell. In fact, many of the features of regular expressions are used by the shell when it performs filename substitutions. Thus, when you submit a command to the shell which contains a regular expression, you should surround it in single quotes to protect the special characters from the shell. For example, the command `grep \<dog\> filename` will not do what you would expect; you need to quote the regular expression: `grep '\<dog\>' filename`. When using `grep` you should also remember that, if it fails to find a match for a regular expression it will not print any diagnostic error. So if you’re using grep and see a message like “`No match`” or “ `... not found`”, it usually means that the shell is trying to interpret the special characters, and that you need to enclose them in single quotes to proceed.

Regular Expressions with Editors

You can use regular expressions to find patterns in files from inside editors like `vi` and `emacs`. In `vi`, use the escape key to get into command mode, and enclose your regular expression in either slashes (`/ /`) for a forward search or backslashes (`\ \`) for a backward search. In `emacs`, the commands “search-forward-regexp” and “search-backward-regexp” provide similar functionality. In addition, substitutions of text can also be performed based on the presence of regular expressions. For example, suppose we wish to remove leading zeroes from a file of numbers. A regular expression which will match a number with leading zeroes can be constructed as `\<00*`, that is a word boundary followed by one or more zeroes. Thus by substituting a null string for such a pattern, leading zeroes could be removed. This function is obtained using escape followed by an “s” command in `vi`, and with the command “replace-regexp” in `emacs`.

In the simple example above, we could perform the necessary change by focusing only on the zeroes in the number, which we wanted to eliminate anyway. A more challenging problem is one in which we need to describe a bit more of the surrounding territory of the pattern on which we want to focus. Suppose we have a file of containing both text and numbers in scientific notation (like `1.3e-5`), and we wish to change the small `e`’s of the scientific notation to capital `E`’s. It’s not too hard to construct a regular expression which will find the numbers in question: `[.0-9]*e[-+0-9][0-9]*`. (Notice that in the character class after the `e`, the dash (`-`) is the first character inside the brackets. This is necessary since it would otherwise be interpreted as part of a range of characters, like `a-z`.) However, in order to just change the `e` while preserving the rest of the number requires a new capability, namely tagging a sub-expression inside of a regular expression. In both `vi` and `emacs`, a tagged subexpression in a substitute command can be surrounded by escaped parentheses (`\( \)`) to tag it for future reference. Then, tagged patterns can be referred to as `\1`, `\2`, and so on, in the “to” portion of the substitute command. So, in `vi` for example, the necessary substitute command is `s/\([.0-9]*\)e\([-+0-9][0-9]*\)/\1E\2/g`. (The trailing `g` in the substitute command causes the substitution to take place for all occurrences within a line, not just the first.) A similar expression could be used with the “replace-regexp” command of `emacs`. Although expressions like this may appear complicated, they provide an efficient way of making complex contextual changes in text with a single command.

A Brief Warning

As I mentioned in the introduction to this article, regular expressions are not completely consistent from program to program. For example, the meaning of the asterisk (`*`) in the shell’s filename expansion is different from that used by `grep` and other programs which support regular expressions. In addition, other versions of `grep` (like `fgrep` and `egrep`) support additional features beyond those described in this article. Programs like `perl` (which will be the topic of future articles in this series) also have many additional extensions to regular expressions beyond what I've described. The features described in this article should be supported by most programs which use regular expressions, but, as always, the online manual pages should be consulted to resolve any discrepancies.

Phil Spector
Applications Manager
Department of Statistics
UC at Berkeley
`spector@stat.Berkeley.EDU`

NET SNOOPING

WWW continues to dominate the Net

by Mike Meyer

Over the last few issues of the newsletter my columns have been dominated by information about the World Wide Web. This merely reflects the enormous impact that the National Center for Supercomputing Applications' Mosaic user interface has had on the way many of us conduct our business. While NCSA continues to develop their version of Mosaic, they have also licensed their code base to several commercial software vendors who will provide enhanced and supported versions of WWW browsers. Many of the original authors of Mosaic have formed their own company to completely rewrite a new WWW browser. They have "an account at the local Domino's and a 'fridge full of cola", so expect to see something exciting from Mosaic Inc. in the near future.

With that introduction it is only natural that I continue to talk about *exploring* the world of WWW. Instead, I'll try to give you some idea of how easy (or difficult) it is to set up your own WWW server and *provide* interesting information to the world. For me, some of the most exciting developments have occurred on StatLib (`http://lib.stat.cmu.edu/`), so I'll talk about $2\frac{1}{2}$ new things. I hope my tight focus on StatLib and Carnegie Mellon will so outrage our readers that you will flood me with e-mail about other Statistics related WWW servers.

1994 Joint Statistical Meeting Abstracts

Have you ever wanted to look at the abstracts for a meeting before the meeting started? What about searching for authors or keywords to try to plan your time at the conference? For the 1994 Joint Statistical Meetings you can at least make some progress. The American Statistical Association office has provided StatLib with the text for 1421 abstracts of talks at the 1994 meetings in Toronto. This year authors were asked to submit their abstracts electronically (on disk, not e-mail) and were admonished not to include formulae in the abstract, but some still did. The abstracts are indexed by number and each abstract has a title, author, address, and key word field. Each abstract is also characterized by a "section" which is either the ASA meeting section (like general methodology, or statistical computing) or the name of a sponsoring society (like ENAR, IMS or the Statistical Society of Canada).

StatLib has made the abstracts available via WWW or e-mail. To see the abstracts either connect to the above URL and explore down the `joint94` path or send the e-mail message

```
send index from joint94
```

to statlib@lib.stat.cmu.edu. You can search the abstracts for specific authors or for specific words in the combined key word and title fields. You can view all of the abstracts in a specific section (like the 65 statistical computing abstracts or the 21 statistical graphics abstracts). The availability of the abstracts was announced just a week before this article was written, and in that time there have been over 2,500 WWW or e-mail transactions related to the abstracts.

> ***You can search the abstracts for specific authors or for specific words in the combined key word and title fields.***

So, how much work did it take to provide this information? The WWW server was already running (and setting that up takes at most a few hours). Getting searching to work took longer than I had hoped—a solid day of hacking at a PERL script and then another half a day of tweaking. The bulk of the work came in cleaning up the data. (That should sound very familiar to a statistician.) The data that the ASA office provided was very well organized but there were lots of small problems that did not surface until I began to test things. My programs were less resilient to very small errors than any one of us would be. Some records had the author and address fields reversed, or one field was duplicated. Some abstracts did not have sections (and my program expected them). Some had a "keywords" field instead of a "key words" field, and so on. I ended up writing a handful of PERL scripts and a number of emacs macros to fix problems in the data. There is a lot more I would like to do, in particular I would like to have the program schedule on line and linked to the abstracts. Fortunately for my other commitments, I do not (yet) have the raw data.

Carnegie Mellon Statistics Graduate Catalog

If you are not interested in abstracts of meetings, you might be interested in the graduate statistics program at, say, Carnegie Mellon. Luckily, you do not have to wait for a paper copy as the full text and pictures are available via the WWW. (Again, connect to StatLib and then follow the `cmu-stats` path.) We have the full text of our brochure, including all the wonderful photographs, but excluding some of the additional art work. The electronic version is certainly not as pretty as the printed version, but all the basic information is there, and the pictures of faculty members are quite impressive.

How much effort did this exercise require? I started with the text of the brochure as Microsoft Word documents, and high quality glossies of the photographs. I had someone read the Word files and save then in Rich Text Format (RTF). There is a public domain RTF to HTML (the WWW base language) converter that did an acceptable job of translating the text into a useful form. I had an undergraduate student scan in the photographs and deposit the images in a place where I could manipulate them. From there it took about a day of my time to hand edit the machine generated HTML and link in the photographs. The total investment was about 2 days of my time. This is small compared to the overall faculty effort required to produce the original paper brochure.

The Data Archive

Now for the $\frac{1}{2}$ of a new thing. Wouldn't it be great if someone would collect together lots of small data sets, with stories and key words. Wouldn't it be great if you could search the data sets for one (or more) that illustrated, say, outliers in regression. Wouldn't it be great if you could do this online, the night before an you have to hand out the next assignment in your statistics class. Well, you can't do that yet. But, stay tuned. Someone has created something like this and StatLib hopes to be able to release it sometime this year.

Mike Meyer
Carnegie Mellon University
`mikem@stat.cmu.edu`

CONFERENCE NOTICES

Interface '95

27th Symposium on the Interface:
Computing Science and Statistics
April 13-16, 1995,
Vista Hotel – Pittsburgh PA,

The Interface Conference is the premier annual conference on the interface of computing and statistics. It is sponsored by the non-profit Interface Foundation of North America, and will be hosted in 1995 by Carnegie Mellon University and the Pennsylvania State University with Michael Meyer and James Rosenberger as joint program chairs. The 1995 conference will be in Pittsburgh Pennsylvania, home of many excellent academic and industrial statistics and computer science research programs, and one of America's most livable cities.

Call for Suggestions

The Symposium is being organized around the theme of "Statistics and Manufacturing," with a sub-theme of green manufacturing, the environment, and quantitative environmental science. Also, sessions will include the broad topics of "Statistical Analysis of Graphical Data", including environmental, geographical, and imaging. Suggestions for sessions and speakers are still welcome.

Contact: Email: `interface95@stat.cmu.edu`
Phone: (412) 268-3108 Facsimile: (412) 268-7828
Mail: Interface 95, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA.

WHAT DO STATISTICIANS SPEND THEIR MONEY ON?
SPEND?! THEY DON'T - THEY JUST THROW IT UP IN THE AIR.

by Andrejs Dunkels

NEWS CLIPPINGS

Election Results

Results are in for the 1995 ASA elections of the Statistical Computing and Statistical Graphics officers.

1995 Officers

Statistical Computing Section

Chair Elect: Sallie Keller-McNulty
Program Chair-Elect: Robert J. Tibshirani
Secretary/Treasurer: Terry M. Therneau
Council of Sections Representatives:
MaryAnn H. Hill (1995-1997)
Michael M. Meyer (1995-1996)

Statistical Graphics Section

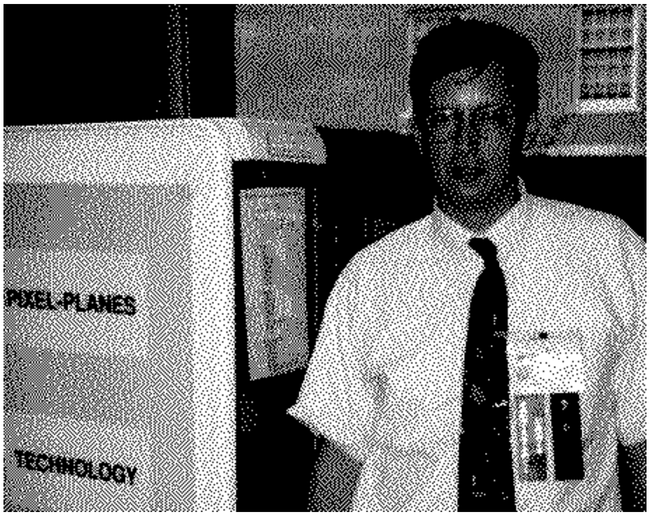
Chair Elect: William DuMouchel
Program Chair-Elect: Stephen G. Eick
Secretary/Treasurer: Robert L. Newcomb (1995-1996)
Publications Officer: Deborah J. Donnell (1995-1996)
Council of Sections Representative:
Colin R. Goodall (1995-1997)

Congratulations to them all!

Interface '94 Highlights

Tom Devlin (left) at the SAS Institute tour, talking with Russ Wolfinger, developer of Proc Mixed.

Program Chair John Sall and the organizing committee put together a great conference for Interface '94 in Research Triangle Park, NC. Approximately 350 people attended three days of courses, tutorials and talks. On Wednesday there was a choice of four short courses sponsored by the Statistical Computing Section of the ASA. In addition to the invited and contributed sessions on Thursday-Saturday, there were three tutorials and a demonstration dealing with the theme of the conference, "Computationally Intensive Statistical Methods". Phil Spector gave an introduction to Perl, Mary Ellen Bock introduced listeners to Wavelets, and Adrian Smith gave a tutorial on Markov chains. Tim Arnold gave a demo on networking innovations and the internet.



John Sall at the UNC visualization Lab.

A delicious continental breakfast was provided each morning, making it easier to attend the 8:15 sessions. In addition there were tours planned to the UNC Graphics and Image Lab and the SAS Institute. The Friday night banquet had some great entertainment. During dinner, there was music by the group "Bluegrass Retreat"—a wonderful treat—and the surprise was that John Sall's secretary Ruth Lee, who had worked so hard handling conference details was the electric bass player for the band. Wayne Lytle of the Cornell University Theory Center was the after dinner speaker. His topic was computer animation, and along with a fascinating discussion of animation he showed videos of animations he had done for a number of different applications. One of the most interesting was a video showing the progression of work on animation of musical instruments playing music he had composed. The end result was fantastic!

SECTION OFFICERS

Statistical Graphics Section - 1994

Roy E. Welsch, Chair
617-253-6601
Massachusetts Institute of Technology
`rwelsch@sloan.mit.edu`

David W. Scott, Chair-Elect
713-527-6037
Rice University
`scottdw@rice.edu`

Richard A. Becker, Past-Chair
908-582-5512
AT&T Bell Laboratories
`rab@research.att.com`

William DuMouchel, Program Chair
617-489-2631
`dumouche@jimmy.harvard.edu`

Sallie Keller-McNulty, Program Chair-Elect
913-532-6883
Kansas State University
`sallie@cecil.stat.ksu.edu`

Michael M. Meyer, Newsletter Editor (93-96)
412-268-3108
Carnegie Mellon University
`mikem@stat.cmu.edu`

Linda A. Clark, Secretary/Treasurer (93-94)
908-582-4807
AT&T Bell Laboratories
`lac@research.att.com`

Howard Wainer, Publications Officer (93-94)
609-734-5623
Educational Testing Service
Princeton, NJ 08541
`hwainer@rosedale.org`

Jane F. Gentleman, Rep.(94-96) to
Council of Sections
613-951-8213
Canadian Centre for Health Information
Ottawa, ON K1A OT6, CANADA
`GENTLEJF.NRCVM01.bitnet`

Sally C. Morton, Rep.(94-95) to Council of Sections
310-393-0411
The Rand Corporation
Santa Monica, CA 90407-2138
`Sally_Morton@rand.org`

Statistical Computing Section - 1994

Trevor J. Hastie, Chair
908-582-5647
AT&T Bell Labs
`trevor@research.att.com`

Mary Ellen Bock, Chair-Elect
317-494-6053
Purdue University
`mbock@stat.purdue.edu`

Sanford Weisberg, Past Chair
612-625-8777
University of Minnesota
`sandy@stat.umn.edu`

Sallie Keller-McNulty, Program Chair
913-532-6883
Kansas State University
`sallie@cecil.stat.ksu.edu`

John A. Rice, Program Chair-Elect
510-642-6930
University of California at Berkeley
`rice@stat.berkeley.edu`

James L. Rosenberger, Newsletter Editor (93-96)
814-865-1348
The Pennsylvania State University
`JLR@stat.psu.edu`

Deborah F. Swayne, Secretary-Treasurer
908-829-4263
Bell Communications Research
Morristown, N.J.
`dfs@bellcore.com`

Karen Kafadar, Publications Liaison Officer
301-496-8556
National Cancer Institute
`kk@helix.nih.gov`

Ronald Thisted, Rep.(94-96) to Council of Sections
312-702-8332/8333
The University of Chicago
`r-thisted@uchicago.edu`

Russell Lenth, Rep.(93-95) to Council of Sections
319-335-0814
University of Iowa
`rlenth@stat.uiowa.edu`

Daryl Pregibon, Rep.(92-94) to Council of Sections
908-582-3193
AT&T Bell Labs
`daryl@research.att.com`

INSIDE

Statistical
COMPUTING & GRAPHICS

The *Statistical Computing & Statistical Graphics Newsletter* is a publication of the Statistical Computing and Statistical Graphics Sections of the ASA. All communications regarding this publication should be addressed to:

James L. Rosenberger
Editor, Statistical Computing Section
Department of Statistics
The Pennsylvania State University
University Park, PA 16802-2111
(814) 865-1348
JLR@stat.psu.edu

Michael M. Meyer
Editor, Statistical Graphics Section
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-1380
(412) 268-3108
mikem@stat.cmu.edu

All communications regarding membership in the ASA and the Statistical Computing or Statistical Graphics Sections, including change of address, should be sent to:

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3402 USA
(703) 684-1221 • FAX (703) 684-2036
ASAINFO@ASA.MHS.COMPUSERVE.COM

PENNSTATE
1 8 5 5

Department of Statistics
University Park, PA 16802-2111

Published by the Penn State Department of Statistics
326 Classroom Building, University Park, PA 16802-2111

This publication is available in alternative media on request.

Penn State is an affirmative action, equal opportunity university.
U.Ed.SCI 95-2