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

April 95 Vol.6 No.1

Statistical COMPUTING & GRAPHICS

A WORD FROM OUR CHAIRS

Statistical Computing

Mary Ellen Bock is the 1995 Chair of the Statistical Computing Section. She follows the able leadership of Trevor Hastie, and welcomes your feedback and comments during the year.

The editors give us free rein to talk about whatever is on our minds in this first column. So here goes.

In the long run no other factor affects our work as strongly as changes in the power and the methods of computing. It would help enormously if we could gaze into a crystal ball to decide where things are going. My own crystal ball nearly cracked when I read a recent article in SCIENCE, November 11, 1994. I realized that in the future I might need a molecular biologist as well as a computer scientist to solve some of my computing problems.

In the article "Molecular Computation of Solutions to Combinatorial Problems", pp 1021-1023. Leonard M. Adelman showed how to encode a large combinatorial problem in DNA molecules. His trick for solving the problem was to use the "enormous parallelism of solution-phase chemistry." (David K. Gifford has a nice commentary article in the same issue of SCIENCE, entitled "On the Path to Computation with DNA", pp 993-994.)

The problem that Adelman considered could be described as an optimal shop scheduling problem. He solved it in a form known as the "directed Hamiltonian path" problem.

CONTINUED ON PAGE 4

A WORD FROM OUR CHAIRS

Statistical Graphics

David Scott is the 1995 Chair of the Statistical Graphics Section. This is his first note to the section, and the editors encourage feedback, directly to the chair, or with letters to the editors of the Newsletter.

It was during the Fall of 1992 that I worked so hard as Program Chair for the 1993 San Francisco Annual Meeting. Time has moved more slowly as I have awaited my turn as your Section Chair. Many thanks to Roy Welsch for his well-organized efforts last year. We will have the usual business meeting and mixer in Orlando this August, and I welcome input and queries on the agenda. Of course, with only one formal meeting per year, the business of a section often seems to move along at a snail's pace. But more on this below.

I would like to welcome the new officers: Program Chair Sally Morton, Secretary/Treasurer Robert Newcomb, and Council on Sections Representative Colin Goodall. Bill DuMouchel as Chair-Elect can take a breather after putting together an excellent program last summer. Stephen Eick will be collecting new ideas as Program Chair-Elect. If you are interested in serving the section, please let us know. In case you have not heard, Sallie Keller-McNulty stepped down as Program Chair after winning selection as Program Officer at the National Science Foundation. Congratulations, Sallie.

Let me outline a few thoughts about the Statistical Graphics Section and its future role. We are a large and prosperous Section.

CONTINUED ON PAGE 3

EDITORIAL

This April 95 issue (Vol. 6 No. 1) of the Newsletter should arrive before the Interface Meeting in Pittsburgh, PA, June 21-24, 1995. The second issue for 1995 should arrive before the annual Joint Statistical Meetings in Orlando this August, and the final issue is a year end issue. The deadline for the next issue is June 25, and for the final 1995 issue October 25. Thanks again to all who have contributed.

This issue introduces the new chairs of our sections, David Scott, Statistical Graphics Section and Mary Ellen Bock, Statistical Computing Section. The Chairs have each introduced themselves to the section through the "Word from our Chairs" column beginning on page one.

Speaking of color, our regular columnist Dan Carr was not pleased with the color rendition in the last Newsletter. We were not pleased either(!). In his letter to the editors, Dan goes on to suggest that authors be given a chance to review their work before publication. We don't currently routinely ship galleys to authors and we are unlikely to start doing this during our term as editors. Furthermore it is unlikely that we would have caught the problem this way as the error occured during the final production phase of the newsletter. Finally we wish to emphasize that this is only a newsletter. We often take some editorial leeway in just making the author's material fit onto the page, and we sometimes resize graphs to make them fit into our format. In an archival journal we would have to be more careful. All the same we deeply regret the error in our last issue and apologise to Dan. We both (Dan and the editors) have learned a lesson in the perils of color printing.

Dan's column this month discusses parallel coordinate plots for cumulative probabilty and quantile plots. His new *pq* and *qp* plots provide yet another tool for comparing distributions.

Mark Monmonier talks about disease atlases and how to use them to forewarn you about potential or real environmental problems. A related reference that Mark does not cite is "The Atlas of Disease Distributions", by Andrew D. Cliff and Petter Haggett, 1988, Blackwell.

Al Liebetrau discusses automatic monitoring of the massive amounts of (visual) data that are now routinely generated. The article makes particular reference to some emerging systems that help to identify complex features in data and even what data is interesting enough to archive. All such automatic systems need to be used with care—every statistician has a horror story where they needed "raw" data but only some pre-filtered version is available. However, there is no doubt that some automation is needed.

In his "Unix Computing" column, Phil Spector talks about three sets of publicly available software (network news readers, PERL, and TEX) that are commonly available on Unix and other systems. Perl is certainly a product of Unix, but some network newsreaders and TEXwere both first available on other operating systems. Some of our readers might remember the TOPS operating system and the SAIL programming language and struggling to produce documents with early versions of TEX. We still foolishly struggle with TEX to produce this newsletter.

In "Net Snooping" Mike Meyer talks about an experiment in virtual conferencing that took place in Heidelberg during March 1995.

Finally, look for the second announcement about Interface '95 on page 22 of this newsletter.

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(did you realize that more and more journals, including the *American Statistician* are now being set in TEX), but will accept plain ASCII.

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

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

LETTERS TO THE EDITORS

Color Figure Disfigured

Letter to the Editors

I was disappointed by the appearance of my row-labeled box plot in the last newsletter. What appeared did not correspond to the postscript file that I sent. The company that performed the color separations for you did a poor job of aligning of the color-separated images and reintegrating the layers so everything would show. This

resulted in numerous problems such as fractured box plots: horizontal lines with jogs in them.

I hope that the fractured image does not discourage readers from seriously considering the positive aspects of proposed design. I chose the example to show how well the proposed box plot design stands up in extreme resolution circumstances, not extreme printing circumstances. I encourage readers to look at my postscript examples and to generate there own plots. The examples and Splus software are available via anonymous ftp to `galaxy.gmu.edu` under `/pub/submissions/rowplot`. In contrast to the 84 box plots shown in the newsletter, the "heart" example has just 12 box plots and shows off the composite symbol in more favorable circumstances. To improve on the cancer rate example a first try might use color to distinguish the box plots for females and males rather than left and right halves of the distributions.

Please consider giving the authors the chance to review their work before it is published. It is particularly important to review graphics. Graphic results can be device specific. For some devices I have had to introduce a variable to control the outlier alignment in box plots. The positioning of symbols and text relative to their supposed plotting coordinates can vary from device to device. What I see on the screen in a postscript viewer is not necessarily what is printed. The discrepancy can be more than low screen resolution artifacts. I need to see the printed output before I can determine if device specific adjustments are necessary.

Half-toning depends not only on the printing device but also on the device's network definition. If I define my printer as a HP LaserJet 4 with postscript capability I get a different result than if I call it a generic postscript printer. The same is true for the color plots on the HP 1200C printer. (I prefer the generic postscript halftones.) The graphic file conversion and interpretation processes do not lead to identical results. I don't have an example where this has directly affected the newsletter but device-definition-induced variation could potentially cause problems.

As a former newsletter editor I understand that the editors have their hands full in putting together the newsletter. The extra effort to send out drafts and gather corrections may simply be too much, but it is worth considering.

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

Editors' Note: *We agree that the color plot contained unacceptable alignment problems, and regret our proofing did not uncover this before printing the Newsletter. Since the graphic was provided in Postscript format, and our camera ready copy of the graphic (printed on a Hewlett-Packard 1200C) was exactly like the copy provided by the author, we falsely assumed the printer's equipment would likewise produce equivalent output.*

FROM OUR CHAIRS (Cont.). . .

Statistical Graphics

CONTINUED FROM PAGE 1

Under the previous ASA constitution, the primary function of the Section was to organize paper sessions at the various Annual meetings—a nontrivial task that has been ably carried out year after year.

Under the new constitution, funds available to the Section from dues and short courses have grown substantially and continue to accrue at a steady pace. I am pleased to announce that the Section officers voted $1,500 towards equipment and services that would move the ASA office from current electronic mail to a true Internet node connection. (Our sister societies such as AMS and SIAM already offer such a level of service and accessibility.) Mike Meyer brought this idea to the Section last summer. More recently, Lorraine Denby reports that the Council of Sections felt Internet connectivity was important, and thus we (and other sections) have acted. While it is nice to be able to move more quickly than the ASA Board on such matters, our Section has the wherewithal to undertake more regular expenditures. At a roundtable luncheon last year, many suggestions were put forth: paying travel expenses of outside speakers; cash awards for best paper, best thesis, and perhaps best presentation; the American Statistics Poster Competition; and others, all appeal to me. We will be assembling a list of ideas to bring to you. Such added responsibility and authority is not easily exercised under the current Charter of the Statistical Graphics Section, in my opinion. We might be better served (in the future!) by a longer term, perhaps two years, for the Chair. The reasons seem clear enough, and I will be polling past Presidents and current officers about its wisdom.

This year's Annual Meeting in Orlando should be especially stimulating for graphics enthusiasts. Some of

the most imaginative and technologically advanced displays and presentations can be found in adjacent amusement parks. I hope you can stay a few extra days and enjoy not only the excellent ASA programs but spend time with family and friends in the area, too. Thank you for the opportunity to serve you.

David W. Scott
Chair, Statistical Graphics Section
Department of Statistics
Rice University
scottdw@rice.edu

FROM OUR CHAIRS (Cont.). . .

Statistical Computing

CONTINUED FROM PAGE 1

A graph with a fixed number of vertices is given along with a number of "directed" edges connecting some of the vertices. The problem is to find a path (i.e. a sequence of edges) that visits each vertex in the graph exactly once. Also the path has to start from a special vertex called the "in" vertex and end at another special vertex called the "out" vertex. When there are lots of edges and lots of vertices, it is hard to find such a path or even tell if one exists. The number of possible paths to check grows explosively in the number of edges and vertices.

Adelman's first idea was a method for labeling the edges and vertices of the graph. It was to encode each edge or vertex by a fixed-length randomly chosen DNA sequence. (This is described by a sequence of letters, each of whose components come from the letters A, C, G, T). For instance all Adelman's labels had length 20 and one of the vertices was labeled

GCTATTCGAGCTTAAAGCTA

while another was labeled

GGCTAGGTACCAGCATGCTT

(The labeling was slightly different for the "in" and "out" vertices.) How does Adelman label a directed edge that goes from the first vertex above to the second one? He labeled the edge by combining the last half of the label for the first vertex with the beginning half of the label for the second vertex, i.e.

CTTAAAGCTAGGCTAGGTAC

The next idea was a way to join compatible edges with appropriate vertices to form enormous numbers of random paths through the graph. It involved physically mixing together multiple copies of each real single-stranded DNA sequence that corresponded to an edge label and each DNA sequence that corresponded to a "complement" vertex label. (Initially this did not include the "in" and "out" vertices.) Getting millions of copies of these DNA sequences is easy and they are very tiny, measured in Angstrom size. The resulting ligation (concatenation) reaction due to the mixing forms double-stranded molecules representing random paths through the graph. This chemical process that formed the numerous random paths plays the role of a vast array of parallel computers.

Below is a representation of the double-stranded molecule representing the ligation of the complement sequences of the two vertices above with the sequence for the directed edge described above that connects them. (The complement sequence is formed from the original sequence by changing every C to a G, every G to a C, every T to an A and every A to a T.) Chemical bonds tying G to C and T to A are represented by the lines in the graph of the resulting double- standed DNA sequence.

```
CGATAAGCTCGAATTTCGATCCGATCCATGGTCGTACGAA
                |||||||||||||||||||
                CTTAAAGCTAGGCTAGGTAC
```

The methods Adelman used for sorting among the resulting random molecules (each representing a path) to actually find a Hamiltonian directed path make use of several other molecular biology tools such as the polymerase chain reaction (PCR), which there is no room to describe here.

There are plenty of open problems involving the method, particularly in deciding the amount of "label" DNA molecules to put in the mixture. This is to insure that with high probability a Hamiltonian path will be formed if it exists in the graph. Opportunities for improving algorithms for the method are wide open. Extensions to other problems would be exciting if the problems are of a type that would benefit from a massively parallel approach.

To learn more about computing with DNA molecules, read the article but I hope I have given you a hint of what might be in our future.

Before closing, I do want to note that I find the sections' newsletter to be a great vehicle for keeping up and learning about the many aspects of statistical computing and

graphics that may affect us. If you are interested in editorial service please let the editors or a member of the executive committee know. Suggestions about short courses and other activities are also welcome. We are very receptive to new ideas and better ways to serve the section.

Mary Ellen Bock
Chair, Statistical Computing Section
Department of Statistics
Purdue University
mbock@stat.purdue.edu

NEWS CLIPPINGS

Results of Student Paper Competition

The results are in! We have selected the four winners of the Computing Section's 1995 Student Paper Competition. In alphabetical order they are

- **Sudeshna Adak**, Stanford University: *Tree based Adaptive Estimation of Time-dependent Spectra for Nonstationary Processes*
- **John Gavin** and Christopher Jennison, University of Bath: *Subpixel Reconstruction in Image Analysis*
- **William Lu**, University of California, Berkeley: *The Expectation-Smoothing Approach for Indirect Curve Estimation*
- **Yingnian Wu** , Harvard University: *Random Shuffling — a New Approach to Match Making*

Before I tell you more about each of the winners, a bit more about the competition itself. This is our first such competition, and we had 15 entrants. Daryl Pregibon and I were the judges, and we had a really difficult task. There were many good submissions, but these four prevailed. If you plan to attend the ASA this year, come and see for yourself.

As part of their prize, each of the four winners will present their papers in a special session at the ASA. The more tangible part of their prize is that their entire conference trip will be covered by the section—airfare, hotel and registration! The approximate value is $1000 each.

We plan to hold such a competition every year, and the papers are due in early January. Registered students are eligible to submit papers. See the December 1994 issue of this newsletter for more details of the conditions.

Details on the Winners

Sudeshna Adak

Sudeshna is a third year Ph.D. student in the Department of Statistics at Stanford University. Her thesis is titled "Time and Frequency domain Analysis of Non-stationary Processes" and her advisor is Professor Iain Johnstone. Sudeshna's research interests include spectral analysis of non-stationary time series, time-frequency analysis, wavelets, cosine packets and best basis algorithms.

Her career plans include research, teaching and consulting.

Tree-based Adaptive Estimation of Time-dependent Spectra for Nonstationary Processes

Modeling of nonstationary time series has found wide applications in speech processing, biomedical signal processing, seismology and failure detection. In this paper, the problem of defining a time-dependent spectrum for a class of locally stationary processes is addressed. A tree-based segmentation method of estimating the time-dependent spectrum is proposed for this class of processes. Results of simulation studies demonstrate that the method has excellent ability to adapt to the rate at which the time-dependent spectra of locally stationary processes change over time.

John Gavin

John Gavin is a final year Ph.D. student in the statistics department at the University of Bath, U.K. His thesis is entitled "Subpixel Image Analysis" and his supervisor is Professor Christopher Jennison. John's current research interests include computational statistics, statistical visualization and non-parametric statistics.

In the future, he intends to work on designs for interactive and information-rich displays for large dynamic databases.

Subpixel Reconstruction in Image Analysis

In statistical image reconstruction, data are often recorded on a regular grid of squares, known as pixels and the reconstructed image is defined on the same pixel grid. This approximation to the true boundary can result in a loss of information which may be quite noticeable for small objects, only a few pixels in size. However, if some prior assumptions are made about the

true image, reconstruction to a greater accuracy than that of the recording sensor's pixel grid is possible. We adopt a Bayesian approach, incorporating prior information about the true image in a stochastic model which attaches higher probability to images with shorter total edge length. In reconstructions, pixels may be of a single colour or split between two colors.

William Lu

William Biao Lu is a final year Ph.D student in the Division of Biostatistics, Statistics Department at UC Berkeley. His thesis is titled "The Expectation-Smoothing Approach with Application to Ill-posed Statistical Inverse Problems" and his advisor is Professor Nicholas Jewell.

His career plan is to be a biostatistician and his research interests include missing and/or incomplete data problems, statistical computing, survival analysis and semi-parametric models.

The Expectation-Smoothing Approach for Indirect Curve Estimation

The EM algorithm is generalized for indirect functional estimation problems by substituting the M-step with a S(moothing) step. This simple method, which we call ES approach, can be justified from a penalized likelihood viewpoint. It makes usual smoothing techniques readily available to indirect estimation problems that are mostly ill-posed. The applications to backcalculation of infection curve from disease incidence data and estimation of mean marker trajectory from prevalent cohort are discussed.

Yingnian Wu

Yingnian is a third year Ph.D. student in the Harvard Statistics Department. His thesis advisor is Professor Donald Rubin. Some recent achievements include work on inconsistent inference, mixture modeling and statistical computing, and he is also interested in missing data, imputation and computer vision.

He would like to teach and do research in a university environment.

Random Shuffling — a New Approach to Match Making

This paper proposes a random shuffling algorithm, which serves as an engine in a matching-fitting algorithm, to solve match making problem and to draw inference involving unknown match in a Bayesian framework. The record linkage problem is studied to demonstrate our method, where the objective is to match two files which contain records for the same sample of units, and perform data analysis on the matched file, when there is no unique identifier, and the records are subject to different sources of errors and omission. The broken sample problem is also considered.

Trevor Hastie
Student Paper Selection Committee
Statistical Computing Section
Statistics Department
Stanford University
trevor@mallet.Stanford.EDU

GRAPHICS JSM PLANS

Graphics Program at Orlando

Invited Sessions

These were primarily organized by Sallie Keller-McNulty, though Sally Morton is responsible now as current program chair.

The sessions are:

- "Vision and Visualization", organized by Rick Becker, AT&T Bell Laboratories. The speaker is Stuart Anstis, Department of Psychology, University California at San Diego and his talk is entitled "To Understand Visualization, It Helps to Understand Vision." Some of Professor Anstis' recent research includes going from observations on optical illusions to devising a test for color blindness in infants.
- " 'X' years from the Grand Tour: Visualizing data in three or more dimensions," organized by Sally Morton, RAND. This session is historical in nature. A compilation of clips from the Statistical Graphics Video Library will be shown. The speakers will be coordinating their talks to cover the developments, high points, and low points in statistical graphics from its roots in the '70s with work on projection pursuit regression and the grand tour to today and beyond. The speakers are Werner Stuetzle (University of Washington, joint work with John McDonald) - "Visualization beyond points and lines" and Andreas

Buja (AT&T Bell Laboratories) and Dianne Cook (Iowa State University) - "Through the windshield in n-dimensions." The discussant will be David Scott of Rice University.

- "Visualizing data on statistical maps," organized by Linda Pickle, NCHS. The speakers are Douglas Herrmann (NCHS) - "Cognitive processes in statistical map reading;" Stephan Lewandowsky (University of Oklahoma) - "Perception of clusters in statistical maps;" Alan MacEachren (Penn State University) - "Representing uncertainty on statistical maps;" and Marc Sebrechts (Catholic University) - "Design issues for a computer-display dynamic map system." The purpose of this session is to bring together a panel of prominent non-statisticians to present cross-disciplinary research that can help statisticians design maps that best display the underlying data patterns. Professors Herrmann, Lewandowsky and Sebrechts are psychologists, and Professor MacEachren is a geographer who has published/edited several books on map design.

- "Frameworks for data display," organized by Deborah Swayne, Bellcore. The speakers are William S. Cleveland (AT&T Bell Laboratories, joint work with Rick Becker and Ming-Jen Shyu) - "Trellis display;" Ted Mihalisin (Temple University, joint work with John Timlin and Jim Mihalisin) - "Visual analysis of very large multivariate databases." The software systems presented in either talk use matrices of conditional plots to represent multivariate data, and the goal of the session is to allow users to compare and contrast the two. The discussant is Leland Wilkinson from Northwestern University.

The Statistical Graphics section is also sponsoring a joint session with the American Accounting Association entitled "Comprehending complex multidimensional financial data using computer graphics: From Chernoff faces to business instrument panels." Speakers are C. Torben Thomsen (California State University at Fresno) and Robert Jensen (Trinity University).

Sally Morton
Statistical Graphics Program Chair
Sally_Morton@rand.org

STATISTICAL COMPUTING NOTICES

Computing Progam at Orlando

The meeting will feature the following invited sessions on Statistical Computing, focusing on the role of statistical computing in areas of current scientific interest:

- Augustine Kong has organized a session on "Statistical Computing and Modern Biology." Charles Geyer will speak on "Annealing Markov Chain Monte Carlo with Applications to Ancestral Inference," Mark Irwin on "Efficient Imputation in Linkage Analysis," and Jun Liu on "Statistical Models for Multiple Sequence Alignments."
- Finbarr O'Sullivan has organized a session on Statistical Computing and Medical Imaging. Speakers will be Paul Sampson, "Morphometric Analysis Derived from 3D Ultrasound Cardiac Images," Yudianto Pawitan, "Estimation of Transmission Factors for PET imaging," and Smarajit Bose, "Image Segmentation with Spatial Contiguity Constraints."
- A session on Statistical Computing and Environmental Sciences has been organized by Peter Guttorp. Dean Billheimer will speak on "Markov Chain Monte Carlo for Species Composition," Patricia Styer on "Variable Selection and Model Building: A Case-Study Comparison of Nonparametric, Semiparametric, and Bayesian Methods," and Colin Goodall on "Models, Computation, and Graphics for Spatio-Temporal Data."
- As those of you who have dealt with very large data sets in statistical contexts without sufficient statistics know, special problems arise, the law of large numbers notwithstanding. Daryl Pregibon has organized a session aptly titled "David meets Goliath: Statistics and Massive Data Sets." Featured speakers are Jon Kettenring "Massive Data Sets: Applications, Barriers, and Opportunities," Colin Mallows "Some (Statistical) Principles for Massive Data Problems," and Dan Carr "Visualization in Massive Data Sets."
- There were a number of excellent papers considered for the Student Competition organized by Trevor Hastie. Those that will be delivered are by William Biao Lu, "The Expectation-Smoothing Approach for Indirect Curve Estimation," John Gavin, "Subpixel Reconstruction in Image Analysis," Yingnian Wu, "Random Shuffling – A New Approach to Match Making," and Sudeshna

Adak, "Tree-Based Adaptive Estimation of Time-Dependent Spectra for Nonstationary Processes."

- Balasubramanian Narasimhan has energetically organized a Special Contributed Session on Statistical Computing in Lisp. Speakers will be Robert Stine on "Wavelets in Lisp-Stat," Frederic Udina on "Interactive Kernel Density Estimation with Lisp-Stat," R. W. Oldford on "Graphical user interfaces for Statistical Applications in Quail," and Christopher Brunsdon on "Exploratory Analysis of Geographical Data." Sanford Weisburg will be a discussant.

John A. Rice
Statistical Computing Program Chair
rice@stat.berkeley.edu

CONTINUING EDUCATON

Continuing Education in the Sections

Interface '95 in June

The Sections are co-sponsoring 3 half-day short courses and 2 two-hour tutorials at Interface '95 in Pittsburgh. All five events will be held on Wednesday, June 21.

The half-day courses are:

- Richard A. Becker, William S. Cleveland and Ming-Jen Shyu, "Trellis Displays"
- John Elder and Paul Hess, "Tools for Discovering Patterns in Data"
- Joseph McKean, John Kapenga, and Thomas J. Vidmar, "Using tcl/tk in Building Interfaces to Statistical Software and Robust Visualization"

The two-hour tutorials are:

- Brian W. Junker, "LaTeX for Statisticians"
- Bob Kuszewski, "HTML and Creating World Wide Web Pages"

More information about the conference appears on page 22 in this newsletter.

Joint Statistical Meetings in August

The Statistical Computing Section will co-sponsor three short courses at the Joint Statistical Meetings in Orlando this August. The topics and presenters are:

- Terry Therneau, "Extending the Cox Model"
- Carl J. Huberty, "Applied Discriminant Analysis"
- Nicole Best and David Spiegelhalter, "Introduction to Complex Bayesian Modeling Using BUGS"

CE Scholarship Program

The Statistical Computing Section has expanded its scholarship program and will offer scholarships to Section co-sponsored continuing education activities at Interface '95 and at the Joint Statistical Meetings. The scholarships now cover the cost of the course and, in the case of students, conference registration fees. Scholarships will be awarded on the basis of need with preference given to students and beginning researchers. Applications should be sent to Thomas F. Devlin, Mathematics and Computer Science Department, Montclair State University, Upper Montclair, NJ 07043, or via email to devlin@mozart.montclair.edu. Applicants should identify the short course of interest, explain the professional importance of the course to them, and include a brief resume.

Last year, the Statistical Computing Section awarded CE scholarships for courses offered at Interface 94 to Shaohsin Chen from Kansas State University, Dave Cummins from North Carolina State University, and Jaekyun Lee from the University of Wisconsin.

Thomas F. Devlin
CE Chair for Statistical Computing Section
Karen Kafadar
CE Chair for Statistical Graphics Section

GEOGRAPHIC INFORMATION SYSTEMS

Pattern Templates as a Geography-Side Approach to Epidemiological Visualization

by Mark Monmonier

Faced with the need to put together a conference paper in hurry, I thought the time was ripe for reworking an idea advanced several years ago (Monmonier 1990a

and 1990b), when my Atlas Touring Project was in full swing, but never developed.

The conference was the International Symposium on Computer Mapping in Epidemiology and Environmental Health, held in Tampa, Florida, in mid-February by the World Computer Graphics Foundation. Themes included mortality atlases, spatial analysis, and applications of geographic information systems. To fit the program, I focused on how atlas touring, scripting, and geographic templates might contribute to the principal goal of disease atlases, namely, identifying potentially meaningful spatial patterns for more detailed epidemiological (and perhaps clinical) follow-up. In an epidemiological context, the primary contribution of atlas touring and geographic templates is vigilance–reducing the risk of not seeing a noteworthy relationship.

A Catalog of Templates

This idea is quite simple really. Compile a catalog of potentially meaningful patterns and attempt to match them up with patterns of disease incidence or mortality. Although EPA's Superfund list and the Toxics Release Inventory would be a start, a comprehensive effort would require a much larger catalog. Templates can be global, national, regional, or local in scope, ranging from worldwide levels of ultraviolet radiation at the earth's surface to the forecast field of groundwater contamination around a landfill, chemical plant, or abandoned gasoline station. Indeed, old directories of manufacturers and certain kinds of retailers would be ideal material for the catalog, as would data on soils, groundwater, surface water, geology, and wind patterns.

This sounds like a fishing expedition, I know, but why not? After all, a vigilant search for potential health hazards is a fundamental duty of environmental health departments. To ignore this responsibility, or pursue it with comparatively ineffective tools, seems both incompetent and unethical. A conscientious public health agency would not only make automated geographic screening a key component in its electronic data system but also develop and maintain a catalog of pattern templates. To me, at least, the catalog ought to be a fundamental part of our national mapping program.

In my paper, I called this approach "cartographic prescreening" because the GIS would attempt matches of all patterns in relevant portions of the catalog with a great many maps of each disease distribution. Rather than offer the viewer a single map, as disease atlases do, or expect the interactive viewer to poke around in faint hope of an illuminating discovery, cartographic prescreening cuts to the chase by presenting potentially interesting views the analyst ought not overlook. A supercomputer would, of course, be handy in dealing with the myriad permutations of measurements, categories, and time periods as well as in providing real-time interactive exploration of large data sets.

Geography-Side and Statistics-Side Strategies

In developing the rationale for prescreening, I drew a crude analogy between spatial analysis and development theory. As economists contrast demand-side and supply-side approaches to fiscal policy, epidemiological analysts might usefully ponder the implications of complementary geography-side and statistics-side approaches to cartographic visualization. By geography-side, I mean template-based prescreening, and by statistics-side, I mean the development of a theoretically determined single-map solution based on the exact Poisson or a similar inferential test.

Let's look at the prevailing statistics-side approach to visualizing geographic data on mortality or morbidity. Rooted in probability theory and underlying assumptions about sampling bias and frequency distributions, strategies based on statistical significance yield a single map of hot spots for each disease-race-sex specific death rate. Used widely in cancer atlases, for which publication cost, page size, and other constraints impose a single-map strategy, statistically-derived approaches balance intensity of mortality against the size of the population at risk. The resulting map typically points out two kinds of places: (1) areas with higher-than-average age-adjusted death rates deemed statistically significant because of large populations unlikely to have had a randomly bad decade, and (2) areas, however sparsely populated, that command attention because of much-higher-than-average death rates. To provide comparable maps for a variety of demographic- and site-specific cancer rates, atlas authors define the first type of place with a constant level of statistical significance (the 5- or 10-percent level, commonly) and identify a second group of hot spots as some intuitively meaningful small proportion of all areas (the top 10 percent, say). Overlap of these two groups is not only possible but informative, and a third hot-spot category accords special recognition to places meeting both criteria. Because very-high rates are more relevant to epidemiological analysis than very-low ones, two remaining categories complete the map: areas with rates significantly below the national average and all other areas, collectively labeled "not significant." Cancer atlases often leave the latter category blank or shade it in gray on a color map that presents the single low-rate group in blue and the three hot-spot

groups in various shades of red and orange.

A serious weakness of the traditional statistics-based approach lies not so much in the adoption of an exact Poisson or similar inferential test as in the use of a single, standardized level of significance that focuses the analyst's attention on one and only one map. In this respect, the weakness is a shortsightedness akin to the equally rigid but clearly more mindless default classifications that software developers inflict on unwitting users eager for a choropleth map. Equal-intervals, the most common default classification, yields a single choropleth map—usually one with five classes. As practiced in recent disease atlases, probability theory also yields a rigid five-category map. Although statisticians can argue convincingly that their maps are more meaningful, and thus inherently superior, such posturing ignores the fact that both approaches yield a single snapshot in contexts where multiple views might be far more revealing.

By contrast, cartographic prescreening avoids an arbitrarily rigid map by adopting a variable cut-point that moves along the measurement scale in small steps. At each position, the cut-point divides the range into two categories. For epidemiological data and similar situations where higher values command more attention than lower ones, the system then compares all places in the higher category with a series of region and trend "masks" representing geographic patterns a savvy investigator might look for. Comparison is based on a matching coefficient that relates hits to non-hits. A "hit" is thus a place with above-average mortality that also appears on the list of places comprising the pattern mask.

For each pattern mask the process yields a rating curve, which can be graphed with the mortality rate scaled along the horizontal axis and the matching coefficient scaled along the vertical axis. The system might simultaneously display rating curves for the more promising templates on a composite graph and invite the investigator to select several for direct display, say, by overlaying templates on a two-category mortality map. A high-interaction graph allowing the experimenter to identify individual masks by pointing to their rating curves would be especially helpful in selecting geographic patterns that in various ways stand out. A cartographic window linked to the rating-curve graph might provide greater understanding than its static counterpart in a printed disease atlas.

Cartographic prescreening can overcome yet another source of arbitrary rigidity of disease atlases—disease rates averaged for fixed time periods, usually one or two decades. Although temporal averaging is a reasonable strategy for generating reliable rates for counties with small populations, environmental factors need not conveniently kick in at the beginning of the decade. Vigilance is served by a system that allows moving averages as well as time periods of variable length.

Cholera Deaths around the Broad Street Pump

I closed my paper by referring to John Snow's famous 1854 map of cholera deaths around London's infamous Broad Street Pump. Snow's map, I suggested, is more a matter of intelligent pattern matching than serendipitous cluster recognition. Snow considered cholera a waterborne disease, after all, and had atlas touring been available in the mid-nineteenth century, his catalog of geographic templates most certainly would have included individual masks around every public water-supply pump in London.

This interpretation of the Broad Street Pump tale is important because geographic templates afford a complementary place- or pattern-driven strategy for dealing with clusters. Indeed, a catalog of a priori patterns is especially useful for recognizing thin linear clusters (in contrast to fat, geometrically compact ones) associated with rivers, fault lines, transport routes, international boundaries, and other potentially important lineations. Even so, cartographic prescreening can never preclude or preempt other strategies (such as statistics-side maps found in disease atlases) for identifying and analyzing clusters. Because we could never be certain that the catalog was complete and up-to-date, the risk of "not seeing" would be too great.

References

Monmonier, M. 1990a. Atlas Touring: Concepts and Development Strategies for a Geographic Visualization Support System. CASE Center Technical Report no. 9011, New York State Center for Advanced Technology in Computer Applications and Software Engineering.

Monmonier, M. 1990b. "Strategies for the Interactive Exploration of Geographic Correlation," Proceedings of the 4th International Symposium on Spatial Data Handling, Zurich, Switzerland, 512-521.

Mark Monmonier
Syracuse University
mon2ier@syr.edu

Filtering on Interesting Data Features

by Al Liebetrau

This column features statistical computing and statistical graphics activities in science and industry. I invite your comments and suggestions for future columns. Please send comments, inquiries, and suggestions to Albert M. Liebetrau, Analytic Sciences and Engineering Department, Battelle-Northwest, MS K5-12, P.O. Box 999, Richland, WA 99352, 509-375-2694, AM_Liebetrau@pnl.gov

First There Were Data . . .

Today, scientists collect data at enormous rates. The Earth Observing System, for example, will collect more than a terabyte of data per day over the fifteen year life of the project. The DOE-sponsored Atmospheric Radiation Measurement Program is designed to receive up to 200 data streams from as many as ten observing sites around the world. Observations are taken at intervals ranging from several seconds to hours, depending on the particular instrument in question. Computer simulation models to analyze grand-challenge computer problems produce immense amounts of data. Data collected through such efforts are often highly dimensional, they typically exhibit complex interrelationships among variables, and they usually span large spatial or temporal regions. With present tools, it is comparatively easy to generate data sets that are too massive to yield to traditional experimental approaches or too complex for traditional theoretical methods.

And There Was Defeat . . .

An important goal is to examine these large data sets for "interesting" features. These features often spark new lines of scientific discovery which, in turn, result in new knowledge and improved understanding. The sheer volume of data to be examined can defeat many of the conventional statistical methods for data exploration and analysis. Even the most efficient graphical procedures can be overwhelmed: Either the interesting structural features are lost in the volume of data, or the investigator is forced to plow through an overwhelming number of visual displays, most of which are "uninteresting." The problem is compounded when the dimensionality of the interesting structural features increases or when those features change with time.

Scientists are typically interested in examining data streams to detect changes in some underlying process. Some changes may be detected by monitoring the mean of one or more variables that reflect the state of the process. For example, the mean of a variable can be monitored to detect outliers in the data, to identify shifts in the level of a process or in the performance of an instrument, or to identify gradual changes in the output of the process. In other cases the investigator is interested in changes that are too subtle or too complex to be captured by merely detecting changes in a process mean. These more general classes of features may appear in the data, for example, as drift among the means of several variables, changes in variability of one or more variables, or changes in the joint correlation structure among several variables of interest. If the number of variables contributing to the change exceeds four or five, then the investigator may miss it with even the most efficient visualization and graphical methods available.

Then There Was Hope . . .

The advantages of some type of automated monitoring system are obvious in situations like those described above where the data are voluminous and the features of interest are comparatively rare. In those cases where the interesting events may be comparatively easy to identify, automated monitoring eliminates errors, saves time, and eliminates drudgery. Automated monitoring is not only convenient, but absolutely essential for detecting the more complex or subtle changes.

A variety of approaches and objectives are possible in creating automated systems to identify and extract interesting features from large and complex data sets. One objective could be to index the data, very much in the same way that a card catalog is an index to a library. The idea is to mark the data in some way when the feature of interest is observed. The investigator can then retrieve the relevant portion of the data for detailed examination at a later date without having to search the entire record again.

Statisticians at Pacific Northwest Laboratory (PNL) have successfully used dynamic linear modeling methods to "mark" interesting portions of multivariate atmospheric data in real time so they may be quickly recovered at a later time. In this approach, features of interest are identified with the states of a suitable state-space model. For each data point, the posterior probability that the model is in each of its possible states is calculated. Large values in the resulting sequences of posterior probabilities serve to identify (mark) those data points where the corresponding feature occurs.

A second objective of an automated data screening system could be to identify complex structural features that are not readily found using conventional methods, as discussed above.

A third objective of an automated data screening capability could be to decide which data to archive in the first place: Data in the "neighborhood" of an observed feature could be stored while data in uninteresting regions could be retained in summary form or discarded altogether.

Researchers at Oak Ridge National Laboratory (ORNL) are developing an automated data monitoring system to achieve these last two objectives. Their goal is to develop a system that will automatically select a "relatively small" subset of the data that contain the features of interest, thereby relieving the investigator of the need to look at the entire data set.

. . . And Soon There Will Be FEaTureS

The ORNL system, called FEaTureS (for Feature Extraction for Long Time Series), is being designed to identify features of interest and extract portions of data that contain these features. The main consideration of the ORNL team is to develop a system that can learn how to efficiently detect selected classes of complex features in the data. A practical requirement is that the methods must be fast enough to apply to very large data sets.

Overview

The ORNL FEaTureS system is based on the concept that changes in the underlying process induce detectable changes in the parameters of a suitable data model. The basic approach is to look for "change points" with the aid of data models similar to those used in time series analysis. The FEaTureS concept has three major components, a feature identification and extraction stage, and analysis and visualization phase, and a learning phase. At the identification and extraction stage, preselected features of interest are identified in the data. Data containing these features, plus that in a surrounding neighborhood (called the context), are extracted from the data and stored. At the visualization and analysis stage, the investigator interactively employs visualization tools to explore the attributes of features generated at the extraction stage. Features are cataloged in terms of their attributes. In the learning phase, the cataloged data are used to optimize feature selection relative to the ability of each feature extraction tool to correctly classify the features.

Feature Identification and Extraction

The software underlying the feature extraction stage contains a tool kit of methods for detecting various changes in parameter values in the data model. Points at which the selected tool identifies a change in parameter values are flagged. The data at these points (called features), plus that in a surrounding neighborhood (called the context), are extracted from the series and stored. Typical features include outliers, the beginning or end of a trend, a change in variability, or a change in correlation structure.

The feature identification tool kit will contain a variety of detection methods including the following:

- **likelihood methods for change point detection.** The basis of the likelihood change point detector the comparison of a forecast based on data before the segment in question and a hindcast based on data that occur after the segment in question.
- **dynamic linear modeling.** The dynamic linear modeling approach is similar in concept to that used by PNL.
- **chaotic time series analysis.** Chaotic time series analysis, which is appropriate for nonlinear systems, involves the reconstruction and analysis of a geometric object in the state space of the series.
- **rescaled range analysis.** Rescaled range analysis employs the Hurst coefficient as a measures of the bias or trend in a time series.
- **the use of artificial neural networks.** Artificial neural networks are first trained to represent normal system dynamics and then used to detect departures in system dynamics from the normal state.
- **stochastic approximation.** This approach is based on the assumption that data derived from solving a complex set of differential equations are noisy data derived from a simpler but unknown nonlinear system.

Observations in segments of the series between features are collapsed over time and binned. Data in a "featureless" section of the series can be described using conventional statistical tools. Through this process, the time series is compressed into alternating sequences of binned segments and features with context. Finally, features are catalogued according to a number of attributes such as length, maximum, minimum, mean or median, variance, difference in the medians of the incoming and outgoing context, and ratios of the variances of the incoming and outgoing context. Other attributes can be defined as appropriate.

Analysis and Visualization

After the extracted features have been identified and classified catalogued, they are examined visually and then analyzed. Features may be marked at this stage for use in the learning phase. Visualization tools include a Feature Dendogram Catalog (FDC), a Time Between Events Plot (TBEP), and a Feature Plot. The FDC Plot provides a graphical catalog of the features clustered according to the attributes used at the extraction stage. For each feature or cluster of features, the TBEP shows the time since the feature cluster was last observed. This plot is useful for detecting whether certain feature clusters are periodic. As its name implies, the Feature Plot shows the segment of the time series that contains the feature and its context.

Learning

At the end of the analysis and visualization stage, all identified features are classified into one of two groups: those that were flagged at the analysis and visualization stage and those that were not. At this stage, the FEaTureS code begins an optimization based on the ability of each feature identification tool to classify the features correctly. The optimization process is repeated for each of the available tools. This process may be viewed as a multivariate optimization in which the objective is to minimize the number of misclassified features. The optimization results will differ both quantitatively and qualitatively from one tool to another. From a summary of these results, the scientist may decide which "optimized" tool to use.

In Conclusion

With today's automated data collection systems, it is possible to assemble enormous data sets with comparative ease. The ability to acquire data is outstripping our capacity to analyze them. The result is that in many cases only a small fraction of the data is ever examined. Automated data monitoring systems such as those described here offer one way method for recovering the interesting features of large data sets before they are lost forever.

Acknowledgments

I thank my colleagues at Oak Ridge National Laboratory for allowing me to share this description of their work in progress.

Albert M. Liebetrau
Battelle Pacific Northwest Laboratories
AM_Liebetrau@pnl.gov

TOPICS IN SCIENTIFIC VISUALIZATION

Parallel Coordinate Variants Of CDF and Quantile Plots

by Daniel B. Carr and Anthony R. Olsen

Introduction

This article describes two new plots for representing cumulative probability (p) and quantile (q) pairs. Traditional quantile and CDF plots represent *pq* pairs using Cartesian coordinates. Given the same *pq* pairs, the Cartesian plots are basically equivalent. The CDF plot puts probabilities on the vertical axis and the quantile plot puts probabilities on the horizontal axis. The proposed parallel coordinate approach uses two vertical axes. We call the plots *pq* plots or *qp* plots depending on the left-to-right order of the two axes.

Parallel coordinate plots provide an alternate approach to representing number pairs. The basic idea is to construct parallel axes and to connect the coordinates of point pairs using straight lines. Two key papers, Inselberg (1985) and Wegman (1990), describe the geometry and interpretation of parallel coordinates plots. The current application is particularly simple. Since the cdf is a function, lines for distinct points pairs never cross between the axes. Lines can intersect on the probability axis. That is, for discrete distributions a probability may connect to an interval on the quantile axis. However, the expected practice is to show connecting lines just for selected jump points. Lines appearing to intersect on the quantile axis can only be low resolution artifacts. For table look-up purposes following straight lines is easy. The absence of crossing lines makes the task even easier.

This article calls attention to two *pq* plot variations, the *pq* density plot and the *pq* piecewise linear plot. Carr and Olsen (1995) describe additional variations and provide construction details. The *pq* density plot represents the surface created by interpolating densities along the *pq* lines between parallel axes. When shown as color images (see Figures 1a and 1b) or as rendered surfaces, such colorful plots draw student interest. The second variation, the *pq* piecewise linear plot (see Figures 2a and 2b), is more of a visual table. This distributional summary retains substantial detail and is suitable for use as a map legend.

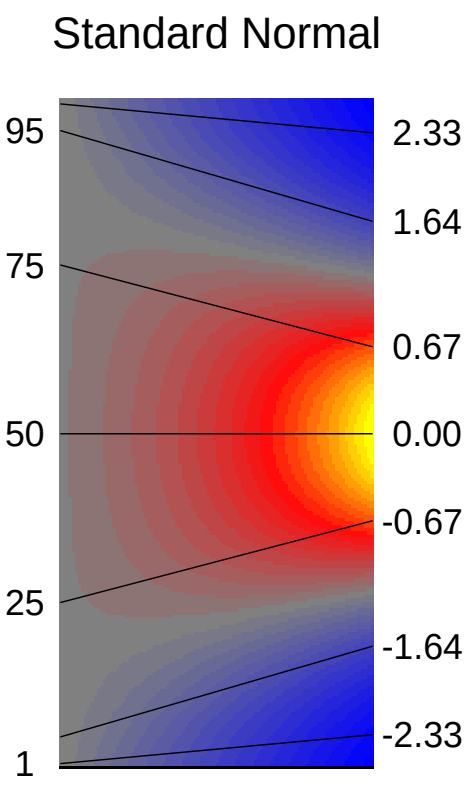

Figure 1a

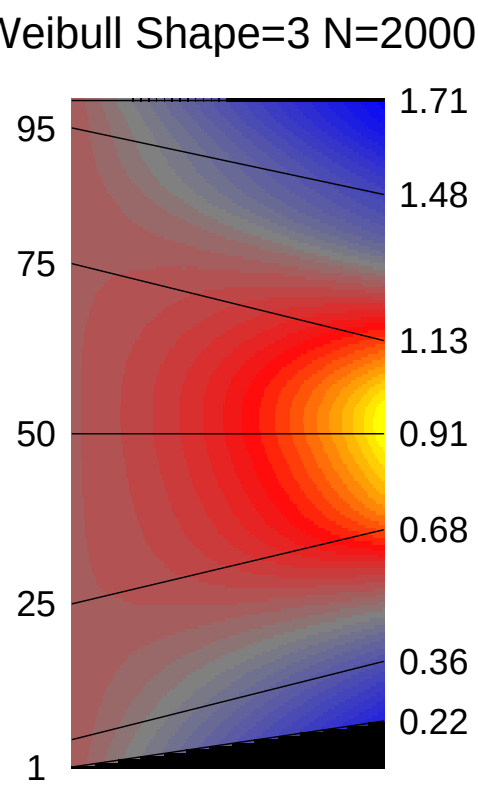

Figure 1b

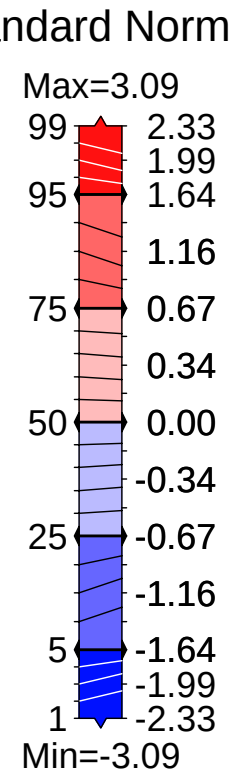

Figure 2a

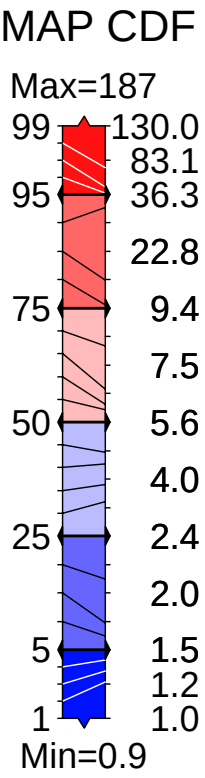

Figure 2b

CDF Plot and Legend Limitations

CDF plots similar to that in Figure 3 (the grid is typically omitted) provide distributional summaries and appear in reports by EPA and other government agencies. While commonly used, such plots prove awkward in regard to several tasks. The awkward tasks include looking up value pairs and, in the map legend context, showing color links to Choropleth map classes.

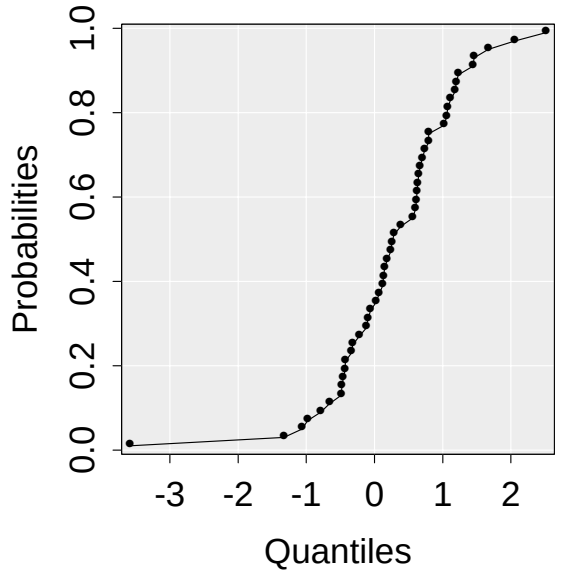

Figure 3: A CDF Plot

Consider the table look-up process in Cartesian coordinate plots. The visual path from quantile to curve to probability (or vice versa) involves a right angle turn. The visual path length differs markedly from small to large quantiles so the treatment of quantiles is not uniform. Standard horizontal and vertical grid lines based on pretty axis values do not typically intersect on the cdf curve and thus do not directly support the reading of specific *pq* pairs.

If the plot includes right-angle reference lines from the quantile scale to the cdf curve to the probability scale, interpolation may be still be required to "read" the values of one or both members of each reference pair. Adding reference lines and labeling their endpoints is a possibility but linear scales often leave little space for labeling especially along the horizontal axis. If skewness provides labeling space for one tail of the distribution, it robs space from the rest of the distribution. Common cdf plots show no reference pairs.

For typical cdf plots, readers must expend mental energy to obtain verbally expressed *pq* (or *qp*) pairs. In an application setting, such energy could be put to better use in memorizing a few values for later reference or in comparing values to other distributions. The Cartesian coordinate representation may seem to be a good storage device but is less than ideal for quick reading of value pairs. We conjecture that few people read more than one or two pairs from a typical CDF plot.

Data analysts often find Choropleth maps more informative if they include statistical summaries of the spatial phenomena. The *pq* pairs can represent a variety of summaries such as the population size in different classes, the map area in different classes, and the number of regions in different classes (see Carr 1993). Having chosen an appropriate summary, a statistician's first thought might be to add class colors to a cdf plot and use it as a legend. However, cdf plots are awkward for this task for two reasons. First, the cdf plot takes up a large area relative to the linear resolution of the two axes. Second, the addition of class colors to a cdf plot is a design challenge. Figure 4 uses gray levels in the plotting region to show class definitions. The disportionately large areas for large quantiles are not acceptable. A second choice is to add colored rectangles along the probability (or quantile) axis when the probabilities (or quantiles) define the classes. The smallest of these rectangles has to be of sufficient area so that the reader can easily perceive its color. Adding colored rectangles along an axis takes up more space as well as complicating the placement of tics and tic labels. The Cartesian coordinate approach is less than optimal for use as a legend.

Typical cartographic legends show the class colors in rectangles. Map makers label these rectangles with quantile (value) bounds or percent (probability) bounds but not typically both (for example see Dent 1992). Goldman (1991) shows both quantile and percent legends. By looking from legend to legend and focusing on corresponding class boundaries one can figure out a few *pq* pairs. Most of the distributional information is lost. Typical legends provide poor statistical summaries.

Estimation Issues

Before further describing the example *pq* plots, a few comments on the estimation of probabilities seem appropriate. Computing probabilities from samples is an essential task. For a simple random sample two estimation approaches are common, the empirical cdf approach and the distribution-of-order-statistics approach.

The empirical cdf for a sample of size n is

$$P(x) = i/n \quad x_{(i)} \le x < x_{(i+1)} \text{ for } i = 0, ..., n$$

where $x_{(i)}$ are order statistics, $x_{(0)}$=-∞ and $x_{(n+1)}$=∞. The definition is troublesome in the tails. That is, the estimated probability for future observations more ex-

treme than the sample extrema is zero. Such probability estimates for extreme values are biased low and hence counter-intuitive. While theory shows that the bias approaches zero as the sample size approaches infinity most people work with finite samples. Recognizing the possibility of more extreme values seems reasonable.

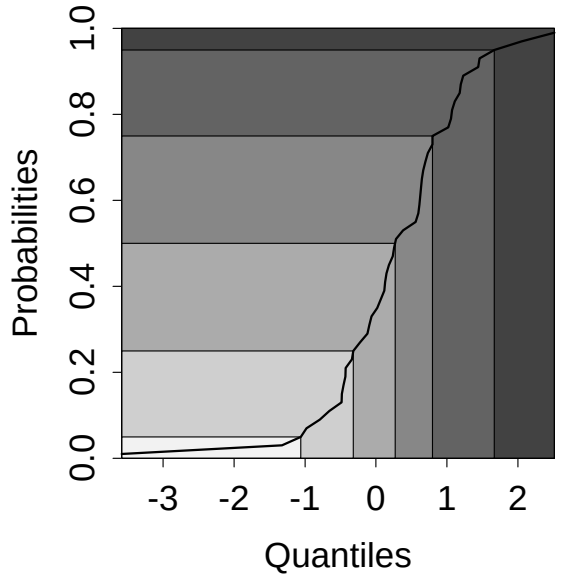

Figure 4: A Legend Attempt

Order statistics results (see Blom 1958, David 1972 and Hoaglin 1983) suggest a more plausible expression:

$$P(x) = (i - a)/(i + 1 - 2 * a) \quad \text{for } x = x_{(i)}$$

$$i = 1, ..., n \text{ and } 0 \leq a \leq .5$$

Chambers et al (1983) suggest obtaining probability estimates for quantiles between the order statistics by linear interpolation and estimates beyond the sample extrema by extending the probabilities at the sample extrema. The estimates for the current graphics use this approach with a=.5. The software referenced by this article will require revision when linear interpolation produces poor estimates.

Even with a distribution-of-order-statistics approach, probability estimates near extrema are uncomfortably close to pure guesses given the amount of scrutiny they may receive. The proposed graphic design allows users to finesse the issue by specifying quantiles or probability limits for the plot. The plot labeling can then list the sample extrema or user-imposed limits without attaching the corresponding probability estimates.

The PQ Density Plot

The *pq* density plot applies to distributions with density functions. The plot construction follows from a few simple observations. First, we can compute a density along each axis. Then we can interpolate the density along *pq* lines between the axes. The construction of density estimates for the quantile axis is a well-studied problem (see Scott 1992). We can chose from many methods. The probability integral transform states that the density is uniform on the probability axis. Since we assume a density function for the quantile axis, the cdf has no jump points. Consequently we can pick any point between the *p* and *q* axes and find the *pq* line that goes through the point. Doing this for a rectangular lattice of points between the axes and interpolating densities between line endpoints yields a density image.

Figure 1a is a *pq* density plot for a truncated standard normal distribution. The figure shows a few standard *pq* lines. The labeling for the *p* axis shows percents rather than probabilities. With minor exceptions the color assignment from blue to gray to red to yellow represents increasing densities with increasingly brighter colors (see Carr 1994). The color assignment fixes the number of color levels so that the "average density" on the *q* axis is gray. Correspondingly the whole *p* axis is gray. The blue regions on the *q* axis indicate below average density. The red and yellow regions indicate above average density. A pair of *qp* lines (right to left) starting in a blue region must converge (or compress the area between the lines) to achieve the uniform density value. A pair of *qp* lines starting in a red or yellow region must diverge (or expand the area between lines) to achieve the uniform density value. The plot can help student intuition. The color scale may not be the students' favorite and the opportunity to experiment with colors may lead to more than a passing glance at such plots.

The construction of Figure 1a uses theoretical density and quantile functions. In contrast Figure 1b uses function estimates based on a sample of 2000 points from a Weibull distribution. This distribution is not symmetric. Figure 1b illustrates a particular scaling choice for the axes. The choice forces the median line to be horizontal. The user specified quantile bound furthest from the median provides a second point and the two points determine the linear graphic scale. This scale leaves the *q* axis empty beyond the short tail and Figure 1b shows the empty region in black. The color assignment in Figure 1b should be adjusted so that the *p* axis is gray. (The process of generating both images directly on the same page introduced a color reassignment puzzle that

we have not yet solved.) The *pq* density plot applies to both theoretical and empirical distributions.

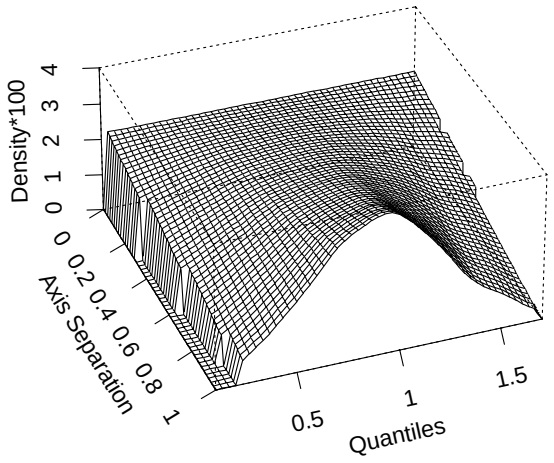

Figure 5: A Perspective View

The construction of the *pq* density plot brings together concepts of cumulative probabilities, quantiles, the probability integral transformation, order statistics, densities, interpolation, image construction and surface representation. Figure 5 shows a perspective view from the quantile side. This view provides more geometric intuition. The mesh would be better if it showed *pq* lines rather than a rectangular grid of lines. A fully rendered color surface with highlights and reference lines would look even better. The construction of different *pq* density views is an instructive exercise.

Inspection of the two plots reveals some omissions. The small plot size pretty well hides the dropped pixels along edges. The pixel problem can be fixed. The distressing omissions are the 5 and 99 percent labels. Due to the lack of plotting space, the software drops the labels. The problem is not just a coding artifact. Percents such as 1 and 5 are going to be close on any small plot with a linear percent scale. This labeling problem motivates the next variation, the *pq* piecewise linear plot.

The PQ Piecewise Linear Plot

The *pq* piecewise linear plot is a generalization of the legends shown in Carr, Olsen and White (1992). The previous legends had a linear *p* axis and represented key *pq* pairs using horizontal lines. Their approach implied a nonlinear *q* axis and readers could not estimate values for other *pq* pairs. The new version shown in Figures 2a and 2b is a set of vertically juxtaposed linear *pq* plots. The black triangles and thick horizontal black lines mark the division between the linear plots. The composite *p* axis is not linear. For example the regions above 95 percent and below 5 percent are larger than they would be on a linear scale. This facilitates labeling and showing more detail in the regions that are often of most interest.

When the plot is a legend for a Choropleth map the key *pq* pairs are the boundaries between the Choropleth classes. The regions between the axes then show the class colors.

Triangles at the top and bottom of the plots point to extrema. The user selects either quantile limits for the quantile axis or the probability limits for the probability axis and this determines the corresponding limits for the other axis. Either approach can exclude the sample extrema, increase resolution for values represented, and finesse uncertainties in calculating probabilities for extrema.

The plot design encourages reading from percents to percentiles (quantiles). The design provides lines for standard percents with 5 percent increments in the center and 1 percent increments near the tails. This generally corresponds to reader interest. For example, the 96^{th} percentile is usually of more interest than the 51^{th} percentile. For visibility the reference line colors need to contrast with the class colors. Figure 2a and 2b uses white lines for the extreme classes. This provides an additional cue about the special treatment of the extreme classes. Following the lines between the axes is easy.

To determine values in addition to the key *pq* pairs readers must interpolate quantiles using the probability-based reference lines. While interpolation is not trivial, the numerous quantile tics and tic labels provide bounds on the answer. The interpolation always occurs on a linear scale with bounding tics.

Close inspection correctly suggests that the number of quantile tics and regular spacing for the quantile labels drives the piecewise space allocation. Figures 2a and 2b suppress some of regularly spaced labels. The additional labels are useful to those interpolating values but begin to make the plot appear complicated. Cognitive studies may suggest an appropriate balance.

The Figure 2a represents a theoretical truncated normal distribution. Figure 2b represents *pq* values computed from probability sampling in an EMAP (EPA Environmental Monitoring and Assessment Program) study. Since the current emphasis is the graphical design, Figure 2b omits the context and measurement units. However, note that the reference lines show considerable angular variation. Large changes in the

quantiles sometimes correspond to small changes in probability. The largest quantile class shows considerable skewness. More resolution can be provided by selecting the 97th percentile as the bounding quantile. Of course this would hide the information about the higher percentiles.

Researchers who collect the data represented on maps are inclined to find most map legends impoverished. Those who study maps often want more distributional detail. The *pq* piecewise linear plot includes more detail without taking up much space. The plot provides a nice compromise between simple legends and extensive *pq* tables.

Closing Remarks

With an already burdensome variety of methodological alternatives, new methods proposed for use should be demonstrably superior to existing alternatives in some significant domain. Newness is not sufficient. (Our interpretation of Pregibon's Razor). In this article we suggest that the *pq* density plot and the piecewise linear plot have sufficient merit to warrant serious consideration.

The ability of plotting methods to generalize is also an important consideration. Can one compare two distributions? What would a parallel *qq* plot look like? How does one add confidence bounds? First thoughts might be that the comparison of Cartesian coordinate curves is so effective that there could not be a viable competitor, but is it so? Cleveland (1985) notes that humans do not judge distances between curves in the correct vertical direction but rather assess distances in a direction roughly normal to the curves. Common confidence lines for CDF plots can be very deceptive. Just maybe there is a better representation for some tasks but that is a topic for another article.

Readers can obtain Splus functions and example script files to conduct their own evaluations or adapt methods to their own applications. Use anonymous ftp to `galaxy.gmu.edu` and look in directory `/pub/submissions/pq`.

Acknowledgements

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

References

Blom, G. 1958. *Statistical Estimates and Transformed Beta-Variables*. John Wiley and Sons. New York.

Carr, D. B. 1993. Constructing Legends for Classed Choropleth Maps." Statistical Computing & Statistical Graphics Newsletter, Vol. 1. No 1. pp. 15-19.

Carr, D. B. 1994. Color Perception, the Importance of Gray and Residuals on a Choropleth Map. *Statistical Computing & Graphics*, Vol 5. No. 1, pp. 17-20.

Carr, D. B. and A. R. Olsen. 1995. "Representing Cumulative Distributions With Parallel Coordinate Plots." Technical Report No. 115. Center for Computational Statistics, George Mason University, Fairfax VA.

Carr, D. B., A. R. Olsen, and D. White. 1992. "Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographical Data." *Cartography and Geographic Information Systems*, Vol. 19, No. 4, pp. 228-236,271.

Chambers, J. M. W. S. Cleveland, B. Kleiner, P. A. Tukey. 1983. *Graphical Methods for Data Analysis*, Wadsworth and Brooks/Cole, Pacific Grove, California.

Cleveland, W. S. 1985. *The Elements of Graphing Data*. Wadsworth. Monterey, California.

David, H. A. 1970. *Order Statistics*. John Wiley and Sons. New York.

Dent, D. B. 1990. *Cartography, Thematic Map Design.* Wm. C. Brown Publishers. Dubuque, Iowa.

Goldman, B A. 1991. *The Truth About Where You Live*, Times Books, Random House Inc. New York.

Hoaglin, D. C 1983. "Letter Values: A Set of Selected Order Statistics" in*Understanding Robust and Exploratory Data Analysis*, Editors: Hoaglin, Mosteller and Tukey, John Wiley and Sons, Inc. New York. pp 33-57.

Inselberg, A. 1985. The Plane With Parallel Coordinates, " *The Visual Computer*, 1, pp. 69-91.

Scott, D. W. 1992. *Multivariate Density Estimation, Theory, Practice and Visualization.* John Wiley and Sons, Inc. New York.

Wegman, E. J. 1990. "Hyperdimensional Data Analysis Using Parallel Coordinates", *Journal of the American Statistical Association*, Vol. 85. No 411. pp. 664-675.

Daniel B. Carr
George Mason University
dcarr@voxel.galaxy.gmu.edu
and
Anthony R. Olsen
U.S. EPA
tolsen@heart.cor.epa.gov

UNIX jewels and perls

Introduction

In this article, I'm going to discuss a few topics which had their birth in the world of UNIX or are often found in UNIX environments, but are not part of the standard UNIX distribution. Even so, you're likely to find some or all of them on any given UNIX system. In addition, most of these programs are in the public domain, or protected by the GNU software license, so they can be obtained from archive sites on the Internet and used freely, as long as you don't charge a fee for their use.

Network News Readers

In the early 1980s (before the advent of such services as America Online and Compuserve), a small group of UNIX users in the Research Triangle of North Carolina experimented with the idea of passing messages among a network of computers, allowing users on any of the computers in the network to read the messages, and potentially contribute their own messages to ongoing discussions. From these small beginnings, the network news, or Usenet system of newsgroups has grown into an international network of machines and users with discussion groups for just about every imaginable (as well as some unimaginable) topics. Unlike bulletin boards, which are typically located on a single computer, there is no "home" for Usenet news; messages are transmitted from machine to machine all across the internet using the Network News Transmission Protocol (NNTP). To read the Usenet news, and to post articles of your own, you must have access to an internet machine which is running appropriate software; ask your system administrator if you have no success with the commands mentioned below.

To view the newsgroups, you'll probably want to use some kind of news reader software. One of the simplest (and oldest) news readers is `rn`; after choosing a newsgroup, each article is displayed and you either read it, or dismiss it and view the next. Alternatively, you can ask for a list of message headers, and pick the ones which interest you. Since `rn` uses a terminal interface, it can be somewhat awkward to pick and choose among the different articles. If you are running Xwindows, the `xrn` program will display a scrollable list of message headers, and will display the articles of your choice in a subwindow. Just about all the commands you could want to use are represented by buttons, so `xrn` is quite easy to use.

For a terminal interface, another alternative is `trn`, the threaded news reader. This program attempts to find articles discussing the same topic (by examining their subject lines), and to present them in a meaningful order. Since the volume of discussion in some news groups is very high, and since you probably won't be interested in every topic which arises, this program can be very helpful. There are other news reader programs available; ask a local guru about ones that might be on your system, if the ones I mentioned don't work out for you.

Even if you never read the network news (and be forewarned that it may be a very time-consuming activity), you may want to peruse the answers to the questions which arise again and again in some of the newsgroups. These questions are assembled into documents known as Frequently Asked Questions (FAQs), and they are posted regularly in each of the groups . If you have access to a World Wide Net browser like `mosaic` or `netscape`, you can view the FAQs for all the newsgroups by opening the URL `http://www.cis.ohio-state.edu/hypertext/faq/usenet`. Many of the FAQs are also available through anonymous ftp at `rtfm.mit.edu` in the directory `/pub/usenet`. It is strongly advised that you read the FAQ for a particular newsgroup before posting a message to the newsgroup; long-time devotees of the newsgroups get very upset when "newbies" ask the same questions over and over again.

Perl

Over the years, a variety of techniques have been developed to process files on UNIX systems, to reformat or extract data and produce reports. Until about 10 years ago, most of these techniques were carried out by writing shell scripts (that is, files containing instructions interpretable by a UNIX command shell) which in turn called such UNIX utilities as `awk`, `sed`, `grep`, `sort` and so on. While many programmers developed arcane skills in this enterprise, two major problems emerged. First, a variety of different shell languages and commands needed to be mastered before even fairly simple jobs could be attempted. This created a daunting barrier to most users, who were more interested in solving problems than learning yet another UNIX utility. But beyond that problem, there is a basic inefficiency in using shell scripts to coordinate the activities of different programs, because each time a new program was called, another shell had to be initiated to carry out the task, and then removed when the task was completed.

In response to these difficulties, Larry Wall of the Jet Propulsion Laboratory created `perl` (Practical Extraction and Report Language) in the early 1980s. The basic idea behind `perl` was to create a superset of the functionality of the most popular UNIX utilities, and to provide a coherent environment to carry the tasks out in an efficient manner. In addition, known problems and limitations of the existing tools were eliminated, and the end result is an extremely flexible tool, which can perform an unbelievably wide range of tasks. I routinely tell users at my site to forget about shell scripts and the other UNIX utilities I've mentioned above, and to focus on learning `perl`.

Although `perl` originated on UNIX systems, it has now been ported to a variety of environments. If you'd like to learn more about `perl`, a great place to start is to read the Frequently Asked Questions document (FAQ) from the Usenet newsgroup `comp.lang.perl`, archived on `rtfm.mit.edu` in the file `/pub/usenet/comp.lang.perl`.

TEX and LATEX

Producing publication quality documents, especially with mathematical equations is a difficult task, made more difficult by the fact that often times such documents are produced by word processing programs which either no longer exist, or whose formats are trade secrets, making it very difficult to electronically share formatted documents with colleagues. Even if one particular word processor were adopted as a standard, most such programs internally use unprintable characters in their documents, and this would create difficulty when attempting to transmit the documents through electronic mail, or to make any sense of them without access to the program which created them. One solution to these problems is a text mark-up language know as TEX, and an extension of the language known as LATEX.

TEX was developed in the 1980s by Donald Knuth of Stanford University as a system "intended for the creation of beautiful books". It is a markup language, as opposed to the more prevalent WYSIWYG (what you see is what you get) word processors which run on personal computers, which means that the object you operate on is an ordinary text file containing special directives (in the case of TEX the directives begin with the backslash character "\") which instruct it to use different fonts or type sizes, to construct tables and mathematical displays, and to place figures and tables in particular locations. The downside of this approach is that you need to learn the language, as well as an editor to create the input files, and that your file must be processed each time you wish to see the changes and additions you make. The upside is that the document you produce is human readable, easily distributable, and, since it contains all the information needed to produce your document, you can exert a great degree of control over the appearance of the final product. TEX and LATEX are quickly becoming standards for electronic distribution of manuscripts and many journals now accept or demand TEX input for their submissions.

LATEX is actually a collection of TEX macros which simplify the production of a number of standard document types. Many, but not all TEX commands are acceptable to LATEX, but most users find that it's easier to learn either TEX or LATEX than to try to become an expert at both. Macros for formatting articles may be available from some journal or book publishers, so if you have a particular target in mind, you might want to find out whether they prefer TEX or LATEX before starting to produce documents.

The FAQ for the `comp.text.tex` newsgroup can be found on `rtfm.mit.edu` in the file `pub/usenet/news.answers/tex-faq`.

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

NET SNOOPING

An experiment in Virtual Conferencing

by Mike Meyer

From March 23 to March 26 I attended a "Workshop on Design and Implementation of Data Analysis Systems" hosted by Günther Sawitzki at the University of Heidelberg. I was fortunate enough to be able to attend in person, but given the workshop nature of the meeting, the attendance was limited to the speakers and a few others. Some of the speakers at the conference included: R. A. Becker (Bell Labs), A. Buja (Bell Labs), W. S. Cleveland (Bell Labs),W. F. Eddy (Carnegie Mellon), P. Groeneboom (TU Delft), D. Keim (München), J. Marais (ETH Zürich), M. Nagel (Bad Elster), L. Tierney (Minneapolis), A. Unwin (Augsburg), F. Wietek (Oldenburg), and A. Wilks (Bell Labs).

The half-baked experiment

As the conference date was approaching Günther and I came up with the half-baked idea of running a virtual conference, so that people who were not actually at the site could at least participate in some way. The idea sounded attractive, but we really hadn't done our homework properly, so the half-baked idea turned into a half-baked implementation. What follows is a short description of what we did, what did and did not work, and what we might do differently the next time.

We had two different audiences and aims. First we wanted to provide a WWW exposure for the conference. The latest version of that material is still available at `http://statlab.uni-heidelberg.de/workshop.html`. We intended (and to some extent succeeded) to keep that material up to date as the workshop proceeded. Our second goal was to "broadcast" the conference over the internet using the CU-SeeMe software.

The facilities

The workshop was held in a small conference center/hotel that is near the old university campus (and in sight of the castle), but far removed from the main campus of the university. Given the network junkies listed above we clearly needed not only computers at the site, but also some good networking. Günther managed (I don't know how) to get hardware vendors to lend us a fleet of workstations for a few days. We had two Silicon Graphics machines (where Bill Eddy quickly discovered the flight simulator program) an HP and a Sun workstation and a few Macintoshes. That sounds like a lot of machines, but at peak times we could have used many more. The machines were connected on a local ethernet and thence, via a 64k ISDN line, connected to the university campus. Again Günther had done his homework by making sure that the remote end of the ISDN connection was only two routers away from the main link between Germany and the US.

The Network

Naturally we worried about only having a 64k connection, but that turned out to be only one of our problems—the effective throughput of the link from the US to Europe proved to be much more of a limitation. The 64k link *was* a problem if we wanted to be broadcasting the video/audio and doing any other remote file operations at the same time. A naive person might expect that at the time we were broadcasting, that is during the sessions, everyone on site would be attending the session and the network would be otherwise quiet. Well, the naive person probably hasn't given a workshop demonstration that required software on a network. We found that at any given time some attendee wanted to fetch some software or data from somewhere remote. That was actually a good sign because people were talking to each other and modifying what they were saying and doing in response to the workshop. But it did mean that our 64k line had to do double duty as both a video feed and for FTP connections. Both the video and the file transfers suffered, so we quickly asked participants not to use the network during sessions. The lesson for us was that we really needed two network channels—one solely for the video portion, and the other for general use.

On the first day of the workshop the network at the University of Heidelberg suffered some major trauma. A router failed, there was a fire in a wiring closet, and general instability on the university network. We could have been in deep trouble, but Günther's foresight of keeping us basically off the university network and near to the German backbone paid off, and our connections to the external world stayed up. Unfortunately, our WWW server was a machine at the university. Thus we, and everyone else, had difficulties accessing the server.

Lessons: Set up your network early, minimize the points of failure, practice first, and buy as much bandwidth as you can afford.

The Video Link

One of the prime reasons for the video link was to allow Paul Velleman of Cornell University to participate in the workshop. Our first experiment was a presentation by Piet Groeneboom at 8pm (German time) of the first night of the conference. I didn't actually hear much of the talk as I spent most of the time being a technician (and still suffering from Wednesday nights red-eye flight). However the WWW page does have a short description of the talk and links to related material.

By 8pm the network had been up for, oh at least an hour or so, so we set up the video camera and started broadcasting. We also opened a CU-SeeMe connection to the return broadcast from Paul Velleman's office. 8pm in Germany was 2pm in the US, so we had hit prime network time in the US but had mostly avoided it in Europe. Overall throughput was not great. Our best indication was that the video images we were seeing of Paul were quite jerky. In hindsight half of the problem was obvious. We were both broadcasting and receiving so we were placing double the demands on our network connection. We eventually (the next morning) realized this and limited ourselves to broadcasts only. However receiving as well had a clear advantage. After a while

Paul held up a sign that read "No Sound." He could obviously see us, but couldn't hear us. We immediately suspected all sorts of software problems, and never did resolve the problem that evening. The next morning we turned on the microphone.

Overnight we also searched for a European site to mirror or reflect our broadcast. The idea was that we would broadcast to the reflector and everyone else would tune into the reflector rather than us. Surprisingly we were able to find two volunteer reflector site in Scandinavia, and eventually used a machine at Lund University for the rest of the conference.

The broadcast had its ups and downs. We discovered the obvious, that clipped video (reduced number of frames per second) is acceptable, but clipped audio is not. We have no firm idea of the number of people who tuned into the workshop, but we know for sure that there were periods with at least a handful of viewers. It was a useful experiment.

Lessons: Practice, practice, practice, hire someone to operate the camera, and buy as much network bandwidth as you can afford.

The WWW Link

Our intention was that during the workshop we would regularly post information about current sessions to the WWW area. Sometimes we would have full electronic versions of the presentations, in other cases we would just have some minimal notes. We would also invite readers to submit comments about the various sessions. These comments would be summarized in the online forum. We planned to update the material on the server twice a day—once near the European morning and then again about 6 hours later near the United States east coast morning.

So the aim was twofold. For us to provide information about what was happening at the conference, and for others to conduct an online discussion with summaries presented at the workshop. That was a fine plan, except I was late telling people about it. I did post to a number of mailing lists and Usenet groups, but only just before the conference started. My plan had been to set up a listserv mailing list on my Carnegie Mellon machine once I get to Germany. Again, a fine plan until I discovered that telnet across the Atlantic (at least from where I was) was very difficult—so painful that I soon gave up any notion of getting real work done on my US computer. I did manage to set up a place for people to mail comments, but not a real discussion group. Hence input from the Internet was minimal.

That left us with the other half of the experiment, providing more or less real time input from the conference. That proceeded reasonably well. We updated the pages a few times a day, and included some notes we took and many links to existing material. Over the few days of the conference a few hundred people visited the WWW pages and looked at various aspects of the program and talks. I did get several complaints from non-europeans who were having difficult accessing the server. I would have to rate the output side of this part of the experiment a moderate success, and the input side a failure.

Lessons: Have a duplicate server in the US, don't count on long distance telnet speeds, announce it early, be prepared.

Conclusion

Don't try to do this with limited bandwidth. Practice first, and practice often. Let the world know what you are doing ahead of time. It takes more than one or two people to do an experiment like this. Don't fly to Europe one night (and not sleep), work all day (and not sleep), work much of the night (and not sleep) and still expect to get anything coherent done.

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

CONFERENCE NOTICES

Interface '95

**27th Symposium on the Interface:
Computing Science and Statistics**

Statistics and Manufacturing with Subthemes in Environmental Statistics, Graphics, and Imaging

June 21-24, 1995
Pittsburgh, PA
Pittsburgh Vista Hotel
David L. Lawrence Convention Center

The Interface Foundation of North America

Hosted by:
Carnegie Mellon University
Mike Meyer, Program Co-Chair
and
Pennsylvania State University
James Rosenberger, Program Co-Chair

General Information

An Invitation

The Interface Foundation of North America cordially invites you to participate in the 27th Interface Symposium, the premier annual conference on the interface of computing and statistics. The Foundation is a non-profit educational corporation founded in 1987 to sponsor the symposium and publish the proceedings. IFNA also co-publishes the Journal of Computational and Graphical Statistics. The theme this year is "Statistics and Manufacturing with Subthemes in Environmental Statistics, Graphics, and Imaging."

Contact:

email: interface95@stat.cmu.edu
phone: (412) 268-7834 fax: (412) 268-7828
mail: Interface '95
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-3890

World Wide Web Info

Updated conference information, directions, etc. is always accessible through our World Wide Web page (`http://www.stat.cmu.edu/interface95/`). For more information about the City of Pittsburgh, look at the Pittsburgh Home Page linked to the Interface95 page.

Schedule

The conference registration begins on Wednesday afternoon, June 21st. The first official conference event is a Wednesday evening mixer. Technical sessions will be held on Thursday and Friday from 8:15 am–5:15 PM and from 8:15 to noon on Saturday. Breaks are scheduled between the sessions and also for lunches. Several short courses will be offered during the day on Wednesday, June 21st. A description of the courses, costs and meeting times is included below.

Program. The program co-chairmen for the conference are Mike Meyer (Carnegie Mellon University) and Jim Rosenberger (Penn State University). The Program committee consists of: David M. Allen (University of Kentucky), Dan Carr (George Mason University), Phillippe Castagliola (IASC), Lawrence Cox (EPA), Noel Cressie (Iowa State University), Tom Devlin (Montclair State University), William F. Eddy (Carnegie Mellon University), Nick Fisher (CSIRO), Alan Genz (Washington State University), Arnold Goodman (UCLA), Jeffrey D. Helterbrand (Eli Lilly), Tim Hesterberg (Franklin & Marshall College), Fred Hulting (Alcoa), Karen Kafadar (University of Colorado), William Kennedy (Iowa State University), John Kettenring (Bellcore), John P. Lehoczky (Carnegie Mellon University), Kanti Mardia (Leeds University), G.P. Patil (Penn State University), and Brian Ripley (Oxford University).

Baseball Game. If you are a baseball fan, the Pittsburgh Pirates will be hosting the San Francisco Giants on Wednesday, June 21st at 3:00 PM. If you'd like to attend the game, please indicate that on your registration card. We need to get the tickets reserved early, to ensure availability. The cost should be $13 per ticket.

Thursday Night Banquet. This year's banquet will take place on board the Gateway Party Liner. The three hour cruise will include sightseeing along Pittsburgh's three rivers, musical entertainment, and the Captain's Buffet Dinner. The cost is included in your registration fee. Please indicate on the registration form if you plan to bring a guest to the banquet. You will be billed for the cost of your guest(s).

Conference Accommodations. The Pittsburgh Vista Hotel, conveniently located in downtown Pittsburgh, has rooms reserved for conference participants (see page 6 for more info). The conference meetings will be held in the David L. Lawrence Convention Center, which is linked to the hotel by an enclosed walkway.

Meals. The Wednesday night reception and the Thursday night banquet are included with your registration fee. All other meals will be on your own. There are five different restaurants/bars located in the hotel, along with many others that are located within walking distance of the hotel.

The City of Pittsburgh. Pittsburgh has changed from the industrial "smoky city" of the 1930s to a major center of business and high technology. In 1985, Pittsburgh was named America's most livable city by the Rand McNally Places Rated Almanac, and it continues to rank high in their ratings. One of the reasons Pittsburgh is considered so livable is its abundance of educational, recreational and cultural activities that are readily accessible. Both Carnegie Mellon University and the University of Pittsburgh, along with other colleges and universities, are located within the city limits. While you are in Pittsburgh, you might want to visit some of our local cultural attractions that are a short distance from the conference site: Andy Warhol Museum, Carnegie Museum of Art, Carnegie Museum of Natural History, University of Pittsburgh's 23 Nationality Rooms, Phipps Conservatory, the

Frick Art and Historical Center, the National Aviary, and the Carnegie Science Center. There are also a number of entertaining activities located nearby, such as shopping at historic Station Square, riding the inclines to Mount Washington where you can overlook the city, swimming at the Sand Castle water park or visiting Kennywood—"the roller coaster capital of the world."

Contributed Papers

The deadline for submitting a contributed paper has been extended until May 15, 1995 for abstracts submitted by email (`interface95@stat.cmu.edu`). Papers can be a demo, presentation, or poster session and are approximately 20 minutes long. Contributors will be notified by 5/16/95, as to whether or not their paper was accepted. Draft copies of the paper should be submitted by May 15th.

Proceedings

Invited and contributed papers should be submitted for the proceedings by July 15, 1995 at the Interface '95 address.

Financial Support

Limited funds are available to support travel and per diem expenses of young researchers and graduate students. Preference will be given to those who will be presenting papers. Please apply to Mike Meyer and Jim Rosenberger (email: interface95@stat.cmu.edu). Applicants will be notified by 4/30/95 as to whether or not they will receive financial support.

Interface '96

Interface '96 will be in Sydney, Australia from July 8-10, 1996 and will be chaired by Nick Fisher. Email inquiries can be sent to sydney96@syd.dms.csiro.au.

Short Courses

Organized by: Tom Devlin, Montclair State University Karen Kafadar, University of Colorado

Location: David L. Lawrence Convention Center - South Meeting Rooms Sponsored by: The Statistical Computing and Graphics Sections of the ASA

- "Tools for Discovering Patterns in Data", John F. Elder and Paul Hess, Wednesday, 9:30-Noon $60
- "Trellis Displays", Richard A. Becker, Ming Shyu, and William S. Cleveland Wednesday, 1:00-5:00 PM $60
- "Using tcl/tk in Building Interfaces to Statistical Software and Robust Visualization", Joseph W. McKean, John Kapenga, Thomas J. Vidmar, Wednesday, 9:30-Noon $60
- "LaTeX for Statisticians", Brian W. Junker, Wednesday, 1:00-3:00 PM $35
- "HTML and Creating World Wide Web Pages", Bob Kuszewski, Wednesday, 3:00-5:00 PM $35

The Invited Program

Keynote Speaker: The keynote address will be given by Raj Reddy, Dean of Carnegie Mellon University's School of Computer Science and the Herbert A. Simon University Professor of Computer Science and Robotics. His research interests include the study of human-computer interaction and artificial intelligence. He was the recipient of the IBM Research Ralph Gomory Fellow Award in 1991, and is the most recent recipient of the A.M. Turing Award.

His address is entitled "Statistics, Computation, and Artificial Intelligence."

Environmental Program

- "Model Uncertainty Assessment for Air Quality Data", Organized by Lawrence Cox, EPA RTP
- "Modelling Systems and Model Validation", Organized by Robert Teitel, Abt Associates
- "Spatial Statistics for Environmental Data", Organized by Noel Cressie, Iowa State University
- "Modelling Environmental Systems", Organized by David M. Allen, University of Kentucky
- "Biodiversity, Geographic Information and Statistics", Organized by G.P. Patil, Penn State University
- "Graphical Methods for Display of Environmental Data", Organized by Dan Carr, George Mason University

Manufacturing Program

- "Quality in Manufacturing", Organized by Nick Fisher, CSIRO, Australia
- "Statistics and Computer Science in VLSI Semiconductor Fabrication", Organized by John P. Lehoczky, Carnegie Mellon University
- "Issues in Coordinate Measurement Analysis", Organized by Fred Hulting, Alcoa, Pittsburgh
- "GUI Interfaces", Organized by Michael M. Meyer, Carnegie Mellon University

Imaging/Graphical Data

- "Statistics and Medical Imaging", Organized by Kanti Mardia, Leeds University (UK)
- "Image Analysis and Remote Sensing", Organized by Jeffrey D. Helterbrand, Eli Lilly, Indianapolis
- "MRI and Medical Imaging", Organized by William F. Eddy, Carnegie Mellon University

Other Sessions

- "Computational Methods for Seismology", Organized by Tim Hesterberg, Franklin & Marshall College
- "Best of the JCGS", Organized by William Kennedy, Iowa State University
- "Beyond Correlation—O' the Many Faces of Cause", Organized by Arnold Goodman, UCLA
- "Huge Data Sets", Organized by Ed Wegman, George Mason University

- "Teaching Statistics through Multimedia and the Internet", Organized by James L. Rosenberger, Penn State University
- "Statistical Numerical Integration", Organized by Alan Genz, Washington State University

Exhibits

There will be an exhibit area in South Meeting Room 10 of the David L. Lawrence Convention Center. The exhibits will be open on both Thursday and Friday, June 22 and 23, from 10:00 AM to 6:00 PM.

If you interested in being an exhibitor, please contact Mari Alice McShane at the Interface '95 office.

Hotel Reservation Request

Interface '95, June 21-24, 1995 Reservations due by: May 30, 1995

Hotel information

The Pittsburgh Vista, conveniently located in downtown Pittsburgh, has rooms reserved for conference participants. Reservations should be made by May 30, 1995. Room rates are: $105 single, $110 double, plus 12% room tax. Please make your reservations directly with the hotel:

By mail: ATTN: Reservation Department
Pittsburgh Vista Hotel
1000 Penn Avenue
Pittsburgh, PA 15222-3873
By phone: (412) 281-3700, or Fax (412) 227-4505

Please mention the conference when making your reservations.

If you need additional alternative hotel information , please contact Interface '95

Interface participants who wish to share hotel rooms may send an email message to Tim Hesterberg at `tim@fnmcps.fandm.edu`. Be sure to include your name, nights you expect to stay, gender, and whether or not you smoke. Tim will contact you about your housing arrangements.

Conference Registration Form

The conference registration form is available by sending e-mail to `interface95@stat.cmu.edu` or the WWW address listed above.

CONFERENCE NOTICES

NTTS-95

International Conference on New Techniques and Technologies for Statistics (NTTS-95)

Bonn, Germany
November 19-22, 1995

The impact of new information and communication technologies on statistics is rapidly growing and there is much interest in new developments within the application framework of official statistics. The purpose of NTTS-95 is to provide a forum for researchers and (official) statisticians to present and discuss new ideas and developments in the application of information and communication technologies for statistics. The main topics of interest will include:

- Survey design and data capture
- Data analysis and knowledge extraction
- Dissemination of results and knowledge

The aim is to bring together researchers and users of these techniques, to share experience, knowledge and enthusiasm in both formal and informal environments. The program will include invited and contributed talks, demonstrations, and panel discussions. The conference proceedings will be published.

Papers (up to 12 pages) are due by June 16, 1995. Further information can be found on WWW: `http://orgwis.gmd.de/explora/ntts.html`

Program Committee

Ph. Nanopoulos, EUROSTAT, Luxembourg (Conference Chair) A. Unwin, University of Augsburg, Germany (Conference Chair) W. Begeer, Barendrecht, The Netherlands; A. Sousa da Camara, New University of Lisbon, Portugal; D. Conniffe, ESRI, Ireland; D. Defays, EUROSTAT, Luxembourg; E. Diday, INRIA, France; J.F. Grandin, Thomson, France; D.J. Hand, The Open University, Milton Keynes, UK; D. Heath, EUROSTAT, Luxembourg; T.C. Jones, CSO, UK; W. Kloesgen, GMD, Germany; W. Kuehn, Statistisches Bundesamt, Germany; H.J. Lenz, Free University of Berlin, Germany; J. Ludley, EUROSTAT, Luxembourg; F. Murtagh, Munotec Systems, Germany; E. Outrata, Statistical Office Czech Republic; G. Piatetsky-Shapiro, GTE, USA; D. Pregibon, ATT, USA; D. Rhind, Ordnance Survey, UK; J.L. Roos, INSEE, France; and B. Sundgren, Statistics Sweden.

Neural Information Processing Systems

CALL FOR PAPERS

Neural Information Processing Systems
-Natural and Synthetic-
Monday, Nov. 27 - Saturday, Dec. 2, 1995
Denver, Colorado

This is the 9th meeting of an interdisciplinary conference bringing together neuroscientists, engineers, computer scientists, cognitive scientists, physicists, and mathematicians interested in all aspects of neural processing and computation. The conference includes invited talks, and oral and poster presentations of refereed papers. There are no parallel sessions. Also planned are one day of tutorial presentations (Nov 27) preceding the regular session, and two days of focused workshops following at a nearby ski area (Dec 1-2). Major categories for paper submission, with example subcategories, are as follows:

Neuroscience: systems physiology, cellular physiology, signal and noise analysis, oscillations, synchronization, inhibition, neuromodulation, synaptic plasticity, computational models.

Theory: computational learning theory, complexity theory, dynamical systems, statistical mechanics, probability and statistics, approximation theory.

Implementation: VLSI, optical, parallel processors, software simulators, implementation languages.

Algorithms and Architectures: learning algorithms, constructive/pruning algorithms, localized basis functions, decision trees, recurrent networks, genetic algorithms, combinatorial optimization, performance comparisons.

Visual Processing: image recognition, coding and classification, stereopsis, motion detection, visual psychophysics.

Speech, Handwriting and Signal Processing: speech recognition, coding and synthesis, handwriting recognition, adaptive equalization, nonlinear noise removal.

Applications: time-series prediction, medical diagnosis, financial analysis, DNA/protein sequence analysis, music processing, expert systems.

Cognitive Science & AI: natural language, human learning and memory, perception and psychophysics, symbolic reasoning.

Control, Navigation, and Planning: robotic motor control, process control, navigation, path planning, exploration, dynamic programming.

Review Criteria: All submitted papers will be thoroughly refereed on the basis of technical quality, novelty, significance and clarity. Submissions should contain new results that have not been published previously. Authors are encouraged to submit their most recent work, as there will be an opportunity after the meeting to revise accepted manuscripts before submitting final camera-ready copy.

Paper Format: Submitted papers may be up to eight pages in length, including figures and references. The page limit will be strictly enforced, and any submission exceeding eight pages will not be considered. Authors are encouraged (but not required) to use the NIPS style files obtainable by anonymous FTP at the sites given below. Papers must include physical and e-mail addresses of all authors, and MUST indicate one of the nine major categories listed above. Authors may also indicate a subcategory, and any preference for oral or poster presentation; this preference will play no role in paper acceptance.

Submission Instructions: Send six copies of submitted papers to the address given below; electronic or FAX submission is not acceptable. Include one additional copy of the abstract only, to be used for preparation of the abstracts booklet distributed at the meeting. Submissions mailed first-class from within the US or Canada, or sent from overseas via Federal Express/Airborne/DHL or similar carrier must be POSTMARKED by May 20, 1995. All other submissions must ARRIVE by this date. Mail submissions to:

Michael Mozer, NIPS*95 Program Chair
Department of Computer Science
University of Colorado
Colorado Avenue and Regent Drive
Boulder, CO 80309-0430 USA

Mail general inquiries/requests for registration material to:

NIPS*95 Registration
Dept. of Mathematical and Computer Sciences
Colorado School of Mines
Golden, CO 80401
FAX: (303) 273-3875
`nips95@mines.colorado.edu`

Sites for LaTex style files: Copies of "nips.tex" and "nips.sty" are available via anonymous ftp at `helper.systems.caltech.edu` (131.215.68.12) in `/pub/nips`, or at `b.gp.cs.cmu.edu` (`128.2.242.8`) in `/usr/dst/public/nips`.
The style files and other conference information may also be retrieved via World Wide Web at: `http://www.cs.cmu.edu:8001/afs/cs/project/cnbc/nips/NIPS.HTML`

NIPS*95 Organizing Committee: General Chair, David S. Touretzky, CMU; Program Chair, Michael Mozer, U. Colorado; Publications Chair, Michael Hasselmo, Harvard; Tutorial Chair, Jack Cowan, U. Chicago; Workshops Chair, Michael Perrone, IBM; Publicity Chair, David Cohn, MIT; Local Arrangements, Manavendra Misra, Colorado School of Mines; Treasurer, John Lazzaro, Berkeley.

Submission Deadline is May 20, 1995 (Postmarked)

SECTION OFFICERS

Statistical Graphics Section - 1995

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

William DuMouchel, Chair-Elect
212-305-7736
Columbia University
dumouch@bayes.cpmc.columbia.edu

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

Sally C. Morton, Program Chair (appointed to replace Sallie Keller-McNulty)
310-393-0411 ext 7360
The Rand Corporation
Sally_Morton@rand.org

Stephen G. Eick, Program Chair-Elect
708-713-5169
AT&T Research Laboratories
eick@research.att.com

Michael M. Meyer, Newsletter Editor (93-95)
412-268-3108
Carnegie Mellon University
mikem@stat.cmu.edu

Robert L. Newcomb, Secretary/Treasurer (95-96)
714-824-5366
University of California, Irvine
rnewcomb@uci.edu

Deborah J. Donnell, Publications Officer (94-96)
206-283-8802 ext 258
MathSoft, Seattle, WA

Colin R. Goodall, Rep.(95-97) to Council of Sections
814-865-3993
The Pennsylvania State University
colin@stat.psu.edu

Jane F. Gentleman, Rep.(94-96) to Council of Sections
613-951-8213
Canadian Centre for Health Information
GENTLEJF@NRCVM01.bitnet

Sally C. Morton, Rep.(94-95) to Council of Sections
310-393-0411 ext 7360
The Rand Corporation
Sally_Morton@rand.org

Statistical Computing Section - 1995

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

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

Trevor J. Hastie, Past Chair
Stanford University
415-725-2231
trevor@playfair.stanford.edu

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

Robert J. Tibshirani, Program Chair-Elect
416-978-4642
University of Toronto
tibs@utstat.toronto.edu

James L. Rosenberger, Newsletter Editor (93-95)
814-865-1348
The Pennsylvania State University
JLR@stat.psu.edu

Terry M. Therneau, Secretary-Treasurer
507-284-1817
Mayo Clinic
therneau@mayo.edu

Karen Kafadar, Publications Liaison Officer
303-556-2547
University of Colorado-Denver
kk@tiger.cudenver.edu

MaryAnn H. Hill, Rep.(95-97) to Council of Sections
312-329-2400 SPSS
hill@spss.com

Michael M. Meyer, Rep.(95-96) Council of Sections
412-268-3108
Carnegie Mellon University
MikeM@stat.cmu.edu

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

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

1855

Department of Statistics
University Park, PA 16802-2111

Nonprofit Organization
U. S. POSTAGE
PAID
Permit No. 1
University Park, PA 16802

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.