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

April 96 Vol.7 No.1

Statistical COMPUTING & GRAPHICS

A WORD FROM OUR CHAIRS

Statistical Computing

Sallie Keller-McNulty is the 1996 Chair of the Statistical Computing Section. She welcomes your feedback and comments on any issue of interest to your section.

I would like to open my first column by thanking Jim Rosenberger, Pennsylvania State University, and Mike Meyer, Carnegie Mellon University, for their dedication and hard work over the last three years as our *Statistical Computing and Graphics* Newsletter Editors. Through their efforts we now have an extremely professional and informative Newsletter that appears three times a year.

Our new Newsletter Editors are Mark Hansen, Bell Laboratories, and Mario Peruggia, Ohio State University. In the process of preparing their first issue (the one you are holding!) I am sure they have become aware of the excitement and challenges that lie ahead. I am confident the Sections will greatly benefit from the fresh perspective they will bring to the Newsletter. I also hope they find this job as satisfying and rewarding as Mike Meyer, Jim Rosenberger, Dan Carr, George Mason University, and I have!

Now I would like to update you on some of the Statistical Computing Section's activities. Most of the activities I will describe occurred during 1995 under the distinguished leadership of Mary Ellen Bock, the Past Chair of our Section.

Our Section will have a high profile at the 1996 Joint Statistical Meetings this August. This is primarily due to the efforts of Rob Tibshirani, University of Toronto, Jim Rosenberger, Tom Devlin, Montclair State College, Trevor Hastie, Stanford University, and Daryl Pregibon, AT&T Laboratories.

CONTINUED ON PAGE 2

A WORD FROM OUR CHAIRS

Statistical Graphics

Bill DuMouchel is the 1996 Chair of the Statistical Graphics Section. In his first note to the section he encourages feedback, directly to the chair, or with letters to the Editors of the Newsletter.

My first order of business as incoming Chair of the Graphics Section is to thank David Scott for his leadership and hard work as Chair last year. It is also very appropriate to thank our outgoing Newsletter Editors, Jim Rosenberger and Mike Meyer, for making the joint Computing-Graphics Newsletter the finest of all the ASA Section Newsletters, and to wish the best to our new Editors, Mark Hansen and Mario Peruggia, as they try to match or even surpass previous achievements.

Next I want to congratulate Stephen Eick, this year's Program Chair, for putting together a great set of sessions for the August Joint Statistical Meetings in Chicago. We have been allotted four invited sessions:

- Information Visualization, organized by Stephen Eick
- Transactional Data Analysis, organized by Daryl Pregibon
- Innovations in Graphics, organized by Stephen Eick
- Statistical Graphics and Multimedia Education, organized by David Scott

(The last-listed session is co-sponsored by the Section on Statistical Education.) The sessions include several speakers from outside the regular statistics community, so we are sure to learn many new graphics tricks in Chicago. Steve describes the sessions in detail elsewhere in this issue.

CONTINUED ON PAGE 3

EDITORIAL

Our First Newsletter!

With this April 96 issue (Vol. 7 No. 1) we initiate our terms as joint Editors of the Newsletter. As we take over from where Jim Rosenberger and Mike Meyer left off, all of our readers familiar with Jim's and Mike's accomplishments will agree that we have our work cut out for us if we wish to maintain the quality standards that the Newsletter has achieved over the years.

In keeping with the time honored adage "If it ain't broke, don't fix it," we have made no changes to the layout of the publication and continue to rely on several of our regular columnists. The only important, logistic novelty, that should be transparent to the reader, is that the Newsletter is now published and printed by Lucent Technologies (the new home of Bell Laboratories) instead of the Statistics Department at PennState. We would like to thank both Jim and Mike for easing the transition with their helpful guidance and technical support.

This issue gives our readers the opportunity to meet our new Section Chairpersons: Sallie Keller-McNulty (Computing) and Bill DuMouchel (Graphics). In their columns, they bring us up to date on the state of our sections and outline their plans and expectations for the future. The ASA's World-Wide Web site is now online and both Sallie and Bill are interested in obtaining feedback from our readers (see also the related articles on pages 5 and 6).

In an article on page 6, David James discusses context-rich graphical displays for data analysis and presents two interesting case studies to illustrate how analysis of variance decompositions and graphical displays can be used in combination to study longitudinal and spatial dependencies. Daniel Carr and Anthony Olsen have contributed an intriguing article (page 10) on the use of multivariate sorting as an aid to simplifying the visual appearance of graphical displays.

In "Net Snooping," Mike Meyer and Dennis Pearl talk about two related educational projects: DASL and EESEE. These are two electronic libraries of data and related stories to be used for teaching. While neither project can be strictly classified as computational or graphical, they both contribute to the important task of familiarizing undergraduate students with data analysis and the use of computers.

The Invited Session Programs at the Chicago Meetings look quite exciting and details can be found in the article on page 18. Also, look for the Interface '96 announcement on page 20 (this year's meeting will be in Sydney, Australia) and for the announcement of the Sixth International Workshop on Artificial Intelligence and Statistics on page 21.

The most difficult task we face, as Editors, is obtaining contributions from our readers. This is your Newsletter and we wish to encourage you to contact us via e-mail if you intend to contribute an article. Letters to the Editors and shorter information pieces would also be welcome. We would prefer to receive contributions in LaTeX but plain ASCII text files would also be acceptable. If you would like to meet us in person to discuss issues related to the Newsletter, there will be a roundtable luncheon discussion on this topic at the Joint Meetings in Chicago.

Mark Hansen
Editor, Statistical Computing Section
Bell Laboratories
cocteau@bell-labs.com

Mario Peruggia
Editor, Statistical Graphics Section
The Ohio State University
peruggia@stat.mps.ohio-state.edu

FROM OUR CHAIRS (Cont.). . .

Statistical Computing

CONTINUED FROM PAGE 1

Our 1996 Program Chair Rob Tibshirani has coordinated an excellent Invited Session Program, as well as organizing all of our contributed papers and posters into synergistic sessions. Jim Rosenberger, our 1997 Program Chair has organized five interesting Roundtable Luncheon Discussions. Tom Devlin has worked with a variety of our members preparing continuing education proposals, resulting in three short courses being accepted into the program. Trevor Hastie and Daryl Pregibon coordinated this year's student paper competition. They had the very difficult task of selecting four winners from among the submissions. The Section is providing financial support for the winners to attend the Meetings and present their papers. Details of the Invited Program and the Student Paper Competition winners can be found elsewhere in this issue.

I expect our Section will have a strong presence in

the 1997 Joint Statistical Meetings. Jim Rosenberger is already beginning to flesh out ideas for invited sessions. He would appreciate any input you can provide (jrl@stat.psu.edu). Jon Kettenring, the 1997 ASA President (and an active Statistical Computing Section member), has chosen the theme, "Shaping Statistics for Success in the 21st Century," for the 1997 Meetings. I will be joining Jim and Jon in their efforts with the 1997 Meetings through my position as overall 1997 Joint Statistical Meetings Program Chair. We hope to bring more technology into the 1997 Meetings, and we hope to use more technology in the planning of the meetings. Keep your eye on ASA's web page (http://www.amstat.org/) for our progress.

Our Section has been instrumental in bringing ASA "online." In addition to providing start-up funding for ASA to gain internet access, our Section's members, such as Lorraine Denby, Bell Laboratories, Mike Conlon, University of Florida, Dan Jacobs, University of Maryland, David Morgenstein, WESTAT, Daniel Solomon, North Carolina State University, and Mike Meyer have provided much of the leadership for the Council of Sections project of creating an ASA web page. This effort has not been as simple as it may sound. Consider the fact that 18 months ago ASA started from square zero: no equipment and no idea of how to design or implement a home page. Now take a look at http://www.amstat.org/. Next, thank Lorraine, Mike, Dan, David, Daniel, and Mike! We MUST continue to provide ASA with leadership and direction to keep ASA moving into the information age. In the articles appearing on pages 5 and 6, Mike Meyer, Lorraine Denby and I provide more information about the status of ASA's web and electronic communication activities and how you can provide input.

In 1995 our Section, along with the Statistical Graphics Section and the Statistical Education Section, provided funding for an Undergraduate Data Analysis competition, organized by David Robinson, St. Cloud State University. We found last year's competition to be a great success, and our Executive Committee has voted to fund the competition again this year. This year's competition is being organized by Esteban Walker, University of Tennessee (ewalker@utkvx.utk.edu). I am sure Esteban will report to us on this year's competition as it progresses.

As you can see from this column, our Section's dues have supported many useful activities. In addition to those discussed above, your dues help the Statistical Computing and the Statistical Graphics Sections jointly sponsor the best mixer of the Joint Statistical Meetings. This year our mixer will be Monday night, August 5th, starting at 7:30 p.m. Don't miss our open bar and stupendous snacks!

Sallie Keller-McNulty
National Science Foundation
smcnulty@nsf.gov

FROM OUR CHAIRS (Cont.). . .

Statistical Graphics

CONTINUED FROM PAGE 1

If you haven't checked out our new web site yet, please do! To reach it, first access the ASA web site, at http://www.amstat.org, next click on "Sections" and then on "Statistical Graphics." Thanks again to Bob Newcomb for creating and tending the site. If you want to place a link directly to our site, its actual address is http://orion.oac.uci.edu/~rnewcomb/statistics/graphics/graphics.html

To my mind, the most important part of the web site is the feedback box. Just click on it and you are prompted to send a message to the officers of the Section on Statistical Graphics. Here are some topics on which we need advice from you:

- Invited sessions that the Graphics Section could present in the 1997 meetings
- Data sets to be used in our 1997 Contest on Statistical Graphics
- More content to include at our web site
- Projects of moderate cost (at most a few thousand dollars) that our Section should undertake
- Graphical ideas or techniques that deserve wider recognition
- Reaction to Newsletter articles

I hope that this list is long enough to give you ideas of the type of advice we need, but all feedback is welcome, of course. Thanks to everyone for helping keep our Section on an upward path.

Bill DuMouchel
Columbia University
dumouchel@columbia.edu

Results of Student Paper Competition

by Trevor Hastie and Daryl Pregibon

The results are in! We have selected the four winners of the Statistical Computing Section's 1996 Student Paper Competition. In alphabetical order they are

- **Dmitrii Danilov**, St. Petersburg State University: *Principal Components in Time Series Forecasting*
- **Ranjan Maitra**, University of Washington: *Estimating Precision in Functional Images*
- **Bob Mau** and Michael Newton, University of Wisconsin, Madison: *Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods*
- **Chris Volinsky**, with David Madigan, Adrian Raftery and Richard Kronmal, University of Washington: *Applying Bayesian Model Averaging to Cox Models*

Before we tell you more about each of the winners, a bit more about the competition itself. This is our second such competition, and this year we had 18 entries. 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. They also get their names and faces in this Newsletter as well as the Amstat news—publicity gives a great kick start to a career!

Details on the Winners

Dmitrii Danilov

Dmitrii is a third year Ph.D. student in the Mathematics & Mechanics Department of the St. Petersburg State University. His scientific advisor is Professor Sergey Ermakov. He expects to receive his Ph.D. next year in numerical methods. His research interests include time series forecasting and stochastic complexity theory.

Principal Components in Time Series Forecasting

A new time series forecasting method is proposed and studied. The main tool of this method is the principal component analysis applied to the data matrix derived from the initial time series by the procedure known as "the delay method." The method under consideration is introduced for a particular class of deterministic functions of a discrete argument, this class consists of functions satisfying linear finite difference equations with constant coefficients. Some results concerning the existence and uniqueness of the forecast as well as numerical examples are considered for this class of functions. Then the method is generalized to the case of general functions of a discrete argument including stochastic time series. Several examples of application of the method to the well known data sets are presented.

Ranjan Maitra

Ranjan is a final year Ph.D. student in the Department of Statistics at the University of Washington. His thesis is titled "Variability Estimation and Model Evaluation in Some Inverse Problems." Ranjan's advisor is Professor Finbarr O'Sullivan. His career goal is to teach and pursue research in a university. His research interests include multivariate analysis, inverse problems, spatial statistics, model validation, analysis of directional data, spatial statistics and image analysis.

Estimating Precision in Functional Images

Functional imaging of biologic parameters like *in vivo* tissue metabolism is made possible by Positron Emission Tomography (PET). Estimating precision of these images is important for drawing inferences on biologic activities. Analytic expressions are intractable and usual bootstrap methods of assessing variability computationally impractical. We suggest an approximate parametric bootstrap approach which eliminates much of the computational overhead. Results on a simplified model chosen to match PET reconstruction characteristics are very encouraging.

Bob Mau

Bob is completing his doctoral thesis under Professor Michael Newton in the Statistics Department at the University of Wisconsin - Madison. Bob plans to build on his experience as a statistical consultant at the Engineering Research Center for Plasma-aided Manufacturing, and pursue a career as an industrial statistician. He would welcome the opportunity to expand the methodology developed in his thesis and make his computer programs useable to researchers in the life sciences.

Bayesian Phylogenetic Inference via Markov Chain Monte Carlo Methods

Using standard probability models for the evolution of discrete data within a group of organisms, we derive a Markov chain to simulate a Bayesian posterior distribution on the space of binary trees that mimic the speciation process. A transformation of the tree into a canonical cophenetic matrix

form suggests a remarkably simple proposal distribution for selecting candidate trees "close" to the current tree in the chain. We illustrate the algorithm with a binary dataset for nine species of plants, then extend to DNA sequences from thirty-two species of fish.

Chris Volinsky

Chris is a fourth year Ph.D. student at the University of Washington under the guidance of Professors Adrian Raftery and David Madigan. His statistical interests include Bayesian model section, graphical models, statistics education, MCMC techniques and any serious statistical applications to baseball. He received his B.A. in Mathematics and Statistics from SUNY Buffalo.

Applying Bayesian Model Averaging to Cox Models

In this paper we apply Bayesian model averaging to evaluate risk factors for stroke and to assess individual stroke risk. We introduce a technique based on the leaps and bounds algorithm which efficiently locates and fits the best models in the very large model space and thereby extends all subsets regression to Cox models. For each independent variable considered, the method provides the posterior probability that it belongs in the model. Using our S-Plus function `bic.surv` to implement this procedure, we show that Bayesian model averaging predictively outperforms standard model selection methods for assessing stroke risk.

The Future...

We intend to hold such a competition every year, and the papers are due in early January. Registered students are eligible to submit papers. See the announcement on page 22 of this Newsletter for more details of the competition and conditions for entry. Professors, please make a note in your calendar to get your students geared up in time for next year's submissions. The deadline for submissions for the 1997 competition is January 10, 1997.

Trevor Hastie
Stanford University
`trevor@playfair.stanford.edu`

Daryl Pregibon
AT&T Laboratories
`daryl@research.att.com`

FROM THE ASA

What electronic services should the ASA provide?

By Sallie Keller-McNulty and Mike Meyer

The ASA Sections came together to provide the start-up funds to set up the World-Wide Web (WWW) server for ASA. Last August the ASA's web site came online (`http://www.amstat.org/`), and ASA now has an official Internet presence. Now that the initial phase of developing a web site for ASA has been completed, it would be useful for members to voice what WWW services they would like to see ASA provide.

Have you ever looked at the combined membership list of the American Mathematical Society (AMS), Mathematical Association of America (MAA), Society for Industrial and Applied Mathematics (SIAM), and American Mathematics Association of Two Year Colleges (AMATYC), provided by the American Mathematical Society at `http://www.ams.org/`? Try it and you will quickly find the names, addresses, phone numbers, memberships, etc. of all 10 members at the University of Western Australia, for example. Would it be appropriate to provide a similar service from the combined membership directory published by the ASA? As an organization that has concerns about data access **and** confidentiality, should the ASA be restrictive in the information it provides about its members? Before you answer that, we encourage you to see if you can find your own phone number and address at `http://www.switchboard.com` or see how much information the American Economics Association provides online about its members `http://www.eco.utexas.edu:80/AEA/`. The ACM also advertises that they will soon have available an online membership directory.

What other electronic services might you like the ASA to provide? How about listing the tables of contents of current ASA journals? What about employment opportunities? (Check out what's available from `http://chronicle.merit.edu/.ads/.links.html` and `http://www.ams.org/committee/profession/employ.html` and `gopher://vuinfo.vanderbilt.edu/11/employment/joe` before you answer.)

The issues of what electronic services the ASA should supply, and at what cost are being addressed by two ASA committees on electronic communications. The two ASA committees are chaired by Lorraine Denby, `ld@bell-labs.com` and Kathleen Lamborn,

lambornk@neuro.ucsf.edu. These committees will be preparing reports to present to the ASA Board of Directors. The issues quickly become quite complicated and one could argue that it is best to allow the respective committees to do their work and write their reports. However, not being known as passive Sections when it comes to discussions involving technology, we felt that our members might like to express an opinion. Send e-mail comments to the ASA board (via Ray Waller, Executive Director of ASA ray@amstat.org), the above committees or to either of us.

Sallie Keller-McNulty
National Science Foundation
smcnulty@nsf.gov

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

Online Membership Listing is APPROVED!

By Lorraine Denby

As Chair of the COS Electronic Communication Committee, I am pleased to announce that the ASA Board of Directors has approved the online membership listing. In the coming months, we will launch a large-scale effort to alert our membership of this venture, giving people the option to be excluded from this listing. The decision to adopt an online listing with "passive permission" was based on our desire to move this effort forward in a conscientious, timely fashion, balancing the concerns of our members with the costs that would be involved in adopting various alternatives. Here's what's been approved so far:

- Ultimately, the online listing can include an individual's name, address, phone, fax, email, section and chapter affiliation.
- For the next three months, expect to see a notice prominently displayed on the front page of the AMSTAT News describing the online membership service and explaining how individuals can have their information excluded from the listing.
- Letters will be sent to corporate chairs informing them of the online membership listing asking them to contact us if they do not want their employee's addresses listed.
- The online membership service and its search capabilities will be demonstrated at the annual meetings in Chicago. The service will be available on the World Wide Web shortly after this demo.
- Members will always have the opportunity to opt out (or back into) the membership listing. The change will be effective within a week of when the staff processes the request.

We are very excited about bringing the ASA online. If you have any comments about the plan outlined above, feel free to contact me at the e-mail address listed below.

Lorraine Denby
Bell Laboratories
ld@bell-labs.com

CASE STUDIES IN INDUSTRIAL STATISTICS

Context-Rich Graphical Displays

By David A. James

1. Introduction

In this short communication we briefly discuss the idea of context-rich graphical displays for exploratory data analysis. We illustrate this concept through two case studies and attempt to distill a handful of useful principles that may be used when analyzing data from other areas.

If we consider graphical techniques such as scatter plots, histograms, boxplots, and quantile plots as context-free in the sense that they are applicable to data sets collected in any field of interest, then context-rich graphics refers to displays that are highly tailored to specific applications. This allows us, for instance, to augment familiar displays with information that is often implicit but not fully exploited by field investigators. By making explicit this information through graphics we are able to better articulate the analysis' goals, methods, conclusions, and limitations.

This idea has a long tradition. Many well-known displays are context-rich; for instance, Napoleon's failed Russian campaign (Tufte, 1983), Playfair's wheat price, wages, and the reigns of British kings from 1565 through 1821 (Tufte, 1983), cholera outbreak in London's 1850's (Cliff and Ord, 1981); more recent examples include EVENTCHARTS (Goldman, 1992), Caveplots (Becker

et al., 1994), and SeeNet (Becker et al., 1991). Extra information is effectively displayed in the context of the application in all of the above mentioned displays, e.g., weather hardships that Napoleon's army experienced, the steady economical improvement of the British empire, censoring in the case of EventCharts, and network topology in the case of SeeNet (see Cleveland 1993; Tufte 1983; Tufte 1990) for many more examples.

2. Case Studies

In the following two case studies we show how graphical displays in combination with analysis of variance decompositions allows us to study the longitudinal and spatial dependence of various manufacturing metrics (routinely collected summary statistics) on measured covariates. This approach is semi-parametric in that no longitudinal/spatial structure is imposed on either the response or its covariates. In both examples, we consider data collected from designed experiments. For each factor level combination in the design we observe a (possibly replicated) multivariate response. In the first example, this response represents measurements taken at various points along an optical fiber, while in the second example our response will be a final functionality test applied to all the devices fabricated on a silicon wafer. These responses are first smoothed and an analysis of variance model is then fitted at each location. We will fit linear models of the form

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{B} + \boldsymbol{E}$$

where all elements in the model are matrices; $(y_{i,p})$ denotes the smoothed response for the i-th observation at the p-th position, $\boldsymbol{X}$ is a design matrix corresponding to some parametrization of the covariates, $(\beta_{k,p})$ is the coefficient for the k-th term at the p-th position, and $(\epsilon_{i,p})$ is the residual from the i-th observation at the p-th position. From this process we get multiple sets of coefficients and effects (one set per location) that we display longitudinally along a fiber and spatially on a wafer.

Optical Fiber

Two important quality characteristics of the transmission along optical fiber cables are the power attenuation and dispersion as the light signals travel along the fiber. Their measurement is done with an optical time-domain reflectometer (OTDR) that sends laser pulses along the fiber and measures their back-scatter intensities at equally-spaced points. To measure the fiber attenuation at a single position on the fiber, multiple signals of specified width are sent and their back-scatter averaged. This process is repeated for positions $p_1, p_2, \ldots, p_n$, to form a collection of power measurements (in dB) that traces the signal attenuation along the fiber. Figure 1 shows one such trace.

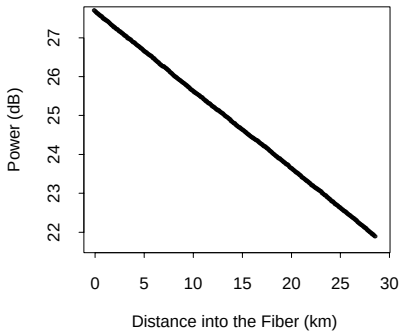


Figure 1. *Power along an optical fiber.*

Although not distinguishable from Figure 1, the variance of the measured attenuation is largerly dependent on position (i.e., as the signal travels further into the fiber, the measurement error is known to increase). At the time of the experiment, various factors were known to have a possible effect on the variance, but the form of this dependence was not understood in the presence of manufacturing variability.

To characterize the OTDR measurement error along the fiber, a 24-run full factorial experiment was conducted varying the power of the input laser or "plugin" (low versus high), the width of individual light pulses (short, medium and long), and the number of pulses that are averaged at each point on the fiber (39, 100, 150, and 200). At each condition 100 replicates were taken and their empirical variances were computed. These somewhat noisy estimates were smoothed using loess (Cleveland, 1979) and are displayed in Figure 2.

Traditionally, experiments like this have been analyzed by first reducing the response (the variance traces) to scalars (e.g., one-dimensional summary statistics like mean or median variance). Then an analysis of variance would be conducted on these summaries. This approach ignores the possible dependence of the factors on positions along the fiber; if the effects are monotone and significantly large, adequate settings may be concluded, but the intrinsic variability along the fiber could not be assessed.

Instead, without too much extra work, we may estimate the effect of the design factors at fixed positions along the fiber. Had we had information as to the functional form of the variance traces as the factors vary, we would had modeled it directly. Since we did not, we fitted models at one-kilometer intervals to $\log(\hat{\sigma})$, and then collected the terms into traces to understand the positional effects of the design factors. Figure 3 shows how the fitted effects vary along position on the fiber.

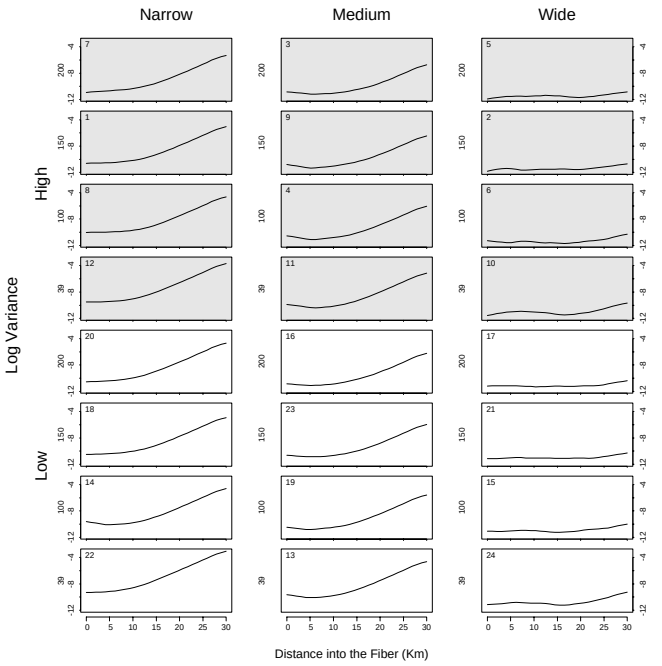


Figure 2. *Design settings from a full-factorial experiment. The top 12 shaded panels correspond to runs with high plugin and the bottom 12 panels with low plugin. Each column depicts the response at one pulse width. Each row shows the response for one level of number of averages: 200, 150, 100, and 39. The number in the upper left-hand corner of each panel indicates the order of experimentation.*

We see, for instance, that the two curves corresponding to the terms in pulse width contribute the most to measurement error, also apparent from Figure 1. Moreover, their contribution is nonlinearly increasing as a function of position. Similarly, we see that the "plugin" effect of input laser power (high better than low), is significant and somewhat uniform along the fiber. Finally, only one term from number of signal averages per measurement is shown to be significant and linearly increasing along the fiber: this term contrasts averaging 39 signals versus 100 or more. We concluded that 100 or more averages is adequate.

Integrated Circuits on a Wafer

Integrated-circuits or "chips" are manufactured on silicon wafers. These wafers are circular disks that may contain from 40 up to 800 or more chips. Before these wafers are cut and the chips packaged, a battery of tests are applied to each chip to assess whether it is defective or not (for this presentation we will not distinguish between the various tests, labeling as defective chips that fail at least one test).

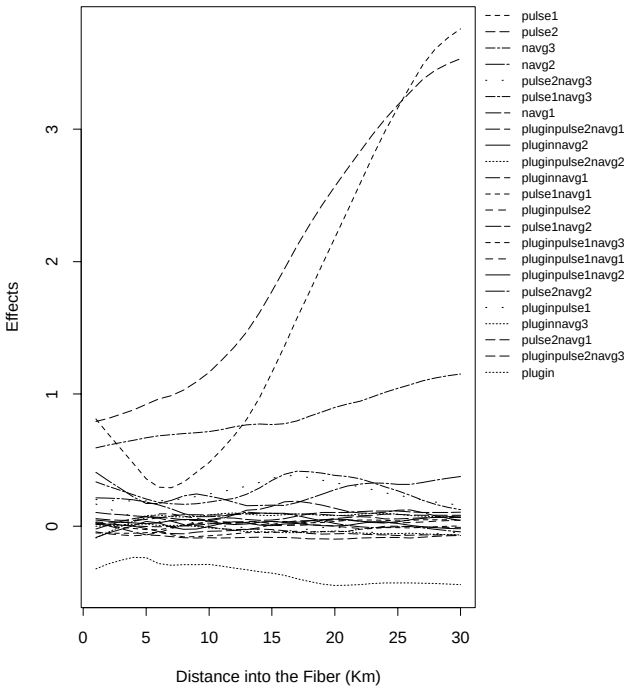


Figure 3. *Effect traces from the full ANOVA model.*

At each site on a wafer, we estimate its probability of being defective by applying a smoother (we use a bivariate kernel smoother, but others may also be used) and then apply a sequence of transformations to stabilize the variance of these estimates. Figure 4 shows one binary wafer where black squares denote defective chips and white squares denote non-defective chips; the wafer on the right shows a smoothed version of the binary wafer where dark areas denote regions of "high" defect probabilities (high with respect the overall proportion of defective chips.)

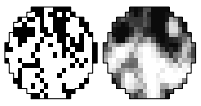


Figure 4. *A binary wafer and its smoothed version.*

In this example, we will use data collected in a designed experiment to relate the patterns we observe in these smooth maps to a set of manufacturing conditions. Traditionally, engineers have summarized each wafer by its yield (i.e., the proportion of good or non-defective chips) and performed data analysis on the yield alone. This approach, like the traditional approach seen in the fiber example, ignores the context (the spatial structure) of the response.

A designed experiment was conducted to study the ef-

fect of two factors on yield; the factors were a diffusion time and a coating step. Three diffusion times were varied, from short, to medium, and long; two coatings were used, thin and thick. In Figure 5 we show the interaction between the two factors by computing the average yield at each factor combination. We augment this well-known display with averaged wafers where the value at the i-th site represents the average defective proportion across wafers for the given factor combination.

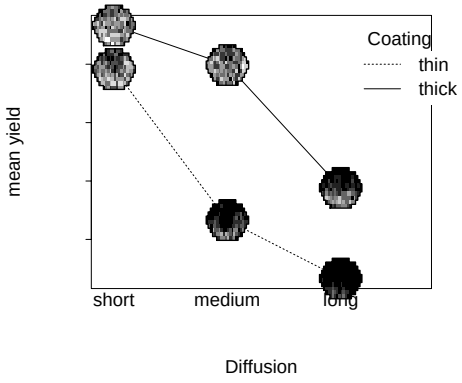


Figure 5. *Interaction plot for the wafer experiment.*

Figure 5 suffices to determine the factor combination that maximizes yield (thick coating and short diffusion times). However, the engineers were also interested in untangling the effects of the factors on the wafer surface, thus an extra step was needed. By following the paradigm shown in the fiber example, we fitted an analysis of variance model at each site on the wafer. We collected the coefficients of coating and the linear term of diffusion from all ANOVA's and displayed them as wafer objects in Figure 6.

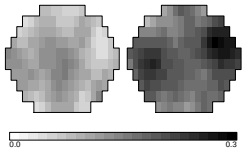


Figure 6. *Coefficients for the coating term and for the linear term of diffusion time.*

We then repeated this process for all parameters of the model and displayed them as wafer objects in a familiar regression analysis summary table in Figure 7. The first column in this figure shows how the coefficients vary across the wafer surface; the second column displays the standard error for each coefficient and site; similarly the third column shows T-values for each site and coefficient; the last column shows p-values — light sites (close to zero) denote chips where the corresponding coefficient is significantly different from zero. The two wafers at the bottom show the estimated standard deviation (labeled `Root MSE`) and the percentage of variability that the model explains (labeled `R-squared`): we note that the model (i.e., changes in diffusion and coating) can explain a large fraction of the variability in yield. The intercept wafer indicates that yield is smallest at the top of the wafers across all experimental conditions.

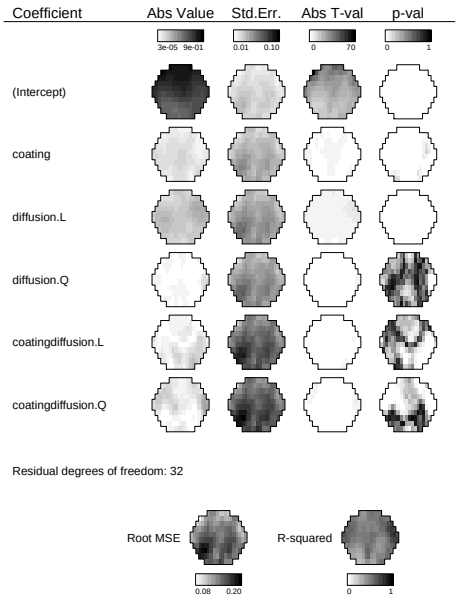


Figure 7. *Coefficients from the ANOVA fit*

The coefficient for coating (better seen in Figure 6) shows smaller yields in the center of the wafer, while diffusion time has large coefficients outside the center. Thus by carrying out the analysis of variance at each site we have been able to separate the effects of coating from those of diffusion time on the wafers.

Predictions and residuals may also be plotted as wafers for model diagnostics.

3. Discussion

As the examples illustrate, by incorporating intrinsic longitudinal/spatial structure into known graphical methods we begin to understand the variation in our data along those dimensions. Notice that in both examples we employed the analysis of variance decomposition as a means for exploratory data analysis rather than an inferential procedure, and then displayed the various components along the length of the fiber and over the surface of a wafer. It was through graphical displays that we assessed the extent of the longitudinal and spatial dependence of the responses. Alternative approaches that could be used include generalized additive models with varying coefficients, (Hastie and Tibshirani 1990; Ha-

site and Tibshirani 1991), and loess models (Cleveland 1979; Cleveland et al. 1993) , among others. In the case of wafers, formal spatial analysis techniques can be used to estimate the extent of spatial clustering and its relation to the covariates (for instance, Taam and Hamada, 1992), but the above graphical displays are more visually effective.

In both examples the models used did not include terms for longitudinal and spatial effects because it was strongly felt that there was no prior knowledge of the interaction between the longitudinal/spatial structure and the design factors. By allowing the effects to freely vary over the fiber and over the wafer surface we guarded against most types of misspecification; yet the graphics we used to display the effects and coefficients effectively reveal their longitudinal and spatial dependence.

If the data are not collected through designed experiments, techniques such as principal components or hierarchical clustering may be appropriate. These are some of the multivariate techniques whose graphical displays can easily be augmented with similar symbols or glyphs, e.g., imagine a cluster dendrogram with wafers at the leaves.

4. Acknowledgments

The fiber study was joint work with Daryl Pregibon, who introduced me to the idea of "pasting" ANOVA models; the wafer case study was joint work with Daryl and Mark H. Hansen – I want to thank both very much.

5. References

Becker, R. A., Clark, L. A., and Lambert, D. (1994). Cave plots: A graphical technique for comparing time series. *Journal of Computational and Graphical Statistics*, 3(3):277–284.

Becker, R. A., Eick, S. G., and Wilks, A. R. (1991). Basics of network visualization. *IEEE Computer Graphics and Applications*, 11(3):12–14.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

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

Cleveland, W. S., Mallows, C. L., and McRae, J. E. (1993). ATS methods: Nonparametric regression for non-gaussian data. *Journal of the American Statistical Association*, 88(423):821–835.

Cliff, A. D. and Ord, J. K. (1981). *Spatial Processes: Models and Applications*. Pion Limited, London, UK.

Goldman, A. I. (1992). Eventcharts: Visualizing survival and other timed-events data. *The American Statistician*, 46(1):13–18.

Hastie, T. and Tibshirani, R. (1991). Varying-coefficient models. Technical memorandum, AT&T Bell Laboratories, Murray Hill, NJ.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Taam, W. and Hamada, M. (1992). Detecting spatial effects from factorial experiments: An application from integrated-circuit manufacturing. *Technometrics*, 35:149–160.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut.

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

David A. James
Bell Laboratories
dj@bell-labs.com

TOPICS IN INFORMATION VISUALIZATION

Simplifying Visual Appearance by Sorting: An Example using 159 AVHRR Classes

By Daniel B. Carr and Anthony R. Olsen

1. The Visual Intimidation Factor

A presumed goal of tables and plots is to communicate to a target audience. Unfortunately, many tables and some plots appear visually intimidating, so fail as a communication device. In the spirit of Tufte (1983), who introduced concepts such as the lie factor and the data ink to total ink ratio, we define a concept called the visual intimidation factor (VIF). The VIF (rhymes with whiff) is the reciprocal of the time (measured in seconds) it takes to decide that the study of a table (plot) is not worth the effort. If the reader studies the table and derives useful information, the time is infinite and the VIF=0. One can't decide faster than a preattentive vision sweep of the table (about 50 milliseconds) so a theoretical upper bound is 1/.050=20. More realistically it may take a whole second to make a decision, so a VIF of 1 is representative of a bad table. Driving down the

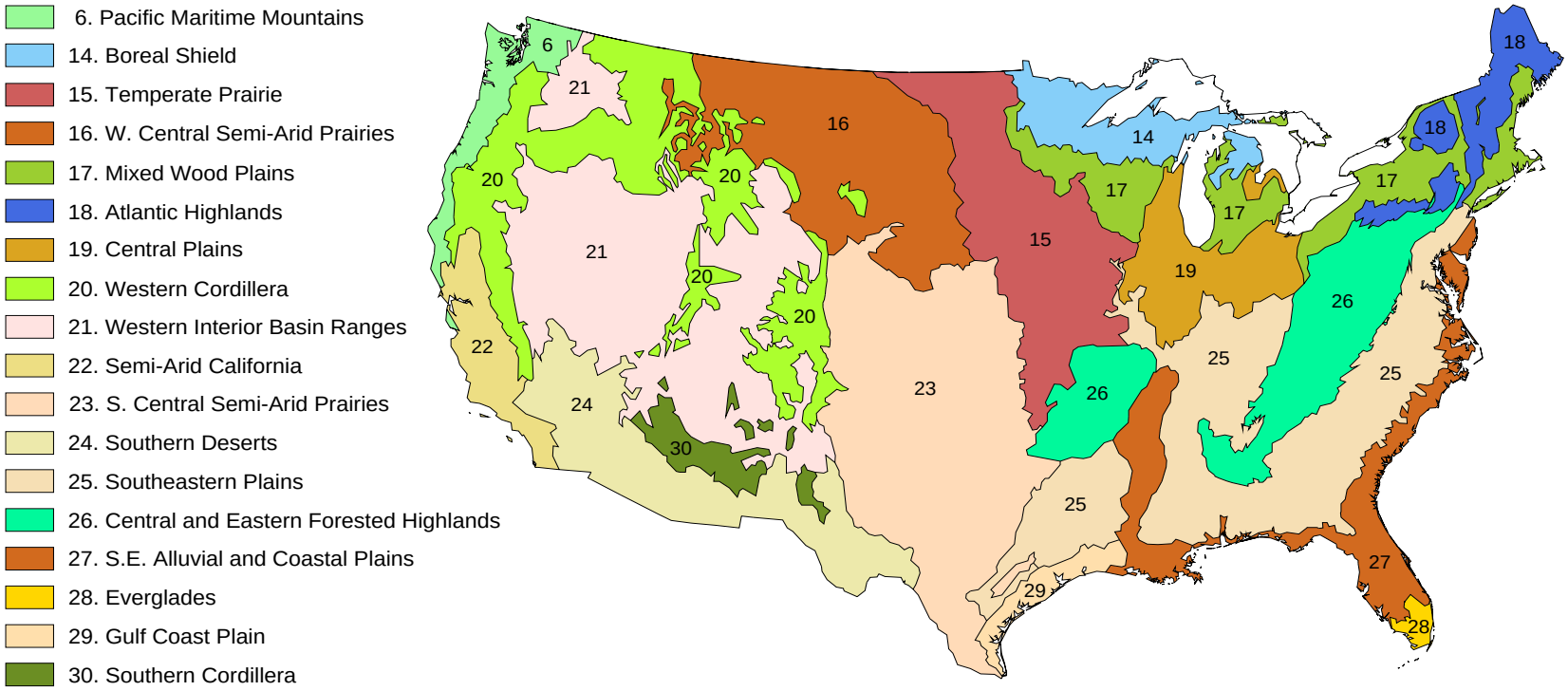


Figure 1. Level II Ecoregions of the Conterminous United States. (Omernik 1995).

VIF for complex tables and plots can be a challenge. Assuming an interested audience, Kosslyn's (1994) adage, "the spirit is willing but the mind is weak," is appropriate. This article focuses on a powerful tool for reducing the VIF, multivariate sorting.

2. Multivariate Sorting

Many statistical graphics writers are proponents of sorting. Cleveland (1985) describes research demonstrating that comparison accuracy increases with the nearness of the comparison items. Sorting brings similar items close together and they become easier to compare. Cleveland (1985, 1993) makes extensive use of sorting to bring out patterns in dot plots. Becker and Cleveland (1994) illustrate the advantage of sorting box plots by medians and Wainer (1993) discusses the advantage of sorting in tables.

While increasing the perceptual accuracy of extraction may have provided Cleveland's explicit motivation to sort data, an amazing consequence is that plots look much simpler. Carr (1994) describes this in terms of shortening the eye traversal path and reducing the number of visual focal points. Sorting boxplots by the median reduces the eye traversal path in moving from median to median and increases the apparent simplicity of the plot. Creating localized blocks in two-way layouts reduces the number of visual focal points and increases the apparent simplicity. Since we process visual information simultaneously on different scales (Marr 1982), our eyes can be drawn to many different places in a plot. An amazing plot reprinted in Marr (1982, page 50) contains patterns at different scales that emerge and disappear as one gazes at the plot. We conjecture that sorting often reduces the number of comparison scales, that limiting the number of comparison scales helps us focus at the same places in repeated viewing, and that stability in repeated viewing is a key to apparent simplicity. Whether or not our conjecture is correct, sorting simplifies.

3. Examples Using Ecoregions and AVHRR Classes

We put sorting to work to simplify the appearance of a two way layout. The challenge comes from the USEPA Western Ecology Division. The levels of the first factor are ecoregions for the conterminous U.S. Omernik (1995) constructs maps that partition North America into ecoregions, on the premise that ecological regions can be identified through the analysis of the patterns and composition of biotic and abiotic phenomena. The partitions integrate extensive knowledge of geology, physiography, soils, vegetation, climate, land use, wildlife, and hydrology. Although ecoregions are available at several scales, our interest is in Level II, which has 18 ecoregions within the conterminous U.S. (Figure 1). We want to communicate the commonalities and differences in biotic and abiotic characteristics across ecoregions.

We focus on a land cover characterization to illustrate how multivariate sorting can reduce the VIF for a large two-way layout. Loveland et al (1991) derives a land cover classification relying mainly on satellite imagery. Using AVHRR imagery spectral intensities for 1 km square pixels and additional information, they assign approximately 8 million pixels into one of 159 land classes. This classification is pixel resolution dependent and does not necessarily reflect the diversity within a pixel. For example, few pixels are classified as water, since few bodies of water dominate a full pixel. At other resolutions, acreage associated with the land classes would differ. The levels of the second factor in the layout are these 159 AVHRR classes. The dependent variable is acreage.

Figure 2 is a line height (thin bar) plot showing the class acreage as a percent of each ecoregion total acreage. For compactness, the plot omits the labels for the 159 classes. To provide labels one could make a larger plot or in an interactive setting handle the labeling by brushing, progressive disclosure or selective magnification. We conclude that Figure 2 has a high VIF. No spatial pattern appears sufficiently interesting to induce further examination. The plot communicates a message of spikes located "randomly" throughout the plot. The order for the row factor levels reflects the general north to south ecoregion numbering pattern. The order for the column factor levels reflects a hierarchical land cover classification scheme that is partially described below.

Figure 3 is a first cut to simplify the appearance of the plot via bi-directional multivariate sorting. While we have not dealt with the column labels and interpretation, the patterns now seem simple enough that maybe we can understand some of the relationships without too much work. In other words there may be a few concepts that characterized the acreage for the ecoregions. With some luck the concepts will mesh well with the existing top levels of the existing hierarchical classification.

Figure 3 illustrates just one of several viable approaches to multivariate sorting are available. Consider sorting rows. One approach is to obtain the median across all the AVHRR classes for each ecoregion and then to sort rows using the median. Cleveland (1993a) has found several examples in which collapsing to one dimension is effective. In environmental applications, the approach often fails because the median is often zero.

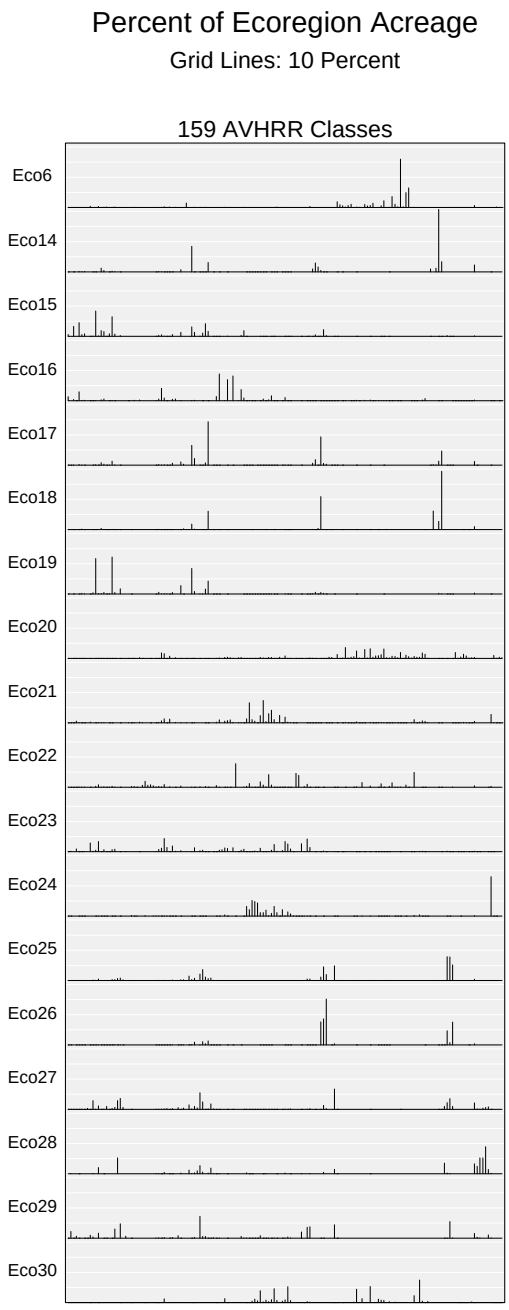


Figure 2:
Original Row and Column Order

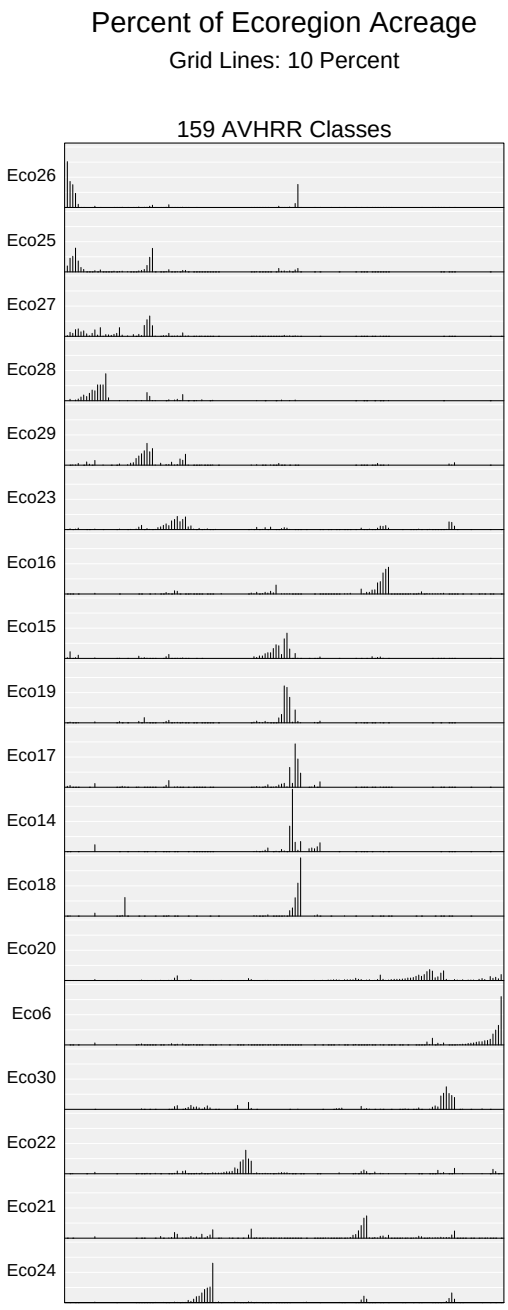


Figure 3:
Sorting of Rows and Columns

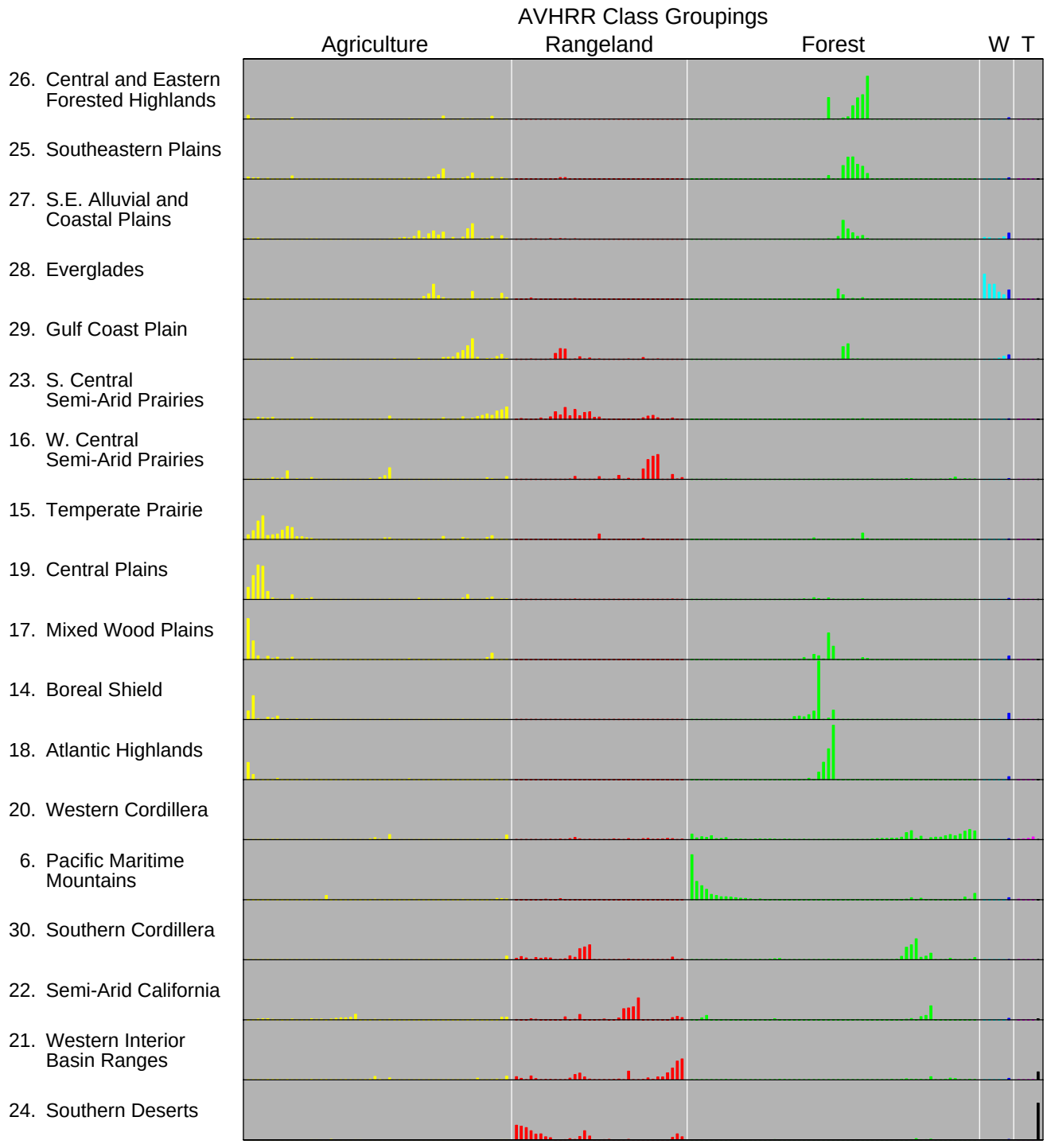


Figure 4: Sorting Classes Within Groupings
W: Wetlands and Water (Cyan and Blue)
T: Tundra and Barren (Magenta and Black)

The goal is to create a smaller number of perceptual groups or focal points. Figure 3 illustrates our use of Friedman and Rafsky's (1979) minimal-spanning-tree breadth-traversal algorithm to sort rows and columns. To sort rows, the algorithm starts by building a minimal spanning tree in 159 space. Then the algorithm establishes two nodes on the tree that have the largest traversal path in going from node to node. The breadth traversal algorithm starts at one of these two nodes and visits nearest subtrees until it eventually arrives at other node. The visiting of subtrees tends to provide effective visual groupings.

Two other approaches for separate row and column sorting are worth mentioning. One simple procedure is to sort rows (columns) by the first principal component scores. Another is to invoke a clustering algorithm such as a single-link algorithm, (see Banfield and Raferty 1992 for some modern clustering options), and to borrow the ordering from the ensuing dendrogram.

Separate row and column sorting is applicable to many crossed two-way layouts. Logical constraints may prohibit sorting both rows and columns, but any time a color matrix appears (from genetic algorithm population descriptions to protein descriptions) one should think about sorting. We don't know of research establishing a perceptually best sorting method. With today's computational power one can optimize over all permutations of rows and over all permutation of columns and iterate if necessary. It is easy to propose various clustering indices for optimization. Anything that puts low values together and high values together will likely help.

Another facet of making row (and column) labeled plots look simpler is to break the labels into groups. Kosslyn (1994) suggests that groups of size four or fewer are best for making within group comparisons. A list of length 12 is more visually intimidating than three groups of size four. In addition, creating smaller groups provides edges. The edges draw visual attention and when readers happen to notice a label of interest at an edge, they begin to get involved. This audience is probably not attuned to "home" ecoregions as it would be to a home state or county. Thus, the comment is not so important for this particular example. We do note in passing that more can be done with the rows in Figure 3. If such were available, a classification based on the labels provides one way of clustering rows. Another approach is to use a clustering algorithm as suggested above. For example, one can make clusters by cutting at the long links in the spanning tree traversal.

Several options provide a graphical representation of the clusters. The natural choice is to add space between clusters. When space is at a premium, we might try indenting every other cluster or alternating two background colors behind the labels. In an interactive setting, a mousing operation might reveal the whole dendrogram. Explicitly showing clusters is not always advantageous. A one-dimensional layout is not conducive to preserving among cluster distances, and the clusters themselves can be somewhat arbitrary. When the explicit clusters are not well supported by the rest of the graphic, the VIF may increase.

In this example, the column labels come with a hierarchical classification. There are seven classes at the lowest resolution, twenty-five at the second resolution, and 159 classes at the highest resolution. The seven classes are agriculture, rangeland, forest, water, wetland, barren and tundra. Figure 4 shows five groups, lumping water with wetland and barren with tundra. Colors indicate all seven groups, using yellow, red, green, cyan, blue, magenta and black, respectively. With grouping by position, the use of different hues in not required for this plot. (Gray and black bars can distinguish the combined groups.) However, the different hues help to strengthen the perceptual grouping and tend to add visual appeal, thus reducing the VIF. Representing as many as seven classes with different hues presses the limits for reducing the VIF. In this case the redundant encoding by position and the almost too-short-to-see magenta lines lessen the interpretation demands. Note that we have sorted the columns within each of the classes. One could also sort the placement of the five classes by the class totals.

Some patterns emerge in Figure 4. The Everglades have wetlands. The Pacific Maritime Mountains have a distinctive pattern of forest types. The Western Cordillera has great diversity. The Southern Deserts are substantially barren. This is partly consistent with what the reader already knows, but perhaps adds some new information. That's a good starting point. The omitted class labeling offers to provide additional information.

4. Variations and Extensions

We have looked at the next higher resolution classification involving 25 groups. One plot variation aggregated acreage into the 25 groups. With only 25 groups, labeling was easy. To save horizontal space and facilitate reading we started the labels over the columns and rotated them counterclockwise 40 degrees from horizontal. With the columns a bit wider than character height thickness, the thin bars became thick bars or else were widely separated. The visual effect was disappointing and the aggregation story was of questionable interest. While including column labels is an important key to

deeper interpretation, we have relegated the plot to electronic access.

Sorting, grouping and labeling are powerful tools. The primary drawback to sorting rows is that people often look up values by their labels (Cleveland 1993b). When labels are not in an alphabetic or another familiar order, then the look-up task becomes complicated. Linking items to a new ordering or back to a map helps people to put the information together. Different hues are only an effective link for a few items. For many items, other approaches are better. Beyond providing written grid coordinates, visual linking methods include marked microplots of 90 degree rotated row-labels, and marked postage stamp maps (see Eddy and Mockus 1995 for a discussion of stamp-sized images). However, linking is a topic for another paper.

5. Access and Comments

Data, Splus functions and script files producing the current examples are available via anonymous ftp to `galaxy.gmu.edu`. Change directory to `submissions/newsletter/sorting`. Examples from other newsletter articles are now stored under this newsletter directory. As always I (Dan) welcome constructive comments plus new graphics challenges.

Acknowledgments and Disclaimer:

This work is part of collaborative research with the USEPA Western Ecology Division and involves Sue Pierson and Pip Courbois. We thank them for their contributions of data and ideas. We invite the readers to see our collective results involving digital terrain data and other variables at the ASA annual meeting in Chicago.

Although the research described in this article has been funded in part by the U.S. Environmental Protection Agency through cooperative agreement CR820820 to George Mason University, it has not been subjected to Agency review. Therefore, it does not necessarily reflect the views of the Agency.

References

Banfield, R. D. and Raferty, A. E. (1992). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803-822.

Becker, R. A. and Cleveland, W. S. (1993). Discussion of Graphical Comparisons of Several Linked Aspects by John W. Tukey. *Journal of Computational and Graphical Statistics*, 2(1):41-48.

Carr, D. B. (1994). Converting Tables to Plots. Technical Report 101, Center of Computational Statistics, George Mason University, Fairfax VA, 22030.

Cleveland, W. S. (1985). *The Elements of Graphing Data.* Wadsworth Advanced Books & Software, Monterey, CA.

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

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

Eddy, W. F. and A. Mockus (1996). An Interactive Icon Index: Images of the Outer Planets. *Journal of Computational and Graphical Statistics,* 5(1):100-111.

Friedman, J. H. and L. C. Rafsky (1979). Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-Sample Tests. *The Annual of Statistics,* 7(4):697-717.

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

Loveland, T. R., Merchant, J. W., Reed, B. C., Brown, J. F., Ohlen, D. O., Olson, P., and Hutchinson, J. (1995). Seasonal land cover regions of the United States. *Ann. Assoc. Amer. Geog.*, 85:339-355.

Marr, D. (1982). *Vision.* W. H. Freeman and Company, New York, NY.

Omernik, J. M. (1995). Ecoregions: A framework for managing ecosystems. *The George Wright Forum,* 12:35-51.

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

Wainer, H. (1993). Tabular Presentation. *Chance*, 6:52-56.

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

Tony Olsen
US EPA
National Health and Environmental Effects Research Laboratory
`tolsen@mail.cor.epa.gov`

DASL—Dazzle—The Data and Story Library

by Mike Meyer

Two years ago, in this column, I wrote,

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

The Data and Story Library (DASL—pronounced "dazzle") has finally been released for public use and is available at `http://www.stat.cmu.edu/DASL/`. The public version of DASL has about 80 stories and associated data files that cover a broad range of simple statistical ideas, from t-tests to regression. Each story describes something about the background of the data and, where appropriate, highlights some feature of the data. The stories often contain a graph of some interesting feature of the data. I am convinced that DASL will be a great help to harried professors who are searching for good examples to illustrate statistical concepts. However, the only way I can convince **you**, is to invite you to try it for yourself. We will know if you like it, because you will either come back and use it again or you will contribute your own datasets. If you do not like it, you will have to tell us, and we look forward to your comments—either in e-mail or as letters to the editors of this forum.

DASL Development

The bulk of the development work for DASL was done by Paul Velleman and a small team of students at Cornell University. *Some* of the funding for DASL was provided by a grant from NSF which funded DASL and a related project at Ohio State University, called the Electronic Encyclopedia of Statistical Exercises and Examples (EESEE). Further details about EESEE are described in the following article.

The rest of the DASL funding has come from the extra effort that all the developers have donated to the project. For some of us, getting DASL "right" has become somewhat of a passion. The early versions of DASL were a purely Macintosh application. My own contribution was the relatively obvious suggestion that DASL be housed on the WWW. That led directly to StatLib being the home of DASL, and indirectly forced me to implement a WWW search mechanism for DASL and the rest of StatLib.

We expect to continue to update the archive's content and interface over the next few years but this depends on you.

The Future of DASL

There is a small amount of continuing NSF funding to support updating and improving DASL. However, the real future of DASL depends on contributions of stories and data from the users. Developing a DASL entry is not difficult, but it is a little more complicated than merely taking your dusty old data decks and giving them to someone. It is much more like preparing a short paper or handout for your employer/clients/students. The handout should contain some (maybe a lot) of scientific background to the data, and a short discussion of some of the interesting data features.

We strongly encourage submissions of stories and data to the archive. To make this relatively easy we have even included a WWW form that will help you format your submissions and then e-mail them in. We hope that the form is useful, but we also understand that some types of story/data combination may require different formats. If you have some data that you think is appropriate for DASL, and the fill-in form does not meet your needs, then please send e-mail to me. We would be delighted if some vendors of statistical software contributed their stories and datasets.

Technical Details

Most of the technical details of DASL are quite obvious. Paul Velleman and a talented student designed most of the user interface. By aggressively tracking improvements in the HTML language, we have been able to end up with an interface which seems quite easy to use and does not overwhelm the user with options and graphics. Early on we decided that we would try to avoid statically generated index entries and rely on dynamic searches. We have stuck with that decision and many of the links on DASL are really requests to the StatLib search en-

gine. I hope that I have coded the searches generally enough that it will be easy to change search engines from the current Isearch engine developed by CNIDR (`http://www.cnidr.org`) to whatever emerges as the best WWW/fulltext search engine.

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

...MORE SNOOPING

EESEE—The Electronic Encyclopedia of Statistical Examples and Exercises

by Dennis Pearl

Many of the stories in DASL are an abbreviated version of the form contained in the Electronic Encyclopedia of Statistical Examples and Exercises (EESEE), a desktop computer based resource of statistical stories in an integrated multimedia environment. EESEE is being built as a fully cross-indexed source of materials for teachers of statistics, as the backbone for independent self-paced learning by students, and as a tool for computer laboratories in a data analysis or statistical concepts course. It consists of an integrated set of referenced stories (including an outline of the original study designs) with multiple levels of comments, questions, solutions, project ideas, and accompanying data sets, pictures, maps, and videos. It currently operates in a user-friendly HyperCard interface.

The stories in EESEE represent a diverse and lively group of issues addressed in the scientific literature and the popular press (to date there are about 75 stories incorporating over 1500 pages of material). A sample of the issues addressed: Do trilobite fossils show these Paleozoic creatures exhibited a trait similar to right-handedness? Does the architectural style of a building make people perceive the building as safe or dangerous? Is there a home field advantage in the World Series? Do emergency room nurses heed warnings about wearing gloves? Is personality an inherited trait? How fast does your blood alcohol level go up when you drink a beer or a glass of wine? Does training in art appreciation really help you interpret paintings like the artist intended? Does the smell of lilac help you think more clearly? Do the spotted owls really have to live in the old-growth forests? Is it possible to make a hamburger without the fat? Do laws about handgun ownership change the rate of burglaries?

Using EESEE, students can take a "guided tour" of these issues by exploring the accompanying materials which map out a statistical critique of the story. Or they can take off on their own - via interactive access to the raw data. If you would like more information about EESEE, you can view selected sample screens at `http://stat.mps.ohio-state.edu/projects/eesee/index.html` or send e-mail questions to `eesee@stat.mps.ohio-state.edu`.

Dennis K. Pearl
The Ohio State University
`dkp@stat.mps.ohio-state.edu`

GRAPHICS AND COMPUTING JSM PLANS

Joint Statistical Meetings

by James Rosenberger and Stephen Eick

The invited program for the Statistical Computing and the Statistical Graphics Sections of the ASA are now in place for the 1996 meetings in Chicago. There is still time to register for the meetings however, and anyone interested should check the AmStat News for the registration form. The programs organized by Rob Tibshirani, Statistical Computing Program Chair, and Stephen Eick, Statistical Graphics Program Chair, look quite exciting.

Statistical Computing Section

The invited sessions planned include a session on "Bayesian inference for High-Dimensional Problems: Models and Computation" organized by Radford Neal. This will feature talks on modern statistical and computation challenges in the statistical and artificial intelligence area. The talks planned are "Computational Methods for Graphical Models" by Jordan, "Regression with Gaussian Processes" by Rasmussen, and a final talk by Wolpert followed by discussion by Blackmond Laskey.

A session on "Statistics and Numerical Integration" is organized by Art Owen, and will cover recent advances in high-dimensional integration. Talks include one by

Owen on this topic, "Number Theoretic Methods in Statistical Inference" by Fang, and "Lattices and Dual Lattices in Optimal Design For Fourier Regression" by Henry Wynn.

"Wavelets and Time-Frequency Analysis" is the theme of a session organized by Jon Buckheit which will feature talks on applications of wavelet methods to statistical problems. Nason et al. will talk on "Wavelet Methods in Statistics", Saito, et al. will talk on "Multipass Classifications by Splitting Data to Good and Bad Sets with LDB" and Buckheit, et al. will present "Best Cumulant Bases for Non-Gaussian Stochastic Processes."

"Resampling and Cross-Validation in Model Building" is organized by Martin Shumacher, and will feature talks on the bootstrap and cross-validation. Talks include "Cross-Validation in Survival Analysis" by Houwelingen, "The Bootstrap and Modulation Estimators" by Beran, and "Reduction of Bias Caused by Model Building" by Schumacher.

"Identifiability and Convergence Issues in MCMC Implementation" is organized by Brad Carlin and will survey some recent results in Monte Carlo Markov Chain simulation methods. Talks in this session include "Orthogeneralizations and Predistributions for Othogeneralized Model Mixing" by Clyde, "Identifiability and Convergence for High-Dimensional MCMC Algorithms" by Gelfand and a final talk by Gidas.

The last session for the section is "Algebraic Algorithms", organized by James Stafford and will feature recent work on symbolic computation for statistical applications. Talks are "Symbolic Ito Calculus: A New Application to the Statistical Theory of Shape" by Kendall, "Stochastic Differential Equations" by Cyganowski, et al., "A Computer Algebra For Sample Survey Methodology" by Stafford, et al. , and "Symbolic Bootstrap and Saddlepoint Correlations in the Investigation of Properties of Bootstrap Estimates" by Andrews.

Statistical Graphics Section

The program for the Statistical Graphics section includes the following invited sessions focusing on new research results in statistical graphics.

On Monday at 8:30am, a session on "Innovations in Graphics" is organized by Stephen G. Eick and includes talks on "Table Lens as a Tool for Making Sense of Data" by Rao, "Visualizing the Distribution of Hierarchical and Categorical Data With Tremaps" by Johnson, and "SDM: Malleable Information Graphics" by Chuah, et al. Interactive graphical displays showing statistical data are computer interfaces specifically tuned for visualization. In the Human-Computer Interface community there has been an extensive focus on designing better interfaces in general and particular interfaces for looking at data. Much of this research occurs in corporate labs. These talks demonstrate three novel techniques and state-of-the-art interfaces for visualizing hierarchical, tabular, and spatial data.

On Tuesday at 2:00pm a session on "Information Visualization: The Next Wave In Statistical Graphics?" is also organized by Stephen Eick and includes talks on "Information Visualization & Exploration Environment" by Ahlberg, "Statistical Bases of Text Visualizations" by York et al. and a discussion by Carr and Mockus. This session introduces the statistical graphics community to a novel technique for visualizing relationships among text documents and software system for visualizing statistical databases.

On Wednesday at 8:30am a session on "Statistical Graphics and Multimedia Education", organized by David Scott, includes talks on "A Multimedia Statistics Course: A Network of Statistical Concepts" by Rosenberger and Heckard, "Stochastic Visualization and the Internet: Research and Instructional Uses" by Rossini et al., and "Interactive Instruction on the Web" by Narasimhan, followed by discussion by David Scott and Sandy Weisberg.

At 10:30am on Wednesday a session on "Transactional Data Analysis" is organized by Daryl Pregibon and includes talks on "Visualizing Credit Card Transactions" by Yuhas, "Opportunities for Visualization in Retail Sales Data" by Wills, and "Visualizing Telephone Network Data" by McIntosh et al. A key and fundamental problem facing corporate researchers involves coping with massive datasets. Current graphics techniques are overwhelmed with small datasets that are easily stored on inexpensive personal computers. Developing scalable graphical techniques is a promising and very important new research area with significant commercial impact and potential.

Call for Suggestions

Between now and the 1996 August meeting the Program Chair-Elect of the respective sections would welcome ideas for organizers and session themes in Statistical Computing and Graphics. Please email your ideas for Statistical Computing sessions to James Rosenberger, and for Statistical Graphics to Dianne Cook, at the addresses given at the back of the Newsletter. Input is

always welcome from the membership, and our sections are key to a wide range of statistical applications and methodology advances.

James L. Rosenberger
Statistical Computing
Program Chair-Elect
JLR@stat.psu.edu

Stephen G. Eick
Statistical Graphics Program Chair
eick@bell-labs.com

CONFERENCE NOTICES

Interface '96

28th Symposium on the Interface: Computing Science and Statistics

Graph – Image – Vision

July 8–12, 1996
Sydney, Australia
http://www.dms.csiro.au/sisc/index.html

Sponsor: Interface Foundation of North America

Program co-chairs: N. Fisher and L. Billard

Cooperating organizations

American Statistical Association (ASA); Institute of Mathematical Statistics (IMS); International Association for Statistical Computing (IASC); Society for Industrial and Applied Mathematics (SIAM); Operations Research Society of America (ORSA).

Symposium Information

Interface '96, the 28^{th} Symposium on the Interface, will be held in Sydney, Australia in conjunction with the Sydney International Statistical Congress, SISC '96. The Congress will be held at The Wentworth Hotel, a venue with outstanding facilities and also very well-located in the centre of Sydney. SISC '96 combines Interface '96 with the 13^{th} Australian Statistical Conference and an IMS Special Topics Meeting on "Contemporary Nonparametrics."

The Program

The Interface meeting is being developed around the theme of Graph–Image–Vision. Topics for sessions include

- Surface reconstruction
- Software quality and metrics
- New directions in programming environments for data analysis
- High-dimensional data
- Saddlepoints and related methods
- Graphics and visualization; Multimedia
- Model selection and time series
- Influence and sensitivity in multivariate analysis
- Hidden Markov models

In addition, there will be a one-day workshop on imaging. Confirmed participants in SISC '96 include Vic Barnett, Andrew Barron, Rick Becker, Lynne Billard, Peter Bickel, Mary Ellen Bock, David Brillinger, John Chambers, Bill Cleveland, Di Cook, Larry Cox, Persi Diaconis, Bill Eddy, Brad Efron, Steve Eick, Ed George, Herb Gish, Peter Hall, David Hinkley, I. Ibragimov, Ross Jeffrey, Iain Johnstone, Genshiro Kitagawa, John Kolassa, Don Kimber, Richard Olshen, Walter Piegorsch, David Scott, Bernard Silverman, Adrian Smith, Yutaka Tanaka, Andreas Weigend, Ed Wegman, Mike West, Lee Wilkinson, and Alastair Young.

Conference Fees and Travel Arrangements

If you have not yet registered for the conference, the general fee is $A375 for members of the participating societies, $A395 for non-members, $A150 for students and $A145 for one-day registrants. ($A1 is approximately $US 0.73.)

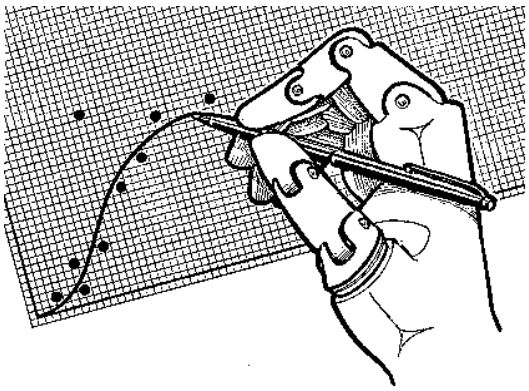
QANTAS is the official airline for the 1996 Sydney Conference. Due to their support of the Conference, the SISC '96 organizers urge delegates to consider QANTAS for travel to and from (and within) Australia.

For More Information

The website listed at the beginning of this article contains the full program for SISC '96, conference fees and accommodation information. You will also find information about day tours and the social program. Details on the Interface meeting as well as SISC '96 can also be obtained from the following sources.

E-mail:	sydney96@syd.dms.csiro.au
Telephone:	+61 2 325-3239
Fax:	+61 2 325-3243
Mail:	The Director, SCISC–96 CSIRO Division of Mathematics and Statistics Locked Bag 17 North Ryde, NSW 2113 AUSTRALIA

CALL FOR PAPERS



Sixth International Workshop on Artificial Intelligence and Statistics

January 4-7, 1997
Ft. Lauderdale, Florida
`http://www.stat.washington.edu/aistats97/`

Program Chair: David Madigan, U of Washington

General Chair: Padhraic Smyth, JPL and UCI

This is the sixth in a series of workshops which has brought together researchers in Artificial Intelligence (AI) and in Statistics to discuss problems of mutual interest. The exchange has broadened research in both fields and has strongly encouraged interdisciplinary work.

To encourage interaction and a broad exchange of ideas, the presentations will be limited to about 20 discussion papers in single session meetings over three days (Jan. 5-7). Focused poster sessions will provide the means for presenting and discussing the remaining research papers. Papers for poster sessions will be treated equally with papers for presentation in publications. Attendance at the workshop will not be limited.

The three days of research presentations will be preceded by a day of tutorials (Jan. 4). These are intended to expose researchers in each field to the methodology used in the other field. The tutorial speakers are A. P. Dawid (University College London), Michael Jordan (MIT), Tom Mitchell (Carnegie Mellon), and Mike West (Duke University).

Topics of Interest

> Automated data analysis and knowledge representation for statistics; statistical strategy; metadata and the design of statistical data bases; causality; multivariate graphical models, belief networks; cluster analysis and unsupervised learning; predictive modeling, classification and regression; interpretability in modeling; model uncertainty, multiple models; probability and search; knowledge discovery in databases; integrated man-machine modeling methods.
>
> Statistical methods used in AI approaches to vision, robotics, pattern recognition, planning, information retrieval, natural language processing, etc., and AI methods applied to problems in statistics such as statistical advisory systems, experimental design, exploratory data analysis, causal modeling.

This list is not intended to define an exclusive list of topics of interest. Authors are encouraged to submit papers on any topic which falls within the intersection of AI and Statistics.

Submission Requirements

Three copies of an extended abstract (up to 4 pages) should be sent to

David Madigan, Program Chair
6th International Workshop on AI and Statistics
Department of Statistics, Box 354322
University of Washington, Seattle, WA 98195

or electronically (either postscript or a WWW address) to `aistats@stat.washington.edu`

Submissions will be considered if postmarked by June 30, 1996.

If the submission is electronic (e-mail), then it must be received by midnight July 1,1996. Please indicate which topic(s) your abstract addresses and include an electronic mail address for correspondence. Receipt of all submissions will be confirmed via electronic mail. Acceptance notices will be mailed by September 1, 1996. Preliminary papers (up to 20 pages) must be returned by November 1, 1996. These preliminary papers will be copied and distributed at the workshop.

Program Committee

Russell Almond, ETS, Princeton; Wray Buntine, Thinkbank, Inc.; Peter Cheeseman, NASA Ames; Paul Cohen, University of Massachusetts; Greg Cooper, University of Pittsburgh; William DuMouchel, Columbia University; Doug Fisher, Vanderbilt University; Dan

Geiger, Techinion; Clark Glymour, Carnegie-Mellon University; David Hand, Open University; Steve Hanks, University of Washington; Trevor Hastie, Stanford University; David Haussler, UC Santa Cruz; David Heckerman, Microsoft; Paula Hietala, University of Tampere; Geoff Hinton, University of Toronto; Michael Jordan, MIT; Hans Lenz, Free University of Berlin; David Lewis, AT&T Labs; Andrew Moore, Carnegie-Mellon University; Radford Neal, University of Toronto; Jonathan Oliver, Monash University; Steve Omohundro, NEC Research, Princeton; Judea Pearl, UCLA; Daryl Pregibon, AT&T Labs; Ross Shachter, Stanford University; Glenn Shafer, Rutgers University; Prakash Shenoy, University of Kansas; David Spiegelhalter, MRC, Cambridge; Peter Spirtes, Carnegie-Mellon University.

For more information see the workshop's Web page given at the top of this notice or write David Madigan at `aistats@stat.washington.edu` for inquiries concerning the technical program or Padhraic Smyth at `aistats@aig.jpl.nasa.gov` for other inquiries about the workshop.

Announcing the 1997 Student Paper Competition

The Statistical Computing Section of the ASA is again sponsoring a Student Paper Session at the Joint Statistical Meetings in 1997. Our 1995 and 1996 competitions were successful, and attracted 16 and 18 entries respectively. The winners present their papers in a special session at the annual ASA meetings. See the article on page 4 concerning the winners of this year's competition.

The topic of the session is *Statistical Computing*. Four students will be selected to participate in this session. All fees associated with registration, accommodation, and travel to the conference will be awarded to the participants in this Session.

Students at all levels (undergraduate, Masters, and Ph.D.) are encouraged to participate. To be eligible, an applicant must be a registered student in the fall of 1996. The applicant must be the first author of the paper.

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

All application materials MUST BE RECEIVED by January 10, 1997. They will be reviewed by the Statistical Computing Section Student Paper Competition Award committee. The topic of the paper should be in the area of statistical computing, and might be original methodological research, some novel application, or any other suitable contribution (for example, a software related project). Selection will be based on a variety of criteria at the discretion of the selection committee, and will include novelty and significance of contribution, amongst others. Award announcements will be made in late January, 1997. The selection committee's decision will be final.

Students not selected for inclusion in the Session may submit their abstract and a registration fee to ASA by February 1, 1997 if they plan to attend the Joint Meetings. Those abstracts must be submitted according to the ASA abstract submission instructions described in AMSTAT News. Students selected for inclusion in the session will receive further information about abstract submission and fee waivers from the award committee.

Inquiries and materials should be emailed or mailed to either one of:

Trevor Hastie
Student Paper Selection Committee
Statistical Computing Section
Statistics Department, Sequoia Hall
Stanford CA 94305
`trevor@playfair.stanford.edu`

Daryl Pregibon
Student Paper Selection Committee
Statistical Computing Section
Room 2C-264, AT&T Laboratories
600 Mountain Ave
Murray Hill, NJ 07974
`daryl@research.att.com`

All electronic submissions of papers should be in postscript.

SECTION OFFICERS

Statistical Graphics Section - 1996

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

Sally C. Morton, Chair–Elect
310-393-0411 ext 7360
The Rand Corporation
Sally_Morton@rand.org

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

Stephen G. Eick, Program Chair
708-713-5169
Bell Laboratories
eick@bell-labs.com

Dianne H. Cook, Program Chair–Elect
515-294-8865
Iowa State University
dicook@iastate.edu

Mario Peruggia, Newsletter Editor (96-97)
614-292-0963
Ohio State University
peruggia@stat.mps.ohio-state.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

Lorraine Denby, Rep.(96-98) to Council of Sections
908-582-3292
Bell Laboratories
ld@bell-labs.com

Colin R. Goodall, Rep.(95-97) 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

Statistical Computing Section - 1996

Sallie Keller-McNulty, Chair
703-306-1883
National Science Foundation
smcnulty@nsf.gov

Daryl Pregibon, Chair–Elect
908-582-3193
AT&T Laboratories
daryl@research.att.com

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

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

James L. Rosenberger, Program Chair–Elect
814-865-1348
The Pennsylvania State University
JLR@stat.psu.edu

Mark Hansen, Newsletter Editor (96-98)
908-582-3869
Bell Laboratories
cocteau@bell-labs.com

Evelyn M. Crowley, Secretary–Treasurer
317-494-6030
Purdue University
crowley@purdue.edu

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

MaryAnn H. Hill, Rep.(95-97) 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) Council of Sections
312-702-8332/8333
The University of Chicago
r-thisted@uchicago.edu

Janis P. Hardwick, Rep.(96-98) Council of Sections
313-769-3211
University of Michigan
jphard@umich.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:

Mark Hansen
Editor, Statistical Computing Section
Statistics Research
Bell Laboratories
Murray Hill, NJ 07030
(908) 582-3869 • FAX: 582-3340
cocteau@bell-labs.com

Mario Peruggia
Editor, Statistical Graphics Section
Department of Statistics
The Ohio State University
Columbus, OH 43210-1247
(614) 292-0963 • FAX: 292-2096
peruggia@stat.mps.ohio-state.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

Nonprofit Organization
U. S. POSTAGE
P A I D
Permit No. 50
Summit, NJ 07901

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3402
USA

This publication is available in alternative media on request.