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

Summer 99 Vol.10 No.1

Statistical COMPUTING & GRAPHICS

A WORD FROM OUR CHAIRS

Statistical Computing

James L. Rosenberger is the 1999 Chair of the Statistical Computing Section. He discusses the many opportunities for those interested in Statistical Computing.

This is my first column as chair of the Statistical Computing Section. It is a pleasure to serve the Section with so many supportive and contributing members. I want to especially thank Mark Hansen, for his continued service as the Newsletter editor and in his role as Program Chair for the exciting program he has put together for the Joint Statistical Meetings (JSM) in August of this year. Details are given in this Newsletter and on the

CONTINUED ON PAGE 2

Statistical Graphics

Di Cook is the 1999 Chair of the Statistical Graphics Section. She points out several new challenges for Statistical Graphics and outlines plans to revamp the Section's Web Site.

As I write this, here in Iowa we're budding into spring. The leaves have almost universally appeared on the trees and shrubs. The apple, plum and cherry trees are in full bloom, the tulips have almost dropped all their petals, and the strawberries are about to fruit, having started flowering several weeks ago. The prairie is coming alive with purple cone flower plants emerging from the bare soil. People are visibly enjoying life outside

CONTINUED ON PAGE 3

SPECIAL FEATURE ARTICLE

Internet Measurement and Data Analysis: Topology, Workload, Performance and Routing Statistics

By kc claffy and Sean McCreary

> *Scientific apparatus offers a window to knowledge, but as they grow more elaborate, scientists spend ever more time washing the windows.*
>
> *– Isaac Asimov*

The infrastructure of the Internet can be considered the cybernetic equivalent of an ecosystem. The *last mile* connections from the Internet to homes and businesses are supplied by thousands of capillaries, small and medium sized Internet Service Providers (ISPs), which are in turn interconnected by "arteries" maintained by transit (backbone) providers. The global infrastructure of the Internet consists of a complex array of competing telecommunications carriers and providers, a very difficult infrastructure to analyze diagnostically except within the borders of an individual provider's network. Nonetheless, insights into the overall health and scalability of the system are critical to the Internet's successful evolution.

Attempts to adequately track and monitor the Internet were greatly diminished in early 1995 when the National Science Foundation (NSF) relinquished its stewardship role over the Internet. The resulting transition into a competitive industry for Internet services left no framework for the cross-ISP communications needed for engineering or debugging of network performance problems and security incidents. Nor did competitive

CONTINUED ON PAGE 4

EDITORIAL

In this edition of the Newsletter, we are pleased to bring you two Internet-related articles. The first is by the founders of an organization called CAIDA (Cooperative Association for Internet Data Analysis). In this special feature article, Kim Claffy and Sean McCreary outline for us some of the challenges faced by researchers trying to characterize the overall health of the Internet. As we begin to rely more and more on Internet-based applications, this kind of investigation will become increasingly important. Kim and Sean describe a mix of measurement tools and related data analysis coming out of CAIDA. I [Mark] asked them to contribute this article because I think this is an area rich with statistical applications. There is a great opportunity here! Within Bell Laboratories, I [Mark] have seen several Internet-, or more broadly network-related projects spring up.

The second Internet-related article comes from Wolfgang Härdle, Sigbert Klinke and Steve Marron. It is about the use of Internet-enabled applications for the teaching of statistics. These authors establish criteria for effective web-based teaching and propose an approach (using XploRe) to meet these challenges. This article contrasts nicely with an earlier piece by Deborah Nolan and Duncan Temple Lang on the use of multimedia for statistics education (this Newsletter, Volume 9, Number 1). It is interesting to compare these approaches both in terms of the technology and the underlying pedagogy. Wolfgang, Sigbert and Steve offer us several well-presented examples of how the connection between Java and XploRe has produced a rich environment for teaching.

With this issue of the Newsletter, we are pleased to introduce Graham Wills as a regular contributor. In his first column, Graham reviews some basics of linked data views. His involvement in the development of several such applications gives him a great perspective to comment on this type of interactivity. Please join us in welcoming Graham to the Newsletter staff!

Other items of interest in this edition include the latest column by Andreas Buja, Editor of the *Journal of Computational and Graphical Statistics* (JCGS). Andreas has a lot of great news to report about JCGS: Submissions are up and the journal is growing! Dan Carr and Ru Sun discuss an alternative to standard Trellis displays, involving layering and perceptual grouping. For the latest in Section news, turn to page 33 where you will read about the Computing Section's Student paper competition winners. You will also find details about a prestigious ACM award given to John M. Chambers for the development of S. John has taken his $10,000 prize and has established an award to promote students developing software for statistical computing.

As usual, we are eager to get feedback about the Newsletter. Send us some e-mail with your comments.

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

Antony Unwin
Editor, Statistical Graphics Section
Universität Augsburg
unwin@uni-augsburg.de

Statistical Computing

CONTINUED FROM PAGE 1

ASA web site at (www.amstat.org). Thanks also to Lionel Galway, our awards officer, who with a committee of the Council of Sections (COS) representatives, has completed the task of selecting the Statistical Computing best student paper awards. The four winners will present their papers at the JSM in Baltimore, and be recognized at the Section business meeting and mixer. I want to also express thanks to Tom Devlin for his continued work as the Electronic Communication Liaison, and webmaster for the Section. See (www.amstat.org/sections/) for the latest information on our Section. Along with the other officers of the Statistical Computing Section I hope we can make this an eventful year for the Section.

The Section needs to hear from you. We welcome your input and suggestions for what the Section can do for you, but even more importantly, we would like to also know what you can (and would like to) do for the Section, and it's more than 2,500 members. We will have a brief survey form on the Section web page for you to complete and express your views and give suggestions.

On the subject of what one can do for the Section: At the Interface '99 Symposium, it was a pleasure to receive, on behalf of the ASA, a gift of $10,000 from John Chambers, of Bell Labs - Lucent Technologies, to endow an annual student prize for a statistical software project. The prize will be awarded to the winning entry, beginning August 2000, with details announced at this year's JSM.

Please check the programs planned for the upcoming JSM announced in this Newsletter. Mark Hansen has planned an excellent program for Baltimore, with five invited sessions, three special topic contributed sessions, and 11 regular contributed sessions. A new feature for the Statistical Computing Section is a competition at the meetings for the best presentation award, based on survey forms the chairs will distribute at each session. Look for the results in the next Newsletter.

On the employment front, Statistical Computing was a featured profession in the Sunday June 6 issue of the Washington Post technology employment section. The article touted "growing opportunities" for exciting employment and "unprecedented demand for professionals with skills in the growing discipline of statistical computing." In addition to quoting the former, Karen Kafadar, and current chair of our Section, the article quoted Kennedy of Iowa State, Bateman of the Census Bureau, Wegman of George Mason, Mauer of MathSoft, Battaglia of SPSS, and Rovner of SAS. We should continue to promote the benefits of our profession.

I'm trying out a new software product I've just purchased (this is not an endorsement) called Dragon Naturally Speaking. This program provides an alternative input device to your computer keyboard, namely the microphone, using statistical speech recognition. Some of you may already be using speech recognition software, however for me this signals the beginning of a new revolution in how we interact with the computer. In addition to dictating text and punctuation, you can speak commands to correct mistakes, open and close windows, and turn the microphone on and off.

Our own profession, statistics, continues to make a significant contribution to the development of this technology. One can read from the software license agreement: "It is understood by you that speech recognition is a statistical process, that recognition errors are inherent in the process of speech recognition, and that speech recognition applications must be designed to allow for such errors in the recognition process." It is reassuring to see the recognition of the statistical technology inherent in this software. As we tackle increasingly difficult problems in dealing appropriately with the uncertainty in scientific processes and in how we model them, we place new and challenging demands on the statistical profession – I predict exciting times ahead.

Another topic that the Statistical Computing Section should put on its radar screen, it is the exploding interest in the data mining software being developed for use by businesses to mine their business data bases. Although this development is driving expanded use of statistical analysis, our community is only peripherally involved in the process. Many of the challenging issues that arise fall between the interface of the computer and the user, and the computer and the databases. Also, due to the size and complexity of the problems, visualization should be an integral part of the research efforts. Analyzing the data is but one aspect, presenting and visualizing the information is perhaps the more important part, especially for the audience most interested in using and interpreting the findings.

James L. Rosenberger
Pennsylvania State University
jlr@stat.psu.edu

Statistical Graphics

CONTINUED FROM PAGE 1

their houses, too. Its a time that I often reflect on where we are and where we could go, after the exhausting final weeks of the semester.

This year I'd like to concentrate on building up our section web pages. We're grateful to Bob Newcomb for getting our web pages up much earlier than many other sections did, and for maintaining them so well. However, it will be hard to find candidates in the future for an office entitled "secretary/treasurer/webmaster," so it is probably time to give Bob a break. I'd like to establish an informal position of section webmaster, a person who will take the lead in expanding and maintaining the information on our pages. To help the webmaster get started, I'd initially like to hire a web page designer to work with us on a new design for the section's web site. Once we have a basic design, we can set up an automatic system to keep a large amount of the material updated. If you have some favorite graphics, examples of pages that you find pleasant to look at and informative to explore, or links to interesting graphics work, please send these to me. It would be wonderful if we could make the Statistical Graphics section web pages the flagship of the ASA web site!

Mike discussed the program for the 1999 joint meetings in the last newsletter, and Debby Swayne included a detailed article on the program (details can be found at `www.research.att.com/~dfs/jsm99/`, and details on the entire JSM program can be found at `www.amstat.org/meetings/jsm/1999/`. The program is wonderfully diverse. In addition to the regular program there will be two short courses sponsored

by Statistical Graphics: "Regression Graphics: Ideas for Studying Regressions Thru Graphics" by R. Dennis Cook, and "Statistical Shape Analysis" by Ian L. Dryden. Both are full-day courses.

The video library has changed hands, from Debby Swayne to David James, who is a member of the technical staff at Lucent Bell labs (`cm.bell-labs.com/cm/ms/who/dj/`). David will think about modernizing the technology of the video library, investigating whether we should begin digitizing videos for distribution over the internet. Information about the video library can be found at (`orion.oac.uci.edu/~rnewcomb/statistics/graphics/library/library.html`). The video library will be featured in the all day video theater at the joint meetings.

This is an exciting time to be involved in statistical graphics. There are some great challenges to be researched: design of web-based graphics, visual exploration of databases, developing interactive techniques for massive data. On this last topic, the National Science Foundation, has just issued a new initiative titled "Large Scientific and Software Data Set Visualization" with details at `www.nsf.gov/cgi-bin/getpub?nsf99105`. You need to have at least 100Gb of data available to test your methods out! Also, this is time when many companies are rapidly developing web pages with statistical summaries, and they need input from statisticians familiar with good statistical graphics, from appropriate color combinations to appropriate reliability/variability representation. In addition, the new language Java keeps developing, and opens up the world of multi-platform computing to graphics! The Java 3DAPI allows us to readily contemplate drawing statistical graphics in a 3D display. This creates an entirely new environment for displaying statistical data, and will result in entirely new types of graphics. Finally, there are always new issues to deal with for graphics when the dimensionality of the data (number of variables) is large.

One final word, thanks to our newsletter editors, Mark Hansen and Antony Unwin. This newsletter is a wonderful document! It is hard work to get the quality articles and beautiful graphics together several times a year.

Di Cook
Iowa State University
`dicook@iastate.edu`

SPECIAL FEATURE ARTICLE (Cont.)

CONTINUED FROM PAGE 1

providers, all operating at fairly low profit margins and struggling to meet the burgeoning demands of new customers and additional capacity, place a high priority on gathering or analyzing data on their networks. This attitude is strengthened by the general lack of quality measurement or analysis tools to support these endeavors, and the absence of baseline data against which an analyst can track changes in the system's behavior.

As a result, today's Internet industry lacks any ability to evaluate trends, identify performance problems beyond the boundary of a single ISP, or prepare systematically for the growing expectations of its users. Historic or current data about traffic on the Internet infrastructure, maps depicting the structure and topology of this amorphous global entity, or projections about how it is evolving simply do not exist.

That is not to say that no measurement of the Internet occurs. There are numerous independent activities in the area of *end-to-end measurement* of the Internet. Typically spawned by end users with an interest in verifying performance of their Internet service, these measurements involve an end host sending active probe traffic out into the network and recording the delay until that packet returns to its source. Unfortunately such traffic measurements involve a large number of parameters that are difficult if not impossible to model independently, and the resulting complexity renders elusive any comparability or useful normalization of the gathered data. There are research groups trying to deploy technology and infrastructure to support more standardized measurement and evaluation of performance and reliability of selected Internet paths, and what specific segments of a given path limit that performance and reliability, but such efforts are slow and have thus far been unable to meet the needs of any of the user, research, or ISP communities.

Network measurements fall into two broad categories: Passive and active. Passive measurements depend entirely on the presence of appropriate traffic on the network under study, and have the significant advantage that they can be made without affecting the traffic carried by the network during the period of measurement. However, it can be much more difficult or impossible to extract some of the desired information from the available data.

Active measurements, on the other hand, directly probe network properties by generating the traffic needed to make the measurement. This allows much more di-

rect methods of analysis, but also presents the problem that the measurement traffic can have a negative impact on the performance received by other kinds of traffic. This tension between measuring the performance of a network and actually using it to carry real traffic necessitates care in the design of any program of active measurements. In the remainder of this paper we will highlight activities and outstanding problems in the areas of passive and active measurements. We will conclude with a focus on near-term research priorities and forecast activities for the next five years.

Passive Techniques: Workload Analysis

> *Everything you've learned in school as "obvious" becomes less and less obvious as you begin to study the universe. For example, there are no solids in the universe. There's not even a suggestion of a solid. There are no absolute continuums. There are no surfaces. There are no straight lines.*
>
> *– R. Buckminster Fuller*

Workload measurements require collecting traffic information from a point within a network, e.g., data collected by a router or switch or by an independent device passively monitoring traffic as it traverses a network link. Network traffic is carried in discrete units called "packets," and they typically vary in size. As a result, measurements are quoted both "per packet" or "per byte." Collection of this data allows for a variety of traffic analyses (e.g., composition of traffic by application, packet size distributions, packet inter-arrival times, performance, path lengths) that contribute to our ability to engineer next generation internetworking equipment and infrastructures. Of particular interest to network operators are traffic flow matrices: tables of how much traffic is flowing from a given source to a given destination network, information that turns out to be vital to optimizing engineering decisions that govern which other transit ISPs to exchange traffic with directly, and where to set up those interconnections.

Figure 1 (page 7) shows a sample matrix of traffic taken from a major backbone site. The units along the horizontal axes are "Autonomous Systems," the organizational units used in routing packets in the backbone. A single Autonomous System may contain many separate networks, but they are all part of a single administrative organization. Since an Autonomous System is the unit at which Internet routing relationships are established and negotiated, a traffic matrix at this granularity is of immediate utility to networking engineers trying to optimize topology or route peering decisions. *Peering* is the relationship between two Autonomous Systems that agree to exchange routing information with each other. A graph at this granularity can be used to determine the traffic balance among ISPs.

Other levels of granularity are interesting as well. For example, a network manager for a corporation or university might want to know which departments exchange the most traffic. Since the entire organization is most likely a single Autonomous System, a traffic matrix could be constructed with individual departmental or workgroup networks along the x and y axes. Figure 2 (page 7) shows a traffic matrix by country, a level of granularity interesting from a public policy as well as an international commerce perspective. Reflecting a two-minute sample taken at a United States peering point location in 1998, this particular image indicates that the U.S. is an international communications hub, as seen by the presence of traffic between points outside the U.S. via the U.S. The log-scale highlights that it is, however, still quite a small fraction of overall traffic, but it is a useful statistic to be able to track. Note the U.S. is almost universally a net exporter of IP traffic.

As another example of relevant workload characteristsics, we present sampled data on Internet packet sizes. Statistics of packet size distribution and arrival patterns are of relevance to designers of network routing and switching equipment since there are both per-packet and per-byte components of the cost of switching a packet, so having metrics for typical Internet workloads allows designers to optimize hardware and software architectures around relevant benchmarks.

Figure 3 (page 7) shows the cumulative distribution of packet sizes, and of bytes by the size of packets carrying them. This figure is a good example of the difference between per-packet and per-byte analyses. There is a predominance of small packets, with peaks at the common sizes of 44, 552, 576, and 1500 bytes. The small packets, 40-44 bytes in length, include TCP (Transport Control Protocol) acknowledgement segments, TCP control segments such as SYN (synchronize sequence numbers), FIN (no more data from sender), and RST (reset the connection) packets, and telnet packets carrying single characters (keystrokes of a telnet session). Many TCP implementations that do not implement Path MTU Discovery use either 512 or 536 bytes as the default Maximum Segment Size (MSS) for nonlocal IP destinations, yielding a 552-byte or 576-byte packet size (Stevens, 1994). A Maximum Transmission Unit (MTU) size of 1500 bytes is characteristic of Ethernet-attached hosts.

Almost 75% of the packets are smaller than the typical

TCP MSS of 552 bytes. Nearly half of the packets are 40 to 44 bytes in length. Note however that in terms of bytes, the picture is much different. **While almost 60% of packets are 44 bytes or less, constituting a total of 7% of the byte volume**, **over half of the bytes are carried in packets of size 1500 bytes or larger**.

Another important workload analysis is the assessment of composition of traffic by protocol type, since some protocols are 'friendlier,' or more responsive to network signals of congestion, than others, and a strong growth in the proportion of such unfriendly protocol traffic would have unsalutary implications for the infrastructure. On the Internet, standard implementations of TCP (Transport Control Protocol) are friendly, while UDP (User Datagram Protocol) implementations are not. Fortunately for the stability of the infrastructure, TCP is the protocol that carries most popular applications known to users today: web pages (HTTP), e-mail (SMTP), and Usenet news (NNTP).

It is often useful to aggregate all packets that represent a single conversation between two endpoints into a single unit called a 'flow.' The distribution of Internet traffic flow lengths, as measured in packets, is heavy-tailed. Our measurements indicate that the majority of flows have a very short duration, e.g., HTTP, SMTP, carrying much less traffic than the kind of bulk data transfer flows for which TCP has been optimized, e.g. FTP and NNTP. Of particular concern is the effect of the increasing popularity of streaming and other multimedia applications that are much larger, often orders of magnitude, than even historically 'bulky' ones. The fairly limited resource accounting and pricing models currently in use along with several other fundamental aspects of the infrastructure make this significant shift in the distribution of flow sizes rather ominous for the stability of the current framework. Indeed, only more accurate resource consumption and concomitant pricing models will help providers grow their infrastructure to keep pace with demand. Changes in this direction would be auspicious for the industry anyway, since moving away from the flat-rate economic model that prevents rational valuation of the utility of Internet service will help maximize the value received by the end user.

We have only provided a few examples of the potential information available via passive monitoring tools. Other applications of passive monitoring include identifying, characterizing, and tracking of: the potential benefit and optimal configuration of web caches and proxies; security compromises to one's infrastructure; the elasticity of flows and effectiveness of congestion control algorithms; the extent to which traffic growth is due to additional users versus an increase in per-user traffic; changes in profile of popular protocols and applications; and penetration and impact of emerging technologies and protocols such as multicast or IPv6.

It is sometimes possible to extract other parameters that normally require active measurement techniques opportunistically from passively collected traffic. This is a very attractive prospect because any information we can obtain through passive techniques is 'free' in the sense that we don't have to impose any extra load on the network under study. More work needs to be done in this area to better identify what can and can't be measured passively so that we can better exploit this valuable source of information.

Active Techniques: Mapping the Internet Ecosystem

> *In an expanding system, such as a growing organism, freedom to change the pattern of performance is one of the intrinsic properties of the organism itself.*
>
> *– Unknown*

New connections among core Internet backbones occur hourly, ranging in capacity from T1 copper (1.55 megabits per second) to OC48 fiber optics (2.48 gigabits per second). This physical structure supports a myriad of new technologies and products, including live (or 'streaming') audio and video, distance education, entertainment, telephony and video-conferencing, as well as numerous new and evolving communications protocols.

Tracking and visualizing Internet topology in such an environment is challenging at best. A particularly ambitious endeavor is underway in our group at CAIDA (Cooperative Association for Internet Data Analysis) through the recent development of `skitter`, a tool for dynamically discovering and depicting global Internet topology. The data collected by `skitter` is useful for more than just topological visualization, however, since it also contains a lot of information about the performance of specific paths through the Internet.

`skitter` works using a process somewhat analogous to medical x-ray tomography, a technique where a three-dimensional image is achieved by rotating an x-ray emitter around the subject and measuring the intensity of transmitted rays from each angle, and then reconstructing the resulting two-dimensional images into a three-dimensional object. Geologists rely on similar techniques to build models of seismic activity using cross-section images (slices) of the earth. Data gathered from tomographic scans play an important role in developing models to analyze and predict select phenomena.

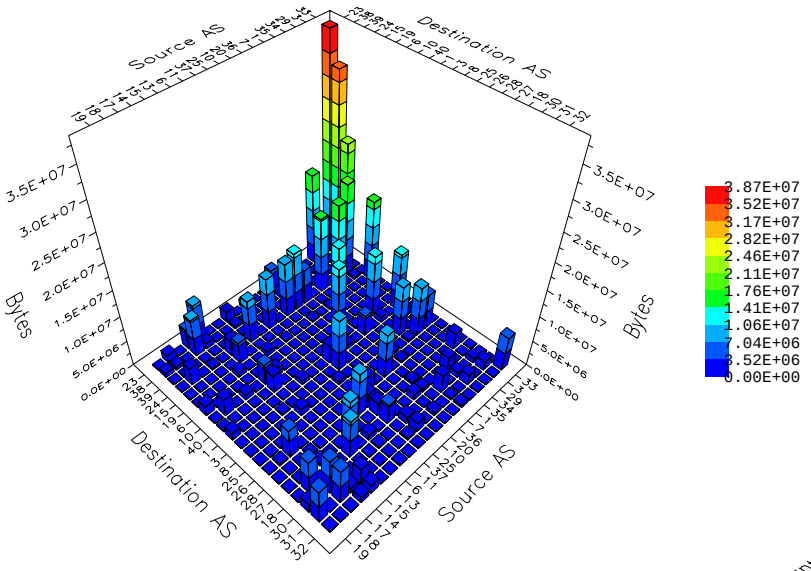


Figure 1: Sample matrix of traffic from source to destination ASes (2 minute sample from FIX-west in April 1998).

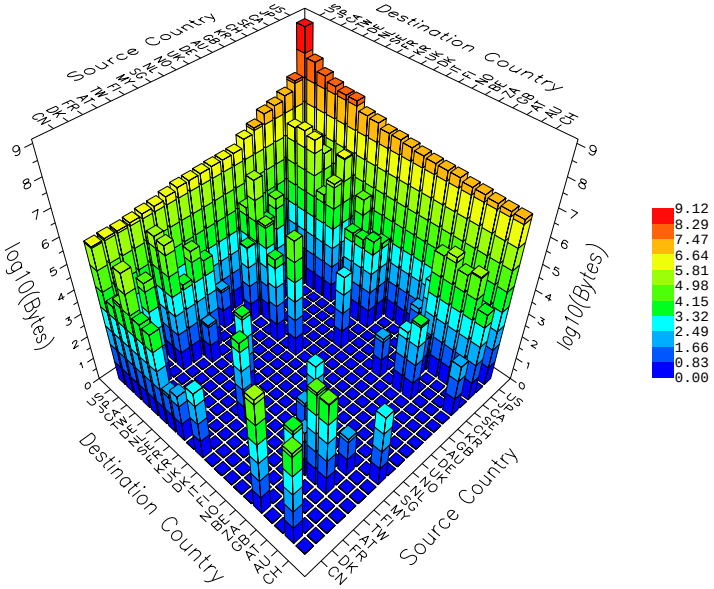


Figure 2: Sample matrix of traffic from source to destination countries (2 minute sample from FIX-west in April 1998).

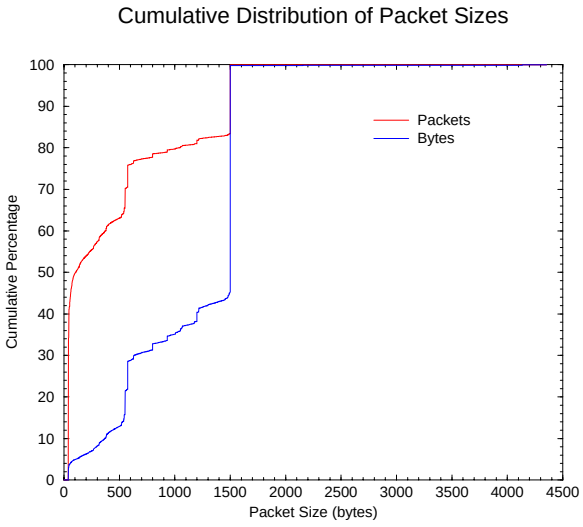


Figure 3: Cumulative distribution of packet sizes, and of bytes by the size of packets carrying them.

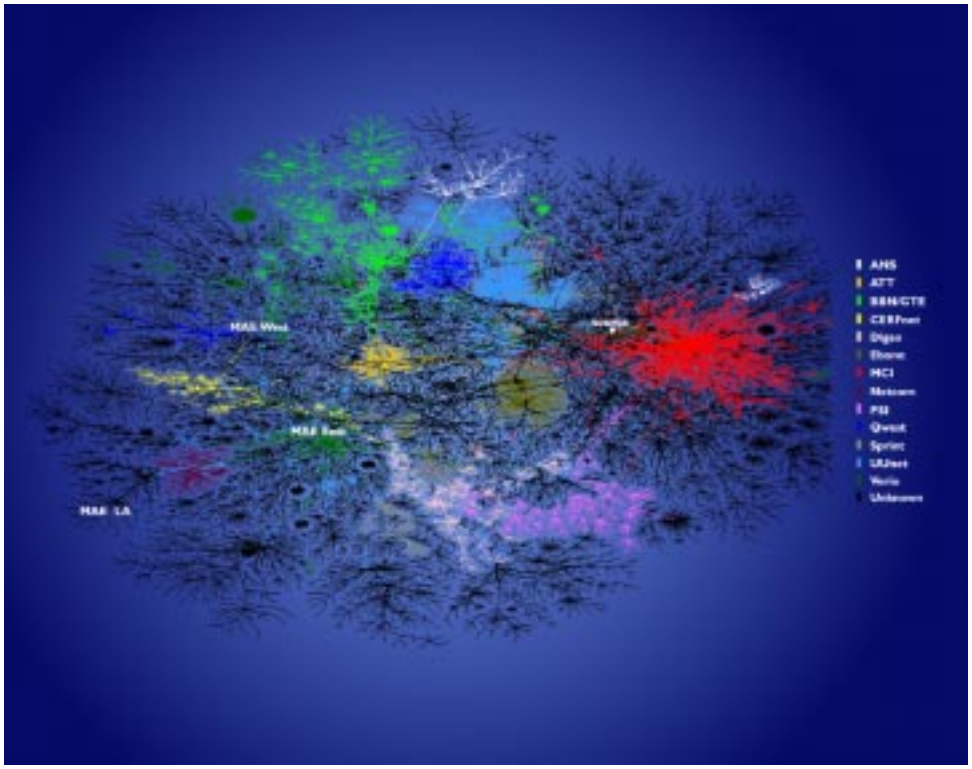


Figure 4: Prototype two dimensional image depicting global connectivity among ISPs as viewed from `skitter`. The layout algorithm used in these images was developed by Hal Burch (CMU) in support of Bill Cheswick's (Bell Labs) Internet Mapping Project (see the Acknowledgements for a URL for this project).

Figure 5: Box & whisker plot of delay values measured from `lancelet.caida.org` (in Ann Arbor, MI) to `www.ucsd.edu` (in San Diego, CA) (log scale).

Figure 6: Histogram of RTTs for 1600 probes from `lancelet.caida.org` (in Ann Arbor, MI) to `www.freebsd.org`.

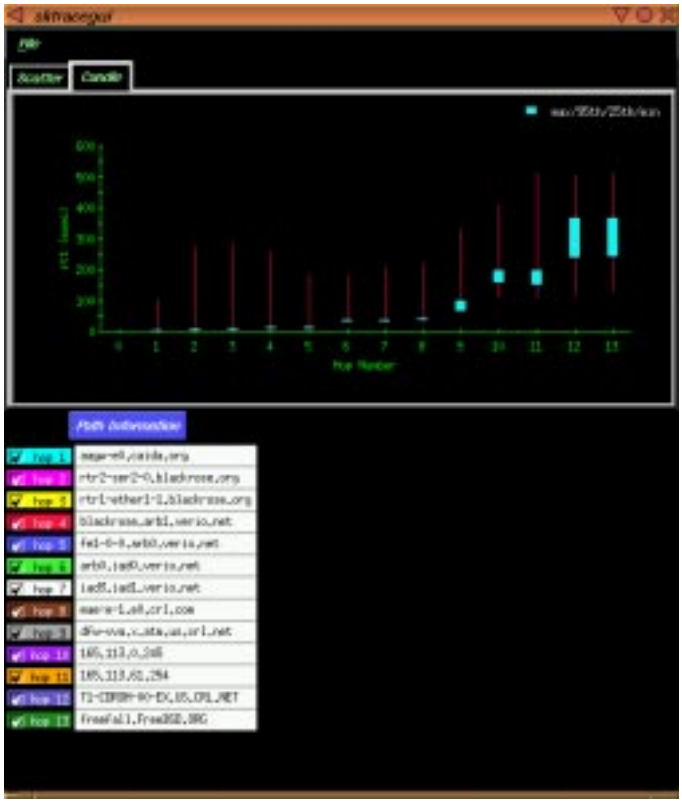


Figure 7: Box & whisker plot of delay values measured along the path to `www.freebsd.org`. For each hop along the path, the blue box delineates the 25th and 75th percentile of RTTs to that hop; the ends of the whiskers delineate the minimum and maximum values.

CAIDA is currently using `skitter` to gather infrastructure-wide (global) connectivity information (what's connected to what?), and round trip time (RTT) and path data (how does a packet get from A to B and how long does it take?) for more than 30,000 destination hosts from six source monitors spread throughout the United States, with additional monitors deployed in the U.S., Europe, and Asia in mid-1999.

`skitter` measures the Internet path to a destination by sending several ICMP (Internet Control Message Protocol) echo request packets towards the destination host in a similar fashion to the common 'ping' utility. However, `skitter` sends these packets with very low 'time to live' (TTL) values. Every router in the Internet automatically decrements the TTL value of each forwarded packet as part of an overall scheme to prevent persistent traffic looping. If the TTL value reaches zero, The router will discard the packet and send an error notification back to the sender. `skitter` sends a series of probe packets from the measurement host with progressively larger TTL values, and as each error message is received the measurement host is able to determine the path taken through the network from the source to the destination. This is essentially the same procedure as used by the `traceroute` utility.

Probing paths from multiple sources to a large set of destinations throughout the current Internet address space allows both topological and geographical representations of a significant fraction of Internet connectivity, the latter admittedly constrained by the abysmal lack of geographic mapping data for Internet address space. Supporting tools also analyze the frequency and pattern of routing changes (when and how often are alternative paths used between the same two endpoints?)

Like any active measurement program, it is essential that skitter measurements impose only a minimal load on the infrastructure as it takes its measurements. `skitter` probe packets are very small, 52 bytes in length, and typically only probe destination hosts at approximately hourly intervals.

Analyzing data from tens of thousands of path measurements can identify critical roles that specific backbones, traffic exchange points, and even individual routers play in transmitting Internet traffic. Figure 4 shows a preliminary two-dimensional visualization of `skitter` data depicting a macroscopic snapshot of Internet connectivity, with selected backbone ISPs colored separately. The graph reflects 23,000 end destinations, through many more intermediate routers.

In addition to collecting overall topological information, active measurement techniques can be used to probe the network for specific problem conditions. Figure 5 shows an example delay distribution, with the common *heavy-tail* characteristic of many Internet end-to-end delay distributions, where many points lie above the lower band of the majority of the data. (Each data point in this plot depicts the distribution of 400 delay values. The blue box delineates the 25th and 75th percentile of those 400 values; the ends of the whiskers delineate the minimum and maximum values. This plot shows a heavy tailed distribution across a fairly long period of time.)

Data from these and other sites over time suggest that even under the best conditions a significant fraction of Internet traffic takes longer than expected to reach its destination. This characteristic produces a tendency for heavy-tailed distributions of round-trip times on the global Internet. Deviations from this nominal behavior are usually indicative of problems in the network.

Figure 6 shows a histogram of round trip times (RTTs) from the same measurement host to `www.freebsd.org`, a host in Northern California, during a workday hour in February 1998. In this example, most of the RTT data is well above the minimum RTT, the cumulative packet loss is around 10%, and the packets not dropped have a relatively wide distribution of RTTs. It seems likely that rather than a single standing queue (where packets are delayed waiting to be processed by some router on the path), this path is subject to one (or more) changing queues.

Figure 6 shows the distribution of RTTs for 1600 probe packets; the green vertical line represents the median value. The distribution is nearly symmetric and reasonably persistent over long time intervals. Such an RTT distribution coincident with packet loss could arise from congestion and global synchronization on a link (Monk, Claffy and McRobb, 1999). A hop-by-hop analysis of the entire path from source to destination can yield some clues to the orgin of the problems.

Figure 7 shows a box & whisker plot for the RTTs to each hop along the path to `www.freebsd.org`. Each point on the graph depicts the distribution of RTTs for probe packets that were discarded at that router. Interpretation of this kind of data is fairly difficult in general. Each data point includes effects from previous data points, since every probe packet must pass through the earlier routers before being discarded at a later one. Furthermore, the RTT distributions also include significant unknown contributions from the unobserved reverse path back from the router to the measurement host. Much of the routing in the Internet today is highly asymmetric, and so differences in the return path taken

by the error response packets from different routers can sometimes dominate the total RTT measured to those routers. In this case however, there is a clear-cut interpretation: The evidence points to congestion between hop 11 and hop 12: the minimum RTT values look similar, but hop 12 has a higher median, wider and more symmetric distribution, and strong correlation to the distribution for the final destination. All of the hops exhibit a heavy-tailed distribution except hops 11 and 12.

The robustness and reliability of the Internet are highly dependent on efficient, stable routing among provider networks. Analysis of real world Internet routing behavior has direct implications for the next generation of networking hardware, software and operational policies. Observations of macroscopic routing dynamics provide insights into:

- effects of outages on surrounding ISPs
- effects of topology changes on Internet performance
- unintended consequences of new routing policies
- potential areas for improving an individual networks' ability to respond to congestion and topology changes .
- infrastructural vulnerabilities created by dependencies on particularly critical paths

One important area of needed work is the comparison of actual behaviour of routed traffic with with those routing policies articulated by the inter-domain routing protocol BGP (Border Gateway Protocol). This is the protocol used to exchange routing information, and it is the primary mechanism for implementing traffic exchange policies among ISPs. Comparing the ambient BGP view of the network with actual traffic routes requires a source of core Internet routing (BGP) data close to the source of the active measurements; tools for acquiring this data with high precision are still a research as well as political challenge, as many ISPs are hesitant to make this kind of information publically available.

Other areas of analysis with strong technical and policy implications: assessing the effectiveness of utilization of the IP address space; extent of asymmetric routing and route instability as a function of service provider and over time; the distribution of traffic by network address prefix lengths; efficiency of usage of BGP routing table space, e.g., via aggregation; favoritism of traffic flow and routing toward a small proportion of the possible addresses/entities; degree of incongruity between unicast and multicast routing; and quantifying effects on connectivity of removal of specific ASes.

Emerging Tools

Performance measurement techniques are often used by network engineers in diagnosing network problems; however, most recently their application has been by network users or researchers in analyzing traffic behavior across specific paths or the performance associated with individual Internet Service Providers (ISPs). A recent development in the industry is the offering of *service level agreements* (SLAs), contracts to guarantee a specified level of service, subject to cost rebates or other consumer remuneration should measurements suggest that the ISP did not adhere to the SLA. SLA's are rather controversial in the community since there is no standard metric or even measurement methodology for calibrating them. CAIDA will focus on tools and techniques for more generic active measurement rather than the typically proprietary tools currently used to monitor SLAs.

CAIDA is among the groups producing tools that utilize active measurement techniques to help visualize network problems, several of which were illustrated in the previous section. There are many other active performance measurements efforts undertaken by various players in the Internet community, the most popular of which are typically user-instigated 'Internet weather reports', a selection of which are described in Nancy Bachman's page `www.caida.org/Weather`. The most important deliverables of most current active monitoring tools focus on either verifying bandwidth or performance stated or implied by vendors and providers, or ascertaining those parameters if the information is not available in the first place. But there are an enormous number of research questions not under concerted investigation at the moment due to the lack of adequate active tools for doing so. Identifying and locating what might be construed as particularly topologically critical pieces of the public infrastructure is one area that the developers of the `skitter` platform hope to accomplish. Others include: finding particular periodic cycles or frequency components in performance data; developing a calculus for describing and drawing the difference between two given 'snapshots' of network performance; finding the topological 'center' of the net; techniques for real-time visualization of routing dynamics; and correlation with passive measurements.

Near-term Priorities

> *Science is not about control. It is about cultivating a perpetual sense of wonder in the face of something that forever grows one step richer and subtler than our latest theory about it. It is about reverence, not mastery.*
>
> *– Richard Powers from the Gold Bug Variations*

Each measurement effort provides a new window on the infrastructure for network operators, designers and researchers. But without well-considered, strategically deployed, and collaboratively maintained measurement tools/infrastructure, these windows are not necessarily offering any useful insight. A particular obstacle is the lack of reasonable knowledge base for mapping IP addresses to more useful analysis entities: Autonomous Systems (BGP routing granularity), countries, router equipment (multiple IP addresses map to same router but without any mechanism for deriving the mapping), geographic location information (latitude/longitude coordinates). There are efforts underway to develop prototype databases for canonical mappings; `www.caida.org/INFO` lists some of them, but their precision, completeness, and concomitant utility will require more concerted community participation.

Indeed, progress in this field requires both top-down and bottom-up pursuit: application developers must scope out what measurements would allow their software to negotiate performance constraints with the network, and Internet service providers need to participate in deploying and evaluating the utility of measurement technology for their own network design, operation, and cost recovery.

The network research community is in a difficult position between these two groups, hoping to design a framework for windows that are useful. For several years the infrastructure was in such a measurement-deprived state that even deploying any data collection tool at all qualified as ground-breaking work. The current state is quite different: there is plenty of measurement occurring, albeit of questionable quality. The current community imperative is rather for more thoughtful infrastructure-relevant analysis of the data that is collected, in particular correlating among data sources/types, and providing feedback into tool design to improve future data acquisition techniques. Unlike many other fields of engineering, Internet data analysis is no longer justifiable as an isolated activity. The ecosystem under study has grown too large, and is under the auspices of too many independent, uncoordinated entities. Nonetheless, the system is evolving rapidly, and prudence would dictate that the depth and breadth of our understanding of it follow in much closer pursuit.

References

Monk, T., Claffy, k, and McRobb, D. (1999) "Internet Tomography: Analysis and Visualization of Global Traffic," in *Proceedings of INET'99*, San Jose, CA (forthcoming).

Stevens, W. R. (1994), *TCP/IP Illustrated, Volume 1: The Protocols*, Addison-Wesley.

Acknowledgements

Thanks to Daniel McRobb for help with the sections on performance and routing, and to Nancy Bachman for helpful editing comments. Many thanks to Bill Cheswick and Hal Burch (Lucent/Bell Laboratories) for providing the graph layout code for Figure 1. For more information see `www.cs.bell-labs.com/~ches/map/`.

kc claffy founded CAIDA, a collaborative organization supporting cooperative efforts among the commercial, government and research communities aimed at promoting a scalable, robust Internet infrastructure. CAIDA is based at the University of California's San Diego Supercomputer Center (SDSC). Support for these efforts is provided by CAIDA members and by the Defense Advanced Research Project Agency (DARPA), through its Next Generation Internet program, and by the National Science Foundation (NSF). More information is available at `www.caida.org`.

kc claffy
University of California, San Diego
Cooperative Association for Internet Data Analysis (CAIDA)
`kc@caida.org`

Sean McCreary
University of California, San Diego
Cooperative Association for Internet Data Analysis (CAIDA)
`mccreary@caida.org`

Connected Teaching of Statistics

By Wolfgang Härdle, Sigbert Klinke, and J. S. Marron

1. Introduction

The study of statistics is commonly considered difficult by students, since it requires a variety of skills including quantitative and graphical insights as well as mathematical ability. Yet an increasing number of people need facility with quantitative methods and students need to acquire statistical capabilities because they are confronted with more and more data sets to be understood. In addition these data sets grow rapidly in size and structural complexity. An example for such data are the files that are collected on mobile phone applications, transaction records, etc. Despite these changing needs the teaching methods used have been surprisingly constant in recent years. An attractive and potentially powerful new way of updating current teaching methodology is via tools based on an intra- or the Internet. In this article we suggest a set of criteria for effective web based teaching and propose the first net based approach to meet these criteria.

New technology is accepted more widely if its use is immediately understandable and easy for everybody. The same is true for teaching statistics in face of the new challenges in structure and size of data. It can only be effective if statistical methods are explained in a way that gives the student easy access to them. Two viewpoints are important for understanding this effective teaching, that of the student and that of the teacher. The new additional component of web based computing in teaching has an impact on both viewpoints. For example, large data sets, interactive graphics and on-line information were unusable for undergraduate statistics teaching a few years ago. Now easy availability of these features requires an update of the criteria behind "what is good teaching?" from both viewpoints.

The student will benefit from

- Quick and easy access to methodology and data via browsers
- Interactive examples since doing is one of the fastest methods of learning
- Smooth transition from classroom to homework to full scale statistical tools

The teacher will benefit from

- Quick and easy broadcast of methodology and data via browsers
- A user friendly environment
- A powerful and flexible environment for dissemination of research

Many current teachers of statistics have a lot of inertia and reservation against changing teaching practice. Hence a stepwise plan towards smooth integration of web based teaching elements will gain the largest following. A series of steps through levels which allows gradual involvement and time commitment is:

Level 1: Display off the shelf class examples via a web browser. This requires only standard display equipment, and minimal effort by the instructor assuming ready made examples are available.

Level 2: Do examples as in Level 1, live on the web and give interested students the link coordinates of exercises, data or further programs and suggest some enriching examples they try on their own. This requires web access for most students and in the classroom. Again the ready made examples can be used and student questions will be minimal because no requirement is made of less capable students.

Level 3: Do examples as in Level 2, but assign homework using web based examples and methodology. This requires web access for all students and much more instructional support to address the inevitable questions and problems.

Level 4: Become a developer of examples. This involves more time and energy on the part of the teacher (the amount depends on the friendliness of the environment and on the integrability of other web based documents), but also yields the most rewards in terms of the customizability that more creative teachers will want. Our goal for teachers who reach this stage is to provide tools which will maximize their individual creativity.

2. The solution

Our approach to meet the above criteria for teaching statistics in elementary courses is based on macros written in XploRe (`www.XploRe-stat.de`). Some examples are discussed in detail in Section 3. Here we develop the general framework and present the various outlets of XploRe for different platforms. XploRe is the interactive statistical computing environment which works equally well on single user machines, intranetworks and web connected clients. This is technically

available via XploRe's server-client concept. The server (professor's machine) makes the course documents (data and statistical methods) available. The client (classroom machine or students PC) connects to the server via the web without additional software downloads. The Java technology (www.javasoft.com) and standard web browsers enable this universal access despite the well known heterogeneity across hardware platforms. The overhead of earlier methods of handing out software and data sets is thus dramatically reduced.

The quick and easy access to methodology and data via browsers comes from the good integrability of XploRe data, macros and tutorials into web documents. Since most students are familiar with using browsers on the Internet, there will be no overhead of learning the environment, which would otherwise distract from their learning the desired statistical lessons.

A set of interactive examples is discussed in Section 3 below. These are intended to illustrate the point that interactive learning is very effective and all based on a standard browser front end (with a Java Swing class from SUN). For example, when a student is involved in choosing numerical parameters for a particular case study, the level of thought needed, followed by anticipation of the answer, which is then immediately displayed, results in a deep type of learning. In particular, doing is the best method of learning.

Some of the earlier approaches (e.g. www.stat.sc.edu/~west/webstat/, www.stat.berkeley.edu/users/stark/Java/ and www.ruf.rice.edu/~lane/stat_sim/) to web based teaching of statistics have included a smooth transition from classroom to homework by allowing the student to use the Java applet shown in class also for homework. A natural next step to producing truly quantitatively equipped students is to also provide a smooth transition to full scale statistical tools that will continue to be useful long after the class is over. Because our examples are based on the statistical computing environment XploRe it is simple to move from classroom examples to more elaborate data analysis.

Traditional methods of conveying data to students, such as writing on a chalkboard or piece of paper, have severe limitations, due to the effort involved at both ends of the process. Exchange of floppy disks allows software, and also larger data sets to be conveyed, but this involves a lot of overhead in terms of effort (e.g. control of homogeneity of hardware platforms) on the part of the teacher. The Internet clearly allows quick and easy broadcast of methodology and data via browsers.

Many teachers of statistics have not learned web development skills, and perhaps may not have even learned other types of computational skills. For such potential users a user friendly environment means class examples must be already completely developed and ready to use. We offer classroom ready examples on (ise.wiwi.hu-berlin.de/statistik/lehrmaterial/statmat.html).

Other teachers of statistics will be higher end users, who have their own ideas for class examples, or else would like to customize those that are provided. For them a user friendly environment means the existing examples are coded in a very high level language, which is easy to modify, and provides a convenient basis for other types of development. XploRe is matrix (array) based and thus development occurs at a higher level than is available from Java programming. An important advantage of XploRe over other high level statistical languages such as SPLUS (www.mathsoft.com/Splus), GAUSS (www.aptech.com) or STATA (www.stata.com) is that XploRe macros may be automatically converted to web transparent methodology via an HTML translator.

Teachers who wish to modify the given examples, or develop their own, will need more than just a user friendly environment. They will also need a powerful and flexible environment, which contains a wide range of quickly usable tools. XploRe has a wide range of statistical tools with the possibility of specialization for different fields like finance, econometrics, etc. Java based approaches to specializing software for teaching cannot provide this full scale since they are based on combinations of the limited set of fixed applets available in the toolbox provided by the applets' constructor.

Statistical Technology on the net

Three hardware platforms are in widespread use for statistical computing and graphical data interaction: Macintosh, UNIX, and Windows. The first has a simple graphically oriented user interface and allows highly interactive dialogues with data. UNIX is used for high-speed and distributed computing but is often less satisfactory in graphical interaction. Windows aims at facilitating both high-speed computing and graphics but is weaker at present than UNIX for Internet access. Distributed computing is simply not possible under Windows unless one uses certain add-ons. An overview of current Internet technology and statistics is given by Symanzik (1998) (www.galaxy.gmu.edu/~symanzik)

Many software platforms for statistical computing exist but are unfortunately not easily interchangeable. The

reasons for this include the history of software development, the targeted user groups, and the optimization of certain software for specific hardware configurations. The original version of GAUSS (www.aptech.com), for example, was optimized for INTEL chips and, therefore, could not be transferred to the Mac or UNIX platforms. Now GAUSS is available on UNIX, but the UNIX version does not have a graphical device that allows, for e.g. interactive changes in the layout of graphs. SPLUS (www.mathsoft.com/Splus) was developed for UNIX systems and was only later transferred to PCs. Consequently, the PC version is different from the UNIX version. EVIEWS was developed for DOS and is now available for Windows but not for UNIX or for the Mac. TSP is a DOS program and is not easily transferred to a Windows/NT platform. SPSS exists for Windows but has still a batch structure that makes many mouse clicks necessary in order to generate implicitly the batch commands. STATA (www.stata.com) , SAS (www.sas.com) and SHAZAM (shazam.econ.ubc.ca) are unusual in that mutually compatible versions exist for all platforms. Besides the software that we mentioned here as examples, there are many other platforms which also share the property of heterogeneity.

Heterogeneity of software platforms creates relatively few problems if there is no need to exchange programs. Exchange of graphs, document files, and ASCII-based data sets can be carried out by FTP, provided that the user has the appropriate graphics plug-in and document reader (e.g., Ghostscript or Acrobat). However, there is also a need for exchangeable computer programs for implementing advanced statistical methods, as these are becoming increasingly complex mathematically, and writing the necessary programs can be a difficult and time-consuming task, which puts it effectively out of reach in many cases.

Graduate-level instruction in statistics provides one example of the usefulness of exchangeability. It is not unusual for a faculty member at one university to give a short course at another. In some cases, a faculty member at one university may use electronic communication to present a course at several geographically dispersed locations. Calculation of an estimator may require heavy computing that is available on the researcher's home machine. During the course, modifications of this estimator and different applications may be discussed, and these may require access to the software at multiple locations. Exchangeability of software is necessary to enable students at all locations to carry out computational and empirical exercises that the instructor has prepared at his own university.

Collaboration among researchers at different locations provides another example of the desirability of exchangeability. In this case, the goal is to enable each collaborator to carry out computations using the same software. Ideally such cooperation should be based on a pool of easily accessible software and computing power for all parties. For effective progress on a project that involves heterogeneous hardware and software, it is desirable for partners to have the ability to contribute methods despite being at different locations and working with different computing environments. In addition it may simply be a problem for a researcher who is a visitor in another establishment to be able to continue using his own programs.

On the other hand, heterogeneity of software does have the important advantage of enabling a developer of new methods to choose the software system that is best suited to the problem under consideration. Therefore the problem of exchangeability should not be solved by standardizing statistical software but by making software from different sources accessible to diversely equipped users.

3. Teachware Quantlet Examples

Example 1

This example illustrates that gathering a random sample is different from "just choosing some", i.e. proper random sampling requires some mechanism to ensure that "all samples are equally likely," which is quite different from "arbitrary human choice."

This is intended for a classroom setting, where students are asked to write down (to avoid changing during the course of the exercise) a "randomly chosen" number among 1, 2, 3, 4. Since most people choose 3, and most of the rest choose 2, the resulting distribution is quite far from the random uniform distribution.

Nonrandomness of the chosen numbers is demonstrated "on line" by entering numbers (the actual counts for each of 1,2,3,4) into the textbox on the left. The numbers currently shown come from an actual class. Clicking "OK" generates the result on the right, which is a bar graph showing the counts (for easy visual interpretation), together with a text window summarizing the results of some simple statistical analysis, including a confidence interval for the proportion of 2's and 3's. While confidence intervals have likely not been explained at this point in the course, it can simply be said "this range gives a feeling for the variability in the data, and it will contain the given proportion if these numbers are actually random." This provides motivation and

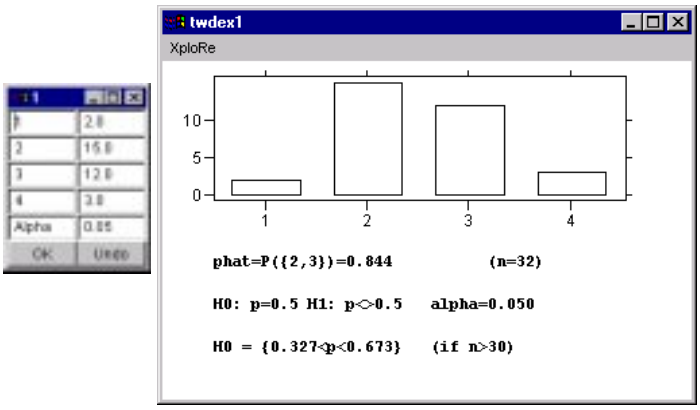


Figure 1: Demonstrates Example 1, "random" is different from "just some." Left: Menu to enter the number of people who have chosen 1, 2, 3 or 4 and the confidence level. Right: Resulting window with graphical and text output, which assesses the amount of "randomness" of the entered numbers.

interest for the time when confidence intervals and hypothesis tests are developed (when this example should be revisited).

For level 2 or level 3 teaching, students should be encouraged to experiment with changing the input values, and watching the change in the interval bracketing 0.5. For example, what happens for (25, 25, 25, 25)? For (0, 0, 100, 0)? What is the difference between (10, 70, 0, 20) and (10, 0, 70, 20)? Students could be challenged to "explore the boundary between random and not" by finding data vectors which are near each other, but give opposite test results. This example may be repeated as many times as desired and may be run directly from the XploRe web site, `www.XploRe-stat.de`. One opens the Java 1.1 interface (Swing classes have to be in the corresponding Java directory), enters `library("tware")` and then enters the quantlet name `twrandomsample()`.

Example 2

This example is intended to illustrate the concept of a p-value for hypothesis testing. For simplicity, it is done in the context of the Binomial(n,p) distribution. The hypothesis tested is: $H_0 : p < c$, for some choice of c.

The example starts with a menu of input boxes, which allows input of:

- The binomial parameter, n (number of Bernoulli trials),
- The binomial parameter, p (probability of success in the Bernoulli trials),
- The observed binomial value, x.

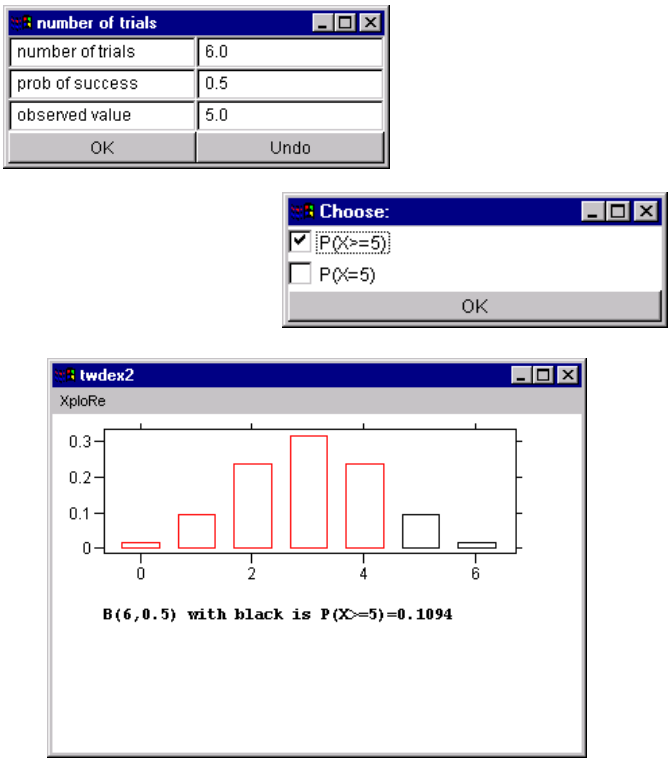


Figure 2: Demonstrates Example 2, how p-values work. Upper: Menu to choose the Binomial distribution parameters: the number of trials, n, the probability of success, p, and the observed value, x. Middle: menu allows choice between $P(X \geq x)$ or $P(X = x)$. Lower resulting plot under Java with area representing the p-value, $P(X \geq 5)$, shown as outline boxes.

The intention is to motivate the p-value, i.e. the "observed significance level" for the observed value x, through graphical display of the region represented by $P(X \geq x)$.

The main graphic is a bar chart, where bar heights show the Binomial(n,p) distribution. The bars corresponding to the event $\{X \geq x\}$ are shown with a black outline, which gives a visual impression for this probability. There is text added to the graph, which gives the numerical value of this probability.

There are two check boxes, which allow choice of the displayed probability as either $P(X \geq x)$ (the usual "p-value") or as $P(X = x)$ (another candidate for "observed significance"). See discussion below about this.

In class it is recommended to demonstrate:

i. When the observed value x becomes larger, the p-value decreases, i.e. the evidence against H_0 becomes stronger.

ii. When p becomes larger, the p-value increases, i.e. the evidence against H_0 becomes weaker. This makes sense, since then the null hypothesis has a better chance of explaining the observed value.

iii. To see why the p-value is $P(X \geq x)$, and not $P(X = x)$, use the checkboxes mentioned above. The parameters $p = .5$, and $x = n/2$ are recommended, and then take several values of n, such as $n = 10, 40, 160$. These show that $P(X = x)$ has the problem that it depends strongly on n, and worse gets small even when there is clearly no strong evidence against H_0. On the other hand, $P(X \geq x)$ is stable for increasing n, and stays large when there is no strong evidence.

This example may be run from the XploRe web site, `www.XploRe-stat.de` directly. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twpvalue()`.

Example 3

This example illustrates two points. First how the normal distribution provides a good approximation to the binomial for large n. Second why it is both important and natural to subtract the mean, and divide by the standard deviation, when doing a normal approximation.

The example starts with an overlay of three theoretical probability histograms (bar graphs where heights are probabilities), representing the Binomial distribution, with a fixed value of p, say $p = 0.6$, and with three choices of n, say $n = 10, 20, 40$, as shown in the left in Figure 3a (bottom). The instructor points out that there is a common "mound shape" to the three graphs, but that they are not close to any fixed distribution, and will not get closer to anything as the sample size n grows, since the probability mass moves to the right. However the effect can be understood, and perhaps adjusted for, by the development of the concept of centerpoint of a probability distribution, e.g. the mean.

When the centerpoint is understood, its effect in the present example is illustrated in the right column of the main graphic. This shows the three theoretical probability histograms of the random variables minus their means. Now it is apparent that mean adjustment overcomes the problem of probability mass moving off to the right, but there is a second problem with the distribution becoming more spread as the sample size grows. Again the effect can be quantitatively understood, and adjusted for, by the development of a concept of spread of a distribution, i.e. the standard deviation.

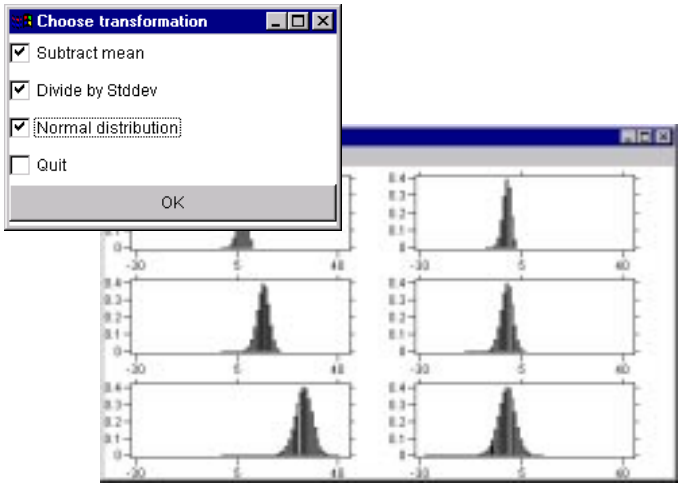


Figure 3a: Demonstrates Example 3, Standardization and Normal approximation of the Binomial. Upper: menus for controlling the transformation of the Binomial distributions. Lower: main graphic window, showing three Binomial distributions in the left column, and the corresponding transformed versions in the right.

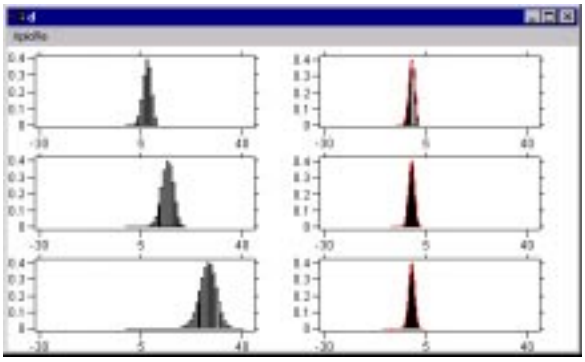


Figure 3b: Shows the effect of adjusting for the scale, on the histograms in the left part of the main output window in Figure 3a. Also shows the effect of overlaying the approximating Normal probability density.

When the spread is understood, its effect in this example is illustrated by checking the "divide by Stddev" box in the control menu. This changes the right column to plots of the probability histograms of the random variables minus their means, divided by their standard deviations as shown in Figure 3b. This shows that the distribution is clearly converging to a common shape. Then the instructor states that with more mathematics, it can be shown that this common distribution is the Gaussian, i.e. normal distribution, which is then overlaid using the "Normal Distribution" checkbox.

This example may be run from the XploRe web site `www.XploRe-stat.de` directly. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twnormalize()`.

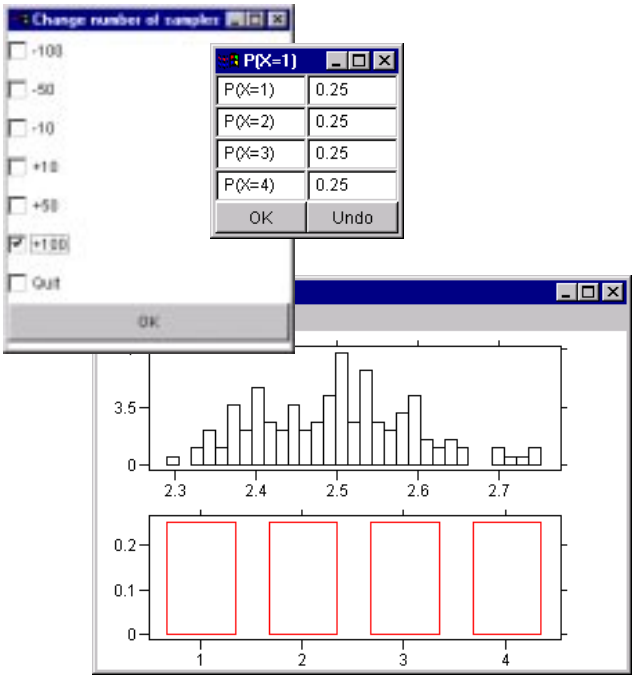


Figure 4: Demonstrates Example 4, Central Limit Theorem. Upper right: menu controlling probabilities of a 4 point distribution. Upper left: menu controlling number of realizations to average. Lower: main graphic window, showing result of repeated convolution, which demonstrates that the distribution of averages converges to the Gaussian.

Example 4

This example shows the main point of the Central Limit Theorem: that averages tend to have a Normal probability distribution, even when the individual underlying distribution is far from Normal. The example starts with a menu containing text boxes (shown in the upper right of Figure 4) for entering an initial discrete probability distribution. This distribution is supported on the integers 1,2,3,4, and after the probabilities are entered a bar graph is displayed (the lower main window shown in Figure 4), showing the probability histogram of the entered distribution. The upper left window controls the number of realizations to average, n.

Clicking OK in the upper left window shows the probability histogram (in the main window) of the average of $X_1, \ldots, X_n$ (computed by simple discrete convolution). This demonstrates how the shape tends towards that of the Normal distribution. Another push button will overlay the approximating normal distribution onto the current probability histogram. For level 2 and level 3 teaching, students could be encouraged to try this with other choices of the underlying probability distribution.

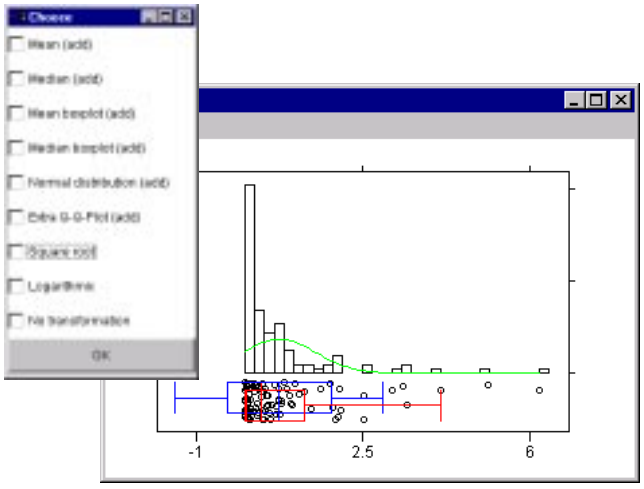


Figure 5: Demonstrates Example 5, Display of 1-d data. Left: Control menu, with checkboxes allowing different displays. Right: Main graphics window, currently showing histogram with overlaid Normal distribution, and jitter plot, with both types of boxplot (mean - standard deviation boxplot, median - quartile boxplot).

They could be challenged to find shapes which give rapid convergence to the Normal, and shapes which give very slow convergence. The student has the possibility to increase and decrease the number of the repetitions of the random drawing. This is designed for discovery of "how" and "when" the Normal limit distribution is a valid approximation as a function of sample size.

This example may be run from the web XploRe site `www.XploRe-stat.de` directly. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twclt()`.

Example 5

This example illustrates the use of visual display devices for one dimensional data. It shows the relationship between the mean and median, shows how transformation can be used to make data have a distribution closer to Normal, and shows the resulting impact of transformations on the mean and the median. Display devices include histograms, jitter plots, and Q-Q (normal probability) plots. Currently considered transformations are the square root, and the logarithm.

The control menu, shown on the upper left of Figure 5, allows choice of which statstical graphics to include. Check boxes will allow the exploration of various notions of "center" and "spread" via overlaid boxplots. The "mean boxplot" is centered at the mean, with the box endpoints showing the mean plus and minus one standard deviation, and with the whiskers showing the mean plus and minus two standard deviations. The "median box plot" is centered at the median, with the box

endpoints showing the quartiles and the whiskers showing the 2.5 and 97.5 percentiles.

The chosen example has substantial skewness which shows that these two boxplots can be quite different, and furthermore that the percentile methods are giving a better notion of "center" and "spread". The square root and the logarithmic (base 10) transformations, show how this situation changes dramatically when the data are transformed. Closeness to normality, in each case, can also be studied via a Q-Q, i.e. normal probability, plot using that checkbox.

This example may be run from the XploRe web site. One opens the Java 1.1 interface, enters `library("tware")` and then enters `tw1d()`.

Example 6

This example gives a visual demonstration of the form of the Pearson correlation coefficient. In particular, it shows why the product moment gives a measure of "dependence," and why it is essential to "normalize," i.e. to subtract means, and divide by standard deviations, to preserve that property.

It uses simulated bivariate Gaussian data, with the number of data points, and the correlation entered through checkboxes as shown at the top of Figure 6. The data are shown with a scatterplot in the main graphics window appearing in the bottom of Figure 6. Text below shows the numerical value of the product moment, $\sum_i (x_i y_i)$, the recentered product moment, $\sum_i ((x_i - \bar{x})(y_i - \bar{y}))$, and the rescaled, recentered product moment, $\sum_i ((x_i - \bar{x})(y_i - \bar{y}))/(s_x s_y)$, which is the Pearson correlation coefficient.

Starting with $N(0,1)$ marginals shows how the ordinary product moment quantifies "dependence," since most values in the first and third quadrants make the product moment positive, and most values in the second and fourth quadrants make the product moment negative, while independence gives cancellation of these effects, so the product moment is essentially zero.

To understand the need for recentering and rescaling, the other menu (shown in the middle of Figure 6) allows changing the center point of the point cloud. When the centerpoint is changed, the point cloud moves accordingly (with the original position shown in gray) and the various moments also updated. The teacher can comment that the original product moment changes dramatically, while the recentered product moment stays the same. Another menu allows changing the scales, and again this change is apparent both visually, and in the product moments, which shows why normalizing by the product of the standard deviations is essential.

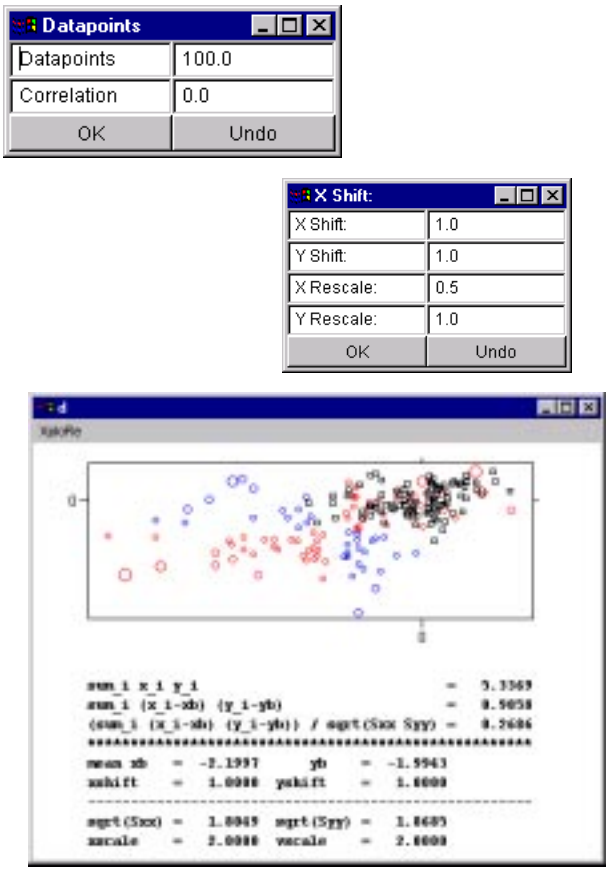


Figure 6: Demonstrates Example 6, Correlation Coefficient. Top: menu for controlling number and correlation of underlying normal data. Middle: menu for demonstrating how shifts and scales affect the product moment, but not the Pearson correlation coefficient.

This example may be run from the XploRe web site `www.XploRe-stat.de` directly. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twpearson()`.

Example 7

This example gives visual insight into how least squares simple linear regression works, and the relationship between the regression of Y on X, X on Y, and total regression.

As for example 6 the data are bivariate Gaussian, and a menu (shown upper right in Figure 7) allows control of the number of data points, and the correlation. Intuitive understanding of least squares fitting is conveyed through interactive manipulation of a candidate fit line. The upper left menu in Figure 7 gives control over this process, through incremental adjustments that are selected by check boxes, followed by a push of the "OK" button. The main graphics window shows the data scatterplot, together with the least squares fit line.

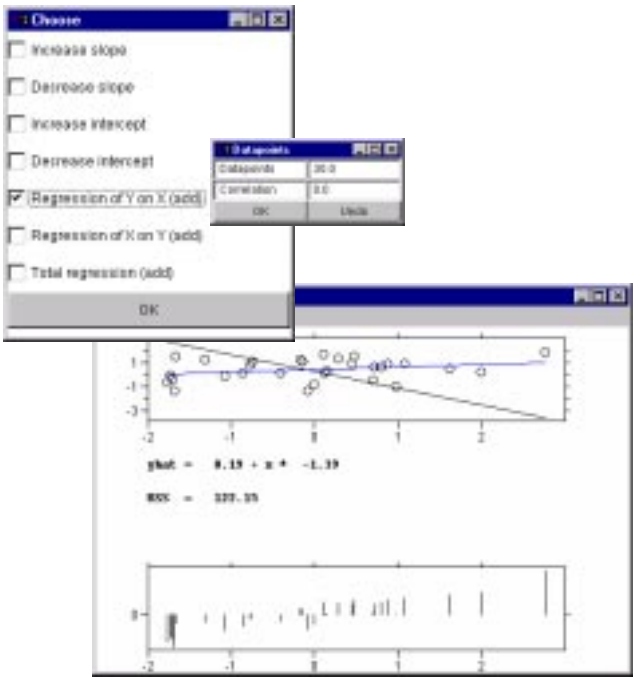


Figure 7: Demonstrates Example 7, Simple Linear Regression. Upper: menus for the changes of the regression line. Lower: main graphic window, showing result of repeated application of changing slope and intercept in comparison with LS line.

A text component shows the equation of the current line (which changes as the line is manipulated), together with the Residual Sum of Squares which gives a numerical summary of the goodness of fit. Very effective visual indication of what RSS means comes from the lower graphics part of this window, which represents the residuals as vertical lines. When the fit is poor (and hence the RSS is large), the residual plot shows why, and give a clear indication of how the line should be moved to improve the quality of the fit to the data.

Additional check boxes allow understanding the variations of regression of X on Y, and total variation, and result in appropriate shifts of the graphics. This example could be modified to allow other types of fitting, such as least L_1, or other types of robust fits.

This example may be run from the XploRe web site `www.XploRe-stat.de` directly. One opens the Java 1.1 interface, enters `library("tware")` and then enters `twlinreg()`.

Acknowledgements

We would like to thank Nathan Derby, Marlene Müller and Bernd Rönz for helpful suggestions and corrections. The paper was financially supported by the Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse," Deutsche Forschungsgemeinschaft. The research of J. S. Marron was supported in part by NSF grant DMS-9504414.

References

Classroom ready examples in XploRe
`ise.wiwi.hu-berlin.de/statistik/lehrmaterial/statmat.html`

GLM tutorial
`www.xplore-stat.de/tutorials/glmstart.html`

GAUSS software
`www.aptech.com`

GAUSS programming for Econometricians
`eclab.econ.pdx.edu/gpe/`

Help system pages
`www.xplore-stat.de/help/_Xpl_Start.html`

Image processing with Java
`www.utdallas.edu/~degroat/javadip/JavaDIP.html`

MD*Tech - Method and Data Technologies
`www.mdtech.de`

Non- and Semiparametric Modeling course text
`www.quantlet.de/~scripts/scripts/spm/spm.html`

SAS software
`www.sas.com`

SHAZAM software
`shazam.econ.ubc.ca`

Splus software
`www.mathsoft.com/Splus`

Stata software
`www.stata.com`

STATLIB server of SPLUS
`lib.stat.cmu.edu/S/`

SticiGui(c) Java Tools
`www.stat.berkeley.edu/users/stark/Java`

SUN's Java Development Kit (JDK)
`www.javasoft.com`

Support Vector Machine
svm.dcs.rhbnc.ac.uk/pagesnew/1D-Reg.shtml

Symanzik (1998)
www.galaxy.gmu.edu/~symanzik

Virtual Stat Lab
www.ruf.rice.edu/~lane/stat_sim/index.html

wavelet book in PDF format
www.quantlet.de/~scripts/scripts/
wav/wavpdf.pdf

Webstat Project
www.stat.sc.edu/~west/webstat/

XLISP-STAT
www.cern.ch/WebMaker/examples/xlisp/
www/cldoc_1.html

XploRe
www.XploRe-stat.de

Wolfgang Härdle
Institut für Statistik und Ökonometrie
Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin
stat@wiwi.hu-berlin.de

Sigbert Klinke
Institut für Statistik und Ökonometrie
Wirtschaftswissenschaftliche Fakultät
Humboldt-Universität zu Berlin
sigbert@wiwi.hu-berlin.de

J. S. Marron
Department of Statistics
University of North Carolina
marron@stat.unc.edu

TOPICS IN INFORMATION VISUALIZATION

Linked Data Views

By Graham Wills

Introduction

I think of a "data view" very generally as anything that gives the user a way of looking at data so as to gain insight and understanding. A data view is usually thought of as a bar chart, scatterplot, or other graphical tool, but I use the term to include a display of the results of a regression analysis, a neural net prediction or a set of descriptive statistics. In a simple case, a scroll bar is a view of a document, linked to a textual representation beside it. Selecting an area in the scroll bar using the thumb links to the associated text view to display new textual information. In general, a data view is a representation the user can look at and study to help understand relationships and determine features of interest in the data they are studying. Typically there are parameters or variations in the method of display so that some way of interacting with the view to modify its behavior is necessary.

Also typical is the desire to explain something of interest found in a view. Do data form two clusters under this particular projection of the grand tour? Is there a change in the relationship between salary and years playing baseball when the latter is greater than five years? When we see something interesting, we want to explain it, usually by considering other data views or by including additional variables. With some types of view, it is not hard to add in variables and see if they can explain the feature, or indeed if they have any effect whatsoever. In a regression analysis, you can just simply add a variable to the set of explanatory variables (taking due care with respect to multicollinearity and other confounding factors). If a histogram of X shows something of interest, you can "add" a variable Y to it by making a scatterplot of X against Y. If you want to explain something in a scatterplot, then it is possible to turn it into a rotating point cloud in 3D, and using projection pursuit or grand tour techniques, you can go to still higher dimensions.

Despite the undoubted utility of this approach, it does present some problems that prevent it from being a complete solution. The main ones are:

- As plots become increasingly complex, they become harder to interpret. Few people have problems with most one-dimensional plots. Scatterplots, tables and grouped boxplots or other displays involving two dimensions are easily learnable. But the necessity of spinning and navigating

a 3D point cloud, or understanding the contributions to a mutivariate projection make these views less intuitive.

- It is harder to accommodate differences in the basic types of the data. High-dimensional projection techniques assume the variables are rational, as do techniques that display multivariate glyphs and, to a large extent, parallel axes techniques. Given a table of two categorical variables, adding a rational variable requires changing to quite a different type of view, such as a trellis display.
- Data that are of a type specific to a particular domain can be impossible to add directly. Exploring relationships in multivariate data collected at geographical locations, on nodes of a graph, or on parts of a text document is very hard because of the difficulty of building views that correlate the statistical element and the structural element of the data. Often, two completely different packages are used for the analysis, with results from one package mangled to fit the input form of the other package - a frustrating task.

A good solution to these problems is the linked data views paradigm. The idea is fairly simple; instead of creating one complex view, create several simpler views and link them together so that when the user interacts with one view (for example, to indicate a feature of interest), the other views will update and show the results of such an interaction. This allows the user to use views that require less interpretation and views that are directly aimed at particular combinations of data. It also allows the easy integration of domain-specific views; views of networks or maps can easily be linked to more general-purpose views.

I do not mean to argue that the linked data views is a uniformly superior method to that of monolithic complex views mentioned above. That is not the case, as there are examples where a single multivariate technique is necessary to see a given feature, and multiple simpler views simply won't do. However, for many problems, especially those where conditional distributions are of interest, the linked data views technique works extremely effectively.

Starting with Scatterplots

One of the earliest linked views work to achieve wide attention was the scatterplot brushing technique of Becker, Cleveland and Wilks (1987). By arranging scatterplots of n variables in a table so that all the $n(n-1)$ ordered combinations of axes are present, the eye can quickly scan a row or column and see how a given variable depends on each other variable. This useful arrangement technique is enhanced by the use of a brush. In this context, a brush is a shape that is dragged around the view by the user, and performs some operation on the graphical elements it passes over. In typical scatterplot brushing tools, the data points brushed over are painted in a different color, both in the panel in which the brush is active, and in all other panels of the window. In our terminology, the brush is the mechanism that links the scatterplot data views.

One of the reasons this technique is so effective is that in each linked view, there is a one-to-one correspondence between cases of the data matrix and graphical representations of these cases, so that in each scatterplot we have complete freedom as to what color or glyph to use to represent this data item. Intuitively, it is easy for us to think of the 'red, square' item, and locate it in each view. The conceptual model (which can easily be the internal data structure, too!) is of adding a few extra columns to the data matrix to represent color, glyph, and visibility and using the brushing technique to modify the values in these columns. The table below shows such a model:

$V1$	$V2$	$V3$	Visible	Color	Glyph
Cork	22.3	5	0	green	circle
Dublin	12.3	5	1	green	circle
Kerry	18.8	6	1	red	circle

Table 1. Sample data $(V1, V2, V3)$, with added variables representing graphical information for display purposes

To manage a brush over a data view, the program must calculate what cases are under the brush and manipulate the values of one or more of the additional graphical variables for each such case. Each linked view must then update to reflect the changes.

Even when restricted to data views that display graphical elements for each case, this is a powerful tool. An example of a successful tool in this area is XGobi (Swayne, Cook and Buja, 1998). XGobi is a X-Windows based tool that presents the user with several glyph-based views (dotplots, scatterplots, rotating plots, grand tour and projection pursuit tours), and uses brushing to link the views along the above lines.

Generalizing the Implementation

The above approach runs into problems with more than small amounts of data. If you have tens of thousands of points, often you want to look at views that aggregate the data, such as bar charts, histograms and fre-

quency tables. Linking them would be useful. Also useful would be a more general method of linking; perhaps we want to link cases with ones in another data matrix, using a user-defined linking function. And, referring back to the opening paragraph, it would be very helpful to be able to link graphical plots of data to models of the data.

The statistical analysis package DataDesk (Velleman, 1997) was originally built as a teaching tool, but is now a full-featured statistical package that has linked views designed in at the core. Brushing works with aggregated views such as bar charts and histograms as well as within unaggregated views, and the outputs of analyses such as regression and correlation analysis can be visualized and are linked to the other views. If a user spots some unusual cases in a view of residuals from an analysis, they can brush those points, see if there is anything that might explain it in other variables, modify the model and instantly see the residual view update to reflect the new model. Figure 1 shows an example of such linking.

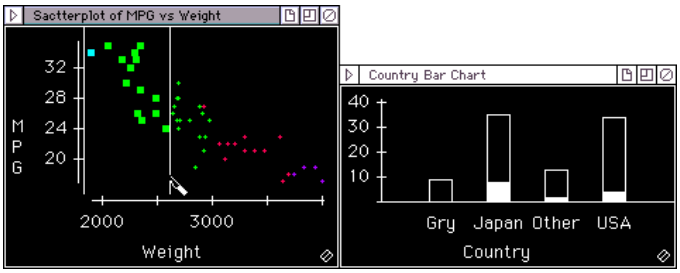


Figure 1. Linked views in Datadesk. Points selected in a scatterplot of Miles per Gallon vs. Weight are highlighted in the bar chart of Country. Selecting the points in the low end of the weight scale shows which country makes the lightest cars.

There are some design decisions that have to be made when linking aggregated views. For example, suppose we wish to link a bar chart and a scatterplot as in Figure 1. Looking at Table 1, how can we represent the Visible attribute consistently in both? Not too hard, items with zero visibility do not appear in the scatterplot and are ignored when creating the bar chart. The Glyph attribute is also simple - bar charts cannot use glyphs and that attribute must be ignored. Color, however, may be dealt with in several ways. Ignoring the coloring by mapping all the colors to a single neutral color is the method used by DataDesk. Another method, used by EDV (Wills, 1997), is to color portions of the bar in proportion to the number of cases in the bar with each color. This is achieved by dividing the bar up into segments, one for each defined color, with the size of these segments proportional to the number of cases in that bar with the given color. The overall affect is to produce a dynamically changing stacked bar chart.

An alternative method is used by LispStat (Tierney, 1990), in which each data item is assigned its own place in the bar and that section of the bar is colored appropriately. This solution is very close to the 'one-to-one' relationship method in the previous section, as each bar is really a set of stacked rectangles, one for each case. Both drawing and brushing over the bars is handled as if the bars were a collection of separate little boxes. Figure 2 shows the difference between the three approaches for some sample data.

One of the more powerful novel features of LispStat is due to its implementation in Lisp - as an interpreted language, the user is free to write any function that can link views together, and indeed can write any data view they wish. If you want interactively to color items in a scatterplot based on their distance from the center of the brush, it is an easy job ... as long as you know Lisp.

MANET (Unwin et al., 1996) is a relatively new environment for exploratory data analysis. MANET is an acronym standing for "Missings Are Now Equally Treated," and this describes an important novel feature of the system; the ability to display information about missing values in views in which they would otherwise appear, and to integrate this information naturally within the view. Thus in a scatterplot of X against Y, cases with missing X values are plotted as a dotplot on the Y axis at locations corresponding to their Y values. This enables the analyst to check for relationships between missing values of X and values of Y. One point to be made here is that the analyst need not do anything special to see if such a pattern exists; it is presented to them as a routine part of the exploration; they are "equally treated."

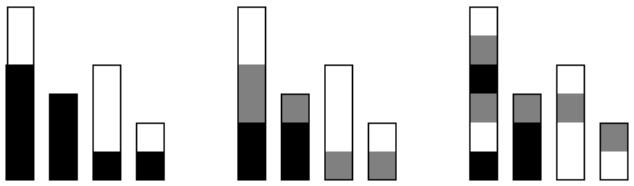


Figure 2. Three methods of linking cases in an aggregated view. Each view shows 4 selected black cases, 4 selected grey cases and 6 unselected cases. In the left bar chart, color is ignored; only the selection state is used. In the middle view, colors are stacked. In the right view, a portion of a bar is allocated to each case, and that portion is colored with the appropriate color.

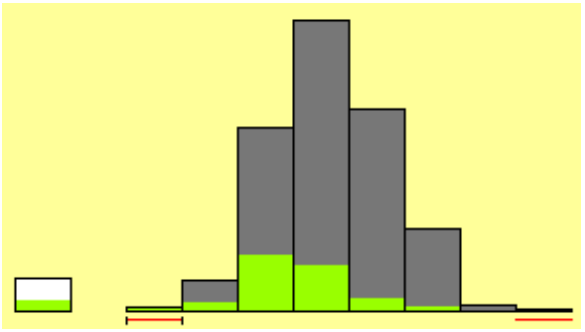


Figure 3. A histogram in MANET. On the left is a bar representing missing values, and under the extremes of the rest of the histogram, lines are displayed to indicate that screen resolution may be causing a misleading display.

Another novel feature of MANET is a technique for dealing with a common problem when analyzing large data sets; screen resolution. A lot of attention has been focused on overplotting for scatterplots and similar views, but MANET is unique in that it addresses underplotting. A very common situation is in drawing bars of a histogram or barchart when either an entire bar or the selected section of it has a logical height of a fraction of a pixel. Especially vexing is the case when a bar has a height of one pixel, but only some of its data cases are selected. Both alternatives - drawing the selection and so filling the bar, and drawing nothing at all - give the false and possibly dangerously misleading impression that the bar contains only unselected or only selected items. MANET helps the analyst avoid drawing such a conclusion by drawing a red line under the bar to indicate the presence of such a condition. Figure 3 shows an example of this technique.

Specializing the Implementation

Moving in a somewhat opposite direction is research aimed at building views and systems for specific domains. For spatial data, Unwin and Wills (Haslett et al., 1990) built a system that combined a number of statistical views with geographical views. REGARD allowed the user to manipulate maps of geographical information, containing layers of data representing sets of geographical information (towns, rivers, countries, etc.). These different layers contain entities with statistical data, so that user can create data views on one or more of these variables, and use the linking system to tie the views together. The interesting part of the tool, from the point of view of this article, is in the linking between geographical layers. A common scenario is the following: A user has been exploring pollution levels in streams and has selected a group of heavily polluted stream segments. They now want to see the result of that selection in a layer containing regions with population data. In a typical Geographic Information System, the analyst would take the identified stream segments and expand around them to create a set of regions. Then the population regions that intersect this stream buffer region can be identified. In REGARD, this process was generalized by using geographical distance to link layers. This encompasses the case above as well as other common examples like inclusion of points in regions, intersection of regions and points in one layer being selected if close to selected points in another layer.

REGARD also pioneered linking in networks which was further developed in NicheWorks (Wills, 1999). Here the data consist of information on nodes and links of a graph, and the goal is to use the linking mechanism to facilitate exploration of relationships between the two sets of data, as in Figure 4. Instead of geographical distance, the concept of distance along the graph becomes the essential factor and the number of useful linking operations is quite large. Depending on the application, selecting a graph node might lead to selecting edges originating from and/or terminating at the node, nodes down or upstream from the node, all nodes and edges in a connected or doubly-connected component containing the node, nodes and edges n steps away, and so on.

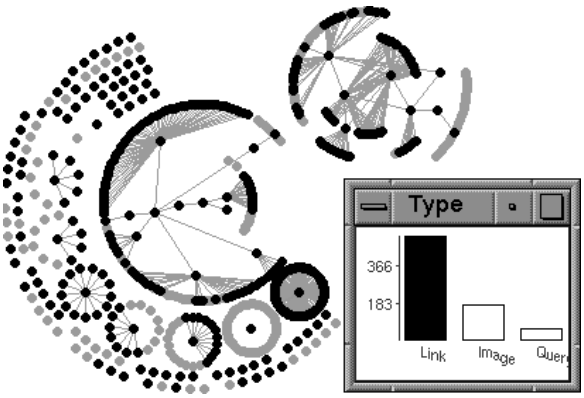


Figure 4. A graph with two components representing sets of URLs (web locations) and their interconnections, linked to a bar chart describing the type of URL. The largest bar, representing “normal” pages (not images or queries/scripts) has been selected.

In a rather different area, strategic computer games are featuring an increasing amount of data view linking. One of the earliest and best implementations is SimCity (Maxis Corporation, 1989), where multiple views of a simulated city continuously change in response to user interactions (and the occasional monster rampage or UFO invasion). SimCity and its descendents feature

a complex model, with many variables that are both outputs and, at the next time step, inputs to the model. As a linked views system with millions of users and several versions, it is an excellent example of using simple linked views to understand a complex model and data set. I like the knowledge that my many attempts to build a perfect metropolis have not just been entertaining.

Onwards

There are several existing linked data views environments, each of which has its own contribution to research into visual exploration of data. Whenever I see a new view or technique that I like the look of, I think how it might be added into the environment I use. It's rare that a view cannot be modified to work with others, and the benefits are large; a new view need only focus on what it does best - its own neat or novel feature, relying on the existing views in the system to do their job. The whole is then greater than the sum of the parts.

In the reference and resources sections below, I have not tried to cite the seminal papers or earliest occurrences; I have instead endeavored to look for references that provide a good overview and introduction of the authors' variation on this powerful and important technique. The linked views paradigm is a vital and expanding part of statistical graphics research, and I have no doubt that I've missed exciting new developments and novel techniques in this brief survey. Drop me a line and let me know about them!

Web Resources

Lisp-Stat program and documentation
`stat.umn.edu/~luke/xls/xlsinfo`

A base page from which to access an introduction to the EDV environment and the NicheWorks graph tool
`www.bell-labs.com/~gwills`

Home for both DataDesk and ActivStats (a teaching tool featuring DataDesk for analysis)
`www.datadesk.com`

Much of the original work on scatterplot brushing was done in the S environment. This is a link to the current commercial version, S-plus.
`www.mathsoft.com/splus`

XGobi program and XGVis program (a version of XGobi for graphs and multi-dimensional scaling)
`www.research.att.com/~andreas/xgobi/`

Maxis' site for SimCity
`www.simcity.com/home.shtml`

A guide to the MANET system
`www1.math.uni-augsburg.de/Manet/`

References

Becker, R.A., Cleveland, W.S. and Wilks, A.R. (1987) "Dynamic Graphics for Data Analysis," *Statistical Science* **2** pp. 355-395.

Haslett, J., Wills, G. and Unwin, A. (1990) "SPIDER - An Interactive Statistical Tool for the Analysis of Spatial Data," *Int. Journal of Geographical Information Systems*, **4**, #3, pp. 285-296.

Maxis Corporation (1989) "SimCity [computer program]," Walnut Creek, California.

Swayne, D. F., Cook, D. and Buja, A. (1998) "XGobi: Interactive Dynamic Data Visualization in the X-Window System," *Journal of Computational and Graphical Statistics*, 7(1) 1998.

Tierney (1990) *Lisp Stat: An Object Oriented Environment for Statistical Computing and Dynamic Graphics*, Wiley.

Unwin, A., Hawkins, G., Hofmann, H., and Siegl, B. (1996) "Interactive Graphics for Data Sets with Missing Values - MANET," *Journal of Computational and Graphical Statistics* 5(2) pp. 113-122.

Velleman, P.F. (1997) *Learning Data Analysis with DataDesk, Student Version 5.0.* Addison Wesley.

Wills (1997) "How to Say 'This is Interesting'," *Proceedings of Section of Stat. Graphics 1997*, pp.25-31, American Statistical Association.

Wills, G. (1999) "NicheWorks - Interactive Visualization of Very Large Graphs," *Journal of Computational and Graphical Statistics*, June 1999.

Graham Wills
Bell Laboratories
`gwills@research.bell-labs.com`

Using Layering and Perceptual Grouping In Statistical Graphics

By Dan Carr and Ru Sun

1. Introduction

This paper concerns the use of layering and perceptual grouping to modify the appearance of statistical graphics. Our motivating example is a Trellis Graphics dot plot with strip labels as shown in Figure 1 (see page 29). The object of our attention is the color-filled strip label boxes. Our subjective opinion is that the color-filled boxes are visually dominant, drawing more visual attention than the dot plot panels. Our impression is of looking through the trellis of strip labels to see the dot plots. We conjecture that the intervening foreground strip labels slow visual comparison of dots from different panels in the same column. While many chose the simple solution of turning off the strip label fill, our redesign goal is to move the dot plot panels into the foreground. We do this conceptually by putting dot plots on a thin marble block. This leaves the strip labels back in flatland. While examples similar to Figure 1 provided our motivation, the methods described in our redesign are applicable to a wide class of statistical graphics.

The notions of figure and ground, foreground and background, and layering are not new. Examples of figure and ground reversal have long intrigued lay people as well as researchers in the field of human perception and cognition. Similarly, researchers in cartography have studied the closely related notion of information layering on maps (for discussions of both topics see MacEachren 1995). The notion of map layers is a basic part of geographic information systems. Tufte (1990) devotes a full chapter to layering and separation. Our first goal is to promote the use of existing layering methods to enhance the appearance of statistical graphics.

We live in a competitive world in which appearances are important. Appearance influences our choices in all facets of our visual life from the food we eat, to the clothes we wear, to the partners we choose. Appearance is not necessarily our strongest criterion. When hungry most of us would choose a spotted banana over a flower despite our notion of what is pretty. Nonetheless at some point of other things being roughly equal, appearance can be the deciding criterion. Our opinion is that statistical community could pay more attention to appearance of graphics. With appearances being roughly equal perhaps the public will choose the graphics with content.

Addressing the appearance of graphics is a bit risky in a profession that strongly focuses on content and perceptual accuracy. For example Cleveland (1993a) presents graphical methods in the context of data examples. This data centric approach is persuasive. He is interested in science and his methods continue to build upon the foundations of perceptual accuracy of extraction promoted by Cleveland and McGill (1984).

Our promotion of methods for the sake of appearance can be viewed with suspicion. We are promoting a three-dimensional appearance and this may seem to run counter to some sound guidance. For example Tufte (1983) disparages area and volume encodings that correspond to the square and cube of variable respectively, and discusses the lie factor. He also discourages display of extra dimensions even when they are constant. The phrase “dimensional puffery” now appears in the human computer interface literature (Card, MacKinlay, and Shneiderman 1999). There is no doubt that frivolous embellishment can detract.

We promote the notion that some embellishment can give the appearance of value added at little perceptual accuracy cost. In particular this paper suggests use of modest 3-D effects to layer information. Further, we consider “perceptual enhancements” that seem inconsistent with common guidance about graphics design. We recommend use of perceptual grouping for its own sake. This runs the risk of people thinking that the groups mean something more. We also connect points to guide the visual flow from point to point and to strengthen the perceptual grouping. Critics will claim that some readers (not themselves) interpret the presence of connection lines as implying linear interpolation, however nonsensical. Nonetheless, this paper suggests that the benefits of layering, perceptual grouping and guidance of visual flow may well outweigh the cost of confusion that can arise until new conventions are established.

The structure of the paper is as follows. Section 2 provides a little background on dot plots, Trellis graphics and strip labels. Section 3 concerns the use of multiple perspectives in plot construction. Section 4 touches on diffuse lighting, shadows, and the appearance of depth. Section 5 revisits labeling issues. Section 6 provides a brief discussion of perceptual grouping and connecting lines. Section 7 concludes by providing links to software and thoughts spawned by Figure 2 (see page 30).

2. Dot plots, Trellis Graphics, and Layering of Strip Labels

Consider dot plots. Every time I (Dan) return to the work of Cleveland and McGill (1984) or Cleveland (1985), I am impressed by the simple elegance of the dot plot concept and the numerous design variations that such simplicity facilitates. My work with federal agencies started with an attempt to promote dot plots. So far most of these efforts have failed. Other than the row-labeled dot plots that I promoted for use in the Atlas of United States Mortality (Pickle et al. 1997), I have yet to see a dot plot in a major federal publication. Perhaps the biggest reason is that dot plots are not a push button option in widely used spreadsheet graphics.

The Trellis graphics in Splus represent a major step forward from the 1970's business graphics that are characteristic of shreadsheet software. The color versions can be quite attractive while conveying information with integrity. The Trellis design provides a general approach to multiway, multipanel and even multipage graphics. This generality necessitated the development of a general approach to labeling. Strip labels emerged as a elegant solution to multiway panel labeling challenge. However, our opinion is that this general labeling solution is not necessarily optimal for special cases.

As indicated in the introduction we think that strip labels between graphics panels impede comparison down the panels within the same column. For this reason, we often prefer some of Cleveland's (1985) earlier dot plot designs that keep the labeling in the margins. Our redesign goal here, however, is to retain the panel strip labels (at least in some form), while attempting to promote comparisons across vertical panels by bringing the panels into the foreground. This is a task in relayering information.

There are different ways to go about relayering information. We conjecture that the strip labels appear in the foreground because the color fill calls attention to areas. It seems that area symbols dominate over line and point symbols. One relayering option simply avoids the color fill. Another option makes the dots even bigger so they effectively become area symbols that compete on an equal footing with strip labels. We think the strip labels provide a visual barrier to vertical visual flow even though they appear in the same visual layer as the dots. Thus we strive to reduce the impact of this visual barrier by using 3-D effects that bring the graphics panels into the foreground.

3. The Marble Block and Perspective Views

As indicated above and shown in Figure 2 (page 30), our simple idea for bringing the graphics panels into the foreground is to place them on thin marble blocks. Simple perspective views that include the block surface and sides provide a 3-D appearance. The details about perspective views can be found in numerous sources such as Foley, van Dam, Feiner, and Hughers (1990), Carr(1994), and Blinn (1996). We want to show the marble blocks in a way that minimizes the area that is consumed by the 3-D construction. Toward this end we digress for a moment.

It is fun to play with perspective views. The impossible tribar of Penrose (1958) and the artwork of M. C. Escher (for example see Ernst 1976) have delighted people around the world. Humans have an amazing ability to maintain a local perspective that is not globally consistent. For many situations local is good enough. For example the construction of the "infinity enlarged" stereo pairs projection (see Carr 1994) effectively moves the center of projection (in this case the mid- point between the eyes) in front of every single point in the plot before projecting the point onto the surface. This provides the key to a space saving view of our marble block.

To save space we position the viewpoint so we see only one block side. As shown in Figure 2, our viewpoint choice is below the center of each block and beyond the block boundary so we see only the lower block side. Maintaining the same relative viewpoint for each block provides the basis for constructing Figure 2. This raises the question of what to do when panels touch left and right. Currently, we move the viewpoint again so the shared edge is right in front of us.

Figure 2 shows our hopping about perspective construction of juxtaposed panels. Some previous examples (see Cleveland 1993b and discussant comments) show the two years of data superimposed on the same panels. For comparing barley yield across years we also prefer superposition to juxtaposition and note that the direct display of differences is helpful. Figure 2 has not been optimized for comparing years.

4. Diffuse lighting, shadows, and the appearance of depth

In Figure 2 we are dealing with a solid block and not a wirefame. Thus we need to pay attention to surface rendering. An actual marble texture would be distracting so we chose to use a plain gray surface. Given the surface, the next step involves the use of lighting models. Lighting can create or enhance the appearance of depth and provide a sense value added relative to flat graphics.

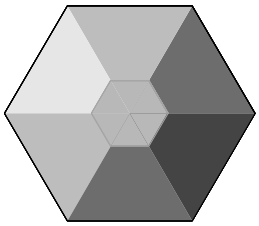


Figure 3 Exploring ambient light and diffuse reflection.

Computer graphic classics (Foley, van Dam, Feiner, and Hughers1990; and Blinn 1996) describe illumination models and surface rendering. Those interested in rigorous treatment should refer to such sources. In brief, there are four basic facets to illumination models, ambient light, diffuse reflection, atmospheric attenuation, and specular reflection. Ambient light is the minimal light emitted by surfaces where the intensity does not depend on a specific light source. Phong shading (normal vector interpolated shading) and specular reflection are useful in producing 3-D appearance. These have been used effectively in density representations (see Wegman, Carr and Luo 1993). Here we focus on the use of diffuse reflection and shadows.

Given the reflective properties of a surface, the relative intensity of diffuse reflection from a point light source shining on a planar surface depends upon the angle between a line from the light source and the normal to surface. The light intensity (lightness, value or brightness depending on the literature) decreases with the cosine of this angle. In contrast to specular reflection, the diffuse light intensity does not vary with viewer's position (except when a surface becomes hidden). Figure 3 provides a simple illustration of ambient light and diffuse reflection. The lighting can make the figure appear as a truncated 3-D hexagon-based pyramid.

Where should the light source be positioned? The convention established by walking under the sun is that light comes from above. The computer graphics convention for simple single-light- source cases is that the light comes from the upper left. Figure 3 should appear to poke out of the surface toward the reader. Putting the light source at the lower right should make the pyramid poke into the surface. However, our visual system can do interesting things. The perception of directionality, above or below the surface, is fragile, and some readers experience the opposite of what is expected.

The construction for Figure 3 provides a nice review of vector operations and is suitable as a student exercise. Suppose we have a right handed coordinate system with the positive z-axis pointing toward the viewer. Suppose the center point of the polyhedron is (0,0,.7) and the vertices are located at $(\cos(ai), \sin(ai))$, where $ai = 0, 60, 120, 180, 240,$ and 300 degrees. The consecutive vertices of a triangular face can define the vectors corresponding to two edges. The cross product of these vectors defined the normal to the face. The dot product of the unit length normal and a unit direction vector (say [-.544,.314,.778]), from a light source at infinity gives the relative diffuse intensity (di) for each triangular surface. Letting the gray level on a 0 to 1 scale be something like $.2 + .7 * di$ provides plausible shading. This can be viewed as a mixture of ambient light and diffuse reflection. The ambient light provides a minimum value for the surfaces. A rough interpretation is that the ambient intensity is $.1 + .8 * \min(di)$. The other gray levels are lighter due to diffuse reflection from the light source. We have avoided using pure white by choosing .7 rather than .8. This leaves a little room for specular highlights.

Figure 2 (page 30) conveys depth by 3-D perspective. The diffuse lighting comes into play when we color the quadrilateral below each panel. We do not bother with a precise diffuse reflection calculation but simply use a darker shade of gray.

Similarly we attempt to provide the appearance of grid lines etched into the surface by pairing the light and dark lines. The side of a groove away from the light (right side for vertical lines and bottom side for horizontal lines) should appear lighter since those surface catch more light. Ideally, a 3-D design should construct thin and shallow "v" shaped trenches and use proper diffuse lighting. The dots are more important than the reference lines, so the construction with etched lines is proper in terms of depth. Care must be taken in terms of line width and contrast so the lines are less salient than dots.

We choose to use red dots because the focal length for red makes the dots appear closer. Shadows for the dots provide another way to convey depth and make the dots appear closer. As Blinn (1996) indicates, simple shadow construction is often suffices to convey depth. Given that the light source is in the upper left shadows of an object appear to the lower right of the object. Drawing a dark gray image to the lower right before drawing the object provides the appearance of depth. The dots in Figure 3 illustrate this.

Lifting the dots off surface in Figure 3 causes some ambiguity. Should the reader judge the center of the dot or the center of its shadow when determining its value? The description here is explicit: the position of

the shadow is altered. However, without an established convention the viewer may reason that shadow on the surface should be judged against grid lines. If shadows are used, the shift should be slight. Our statistical Puritanism may incline us not to use shadows, but we confess that a little bit of shadow is fun.

Note that web page design makes heavy use of lighting models to produce modest 3-D effects. Boxes routinely have light edges on the left and top and dark edges at the right and bottom. Such effects are sometimes applied with little thought. In particular standard tabling methods put 3-D ridges around numbers. This impedes visual flow and deters comparison of numbers. The choice of ridges is consistent with drill-down mentality. We are supposed to be happy just to have found a number. The notion that graphics (include tabular graphics) is about comparison has not yet been reinvented some communities.

5. Labels Revisited

Labeling is perhaps the hardest of design challenges. The labels for barley variety in Figure 2 (page 30) posed a difficult challenge until we figured out that they had to go on the marble block to keep aligned with dots. This leaves the strip labels back in flatland by themselves. This is not so bad since strip labels refer to the whole panel below rather than to specific dots. Bump-mapped text could put the strip labels on the same visual plane as the dots. From our point of view the strip labels deserve their fate for getting between our precious graphics panels.

In Figure 2 we do not show the factor levels, and that appear as dark shaded rectangles in Figure 1 (page 29). These shaded rectangles complicate the reading of text. Our visual system sees lines at the edges of regions with contrasting lightness (see Friedhoff and Benson 1991). When text overplots such edges it is like plotting a character on a line. Adjusting to a background of changing contrast when reading is also extra work. A possible solution, when there is space, is to plot little dots near the edge of the strip frame. For example plotting one dot could indicate level 1, two dots level two, and so on. Perceptual grouping in units of five would aid in fast perception for numbers over five. Printing the levels as text might be considered but may get confused with the labels. We do not consider the numeric level information important in Figure 1 so do not show it.

Cleveland (1993a) discusses on multivariate sorting by medians to establish the plotting order for barley varieties, sites and years. Sorting is important to simplify appearance and facilitate comparison. Figure 1 is similar to a figure that appears in a copyrighted user manual (Mathsoft 1995). In the effort to avoid the appearance of producing an exact copy of such there are some changes. The obvious one is putting 1931 to the left of 1932. In this particular case we lean toward following the left to right reading convention rather to maintain full consistency in the sorting of factors.

6. Perceptual Grouping

The paper also presents a second facet of graph redesign, perceptual grouping. Having seen the merits of perceptual grouping illustrated by Kosslyn (1994), it is hard not to promote wider usage of various forms of perceptual grouping in statistical graphics. A few examples now appear in newsletter articles (Carr and Pierson 1996, Carr and Olsen 1996, and Carr et al. 1998).

One direct application of perceptual grouping is in table-lookup process that Cleveland (1993b) calls matching. Consider the visual task of matching labels and points. Figure 1 uses lines across the panels to connect labels and dots. This makes the plot look busy. Further horizontal lines inhibit vertical visual flow. The line linking works to the extent that one has the patience to track a line all the way across the page. Figure 2 groups the labels and dots in a 3-4-3 pattern. It is trivial to pick out the middle label in the top group of three and the middle dot in the top group of three on the right panel. We suggest that for remote points this label to point matching is easier than individual line tracking. It is also trivial to pick out the middle group or the bottom group on each panel and match the respective labels and points within each group.

There is a possible drawback to the perception grouping of barley varieties in Figure 2. Readers may look for some other meaning to the grouping than simple perceptual grouping. This is in part a matter of convention. If perceptual grouping were routine, the default interpretation for unlabeled groupings would suffice.

The design of Figure 2 strengthens the perceptual grouping by connecting points in a group with lines. This makes it relatively easy to compare the middle four points in the top panel with the middle four points in the fourth panel below. One can easily focus attention of the third points in these two sets of four points to make comparison. We think it is harder to find, remember and compare the two 6th points using Figure 1 with all symbols. The connecting lines help to guide the visual flow down the page. Each line leads the reader to the next point. Searching to find the next point is necessary only for disconnected points. Thus connecting points serves a worthy purpose.

Figure 1: A Trellis Graphics dot plot with strip labels.

There are at least two drawbacks to the connecting lines. First, treatment of points is not uniform. Some consecutive points are connected and others are not. Connecting all the points in a panel solve this problem. However, experience with such examples (see Wilkinson 1993 and Carr 1994) suggests that ten items provide too complicated a shape to store in short term memory. It is not a small perceptual group. The unequal treatment becomes a price that we are willing to pay for the advantage of rapid comparison.

The second drawback is that there seems to be a graphic convention to avoid connecting points unless linear interpolation is implied. This convention is routinely violated in time series line plots. The designers of such time series graphic do not typically plan to defend the idea of precise linear interpolated between values even if such values are plausible. In Figure 2 we connect consecutive points based on the multivariate-sorting-established rank order. One could interpret the points along the connecting lines as being interpolated values for mixing barley varieties, but this is not what we intend. We want to guide the visual flow down the page. We are willing to pay the price of confusing those who might spatially interpolate population values between New York and Paris because the rank ordered population dots are connected on a graph.

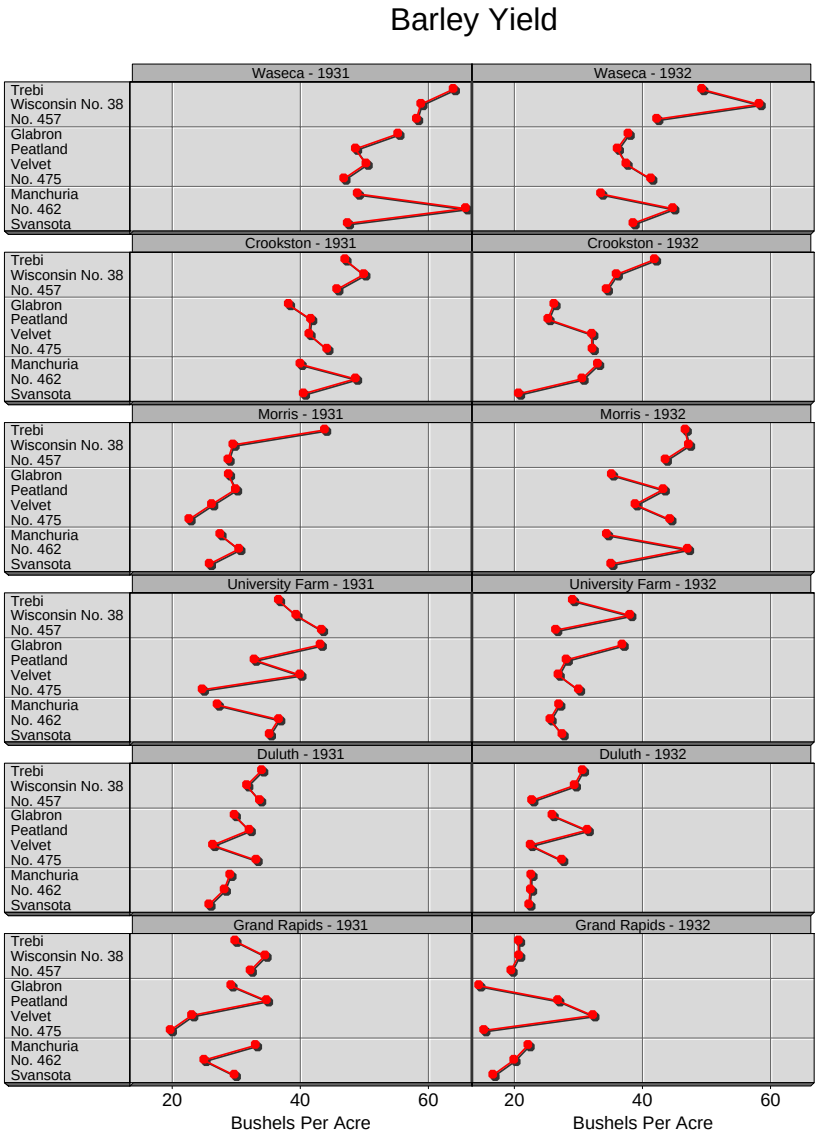


Figure 2: An enhanced dot plot, making use of lighting and perceptual grouping.

7. Access and Thoughts Spawned by Figure 2

We produced the graphics in this paper using Splus. Script files for the examples are available via anonymous `ftp` to `galaxy.gmu.edu`. Change directory to `pub/dcarr/newsletter/lightlayer`.

Lighting models, doubling buffering for animation, map layer composition and a host of other graphics tools have long been used in computer graphics or in GIS environments. The computer human interface community is experimenting with some very interesting interaction methods. This gives pause for thought. Yes, the business graphics in spreadsheet software are very old designs. However, the graphics that many of us statistical packages user cope with is also very old. One can go only so far by putting new wrappers on old graphics engines. Our community could use more modern environments for combining statistical transformation with visual representations.

This paper skips over several variations and extensions of methods used in the design of Figure 2. Some very old density and texture applications that use 3-D lighting models have yet to see the light of publication. Looking forward, the power of using lines and "ribbons" and translucent surfaces in statistical graphics is far from been fully appreciated. Much work remains.

Some of what remains to be done may seem surprisingly simple. A perceptual question raised by Figure 2 is, "how close is the graphic." Vocal presentations can be too soft, meet the listener half way, or be too loud. Graphics can go all the way from being too remote to being "in your face." An example of the latter is a full

screen 3-D three-valued pie chart. Too loud or too close is often associated with lack of content and purposeful inhibition of thought. Now immersive graphics are an option. Where this fits in the spectrum remains to be seen.

The desire here has been to lift the information off the page but not very far. Our goal is to actively present the information (move it forward from flatland), yet present it gently. The visual flow and perceptual grouping correspond to efforts in clear articulation. That is, we don't want the reader to work harder than necessary to translate the message and want to provide the reader with the freedom to think about the message. There are simple facets of the graphics presentation process that seem to be little studied. For starters, how big should dots be for us to easily see the color and how small should they be for us to take the dot center for granted. More generally, what is the right balance in terms of perceptual proximity to serve the intended communication purpose?

To a designer's eyes graphics are rarely finished but rather opportunities for further experimentation. Perhaps readers will look at both Figures 1 and 2 and come up with something better.

Acknowledgements

S-Plus is a register trademark of Mathsoft, Inc. This work was supported by the EPA under cooperative agreement No. CR8280820-01-0. The article has not been subject to review by EPA, does not necessarily reflect the view of the agencies, and no official endorsement should be inferred.

References

Blinn, J. (1996) *Jim Blinn's Corner: A Trip Down the Graphics Pipeline*, Morgan Kaufmann, SF, CA.

Card, S. K., J.D. Mackinlay and B. Shneiderman, Editors (1999) *Reads in Information Visualization: Using Vision to Think*, Morgan Kaufmann, SF, CA.

Carr, D. B. (1993) "Production of Stereoscopic Displays for Data Analysis," *SCGN*, Vol.4 No. 1, pp. 2-7.

Carr, D. B. (1994) "Converting Tables to Plots," Tech Rpt No. 101, Center for Computational Statistics, GMU.

Carr, D. B. and A. R. Olsen (1996) "Simplifying Visual Appearance By Sorting: An Example Using 159 AVHRR Classes," *SCGN*, Vol. 7 No. 1 pp. 10-16.

Carr, D. B., A. R. Olsen, J. P. Courbois, S. M. Pierson, and D. A. Carr (1998) "Linked Micromap Plots: Named and Described," *SCGN*, Vol. 9 No. 1 pp. 24-32.

Carr, D. B. and S. Pierson (1996) "Emphasizing Statistical Summaries and Showing Spatial Context with Micromaps," *SCGN*, Vol. 7 No. 3 pp. 16- 23.

Cleveland, W. S. (1985) *The Elements of Graphing Data*, Hobart Press, NJ.

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

Cleveland, W. S. (1993b) "A Model of Studying Display Methods of Statistical Graphics," *JGCS*, Vol. 2, No. 4.

Cleveland, W. S. and R. McGill (1984) "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *JASA*, 79:531-554.

Ernst, B. (1976) *The Magic Mirror of M.C. Escher*, Ballantine Books, NY.

Foley, J. D., A. van Dam, S. K. Feiner, and J. F. Hughes (1990) *Computer Graphics: Principles and Practice, 2nd Edition*, Addison-Wesley, NY.

Friedhoff, R. M and W. Benson (1991) *The Second Computer Revolution - Visualization*, Freeman and Company, NY.

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

MacEachren, A.M. (1995) *How Maps Work: Representation, Visualization, and Design*, The Guilford Press, NY.

Mathsoft (1995) *S-PLUS Trellis Graphics User's Manual, Version 3.3*, MathSoft, Inc. Seattle, WA.

Penrose, R. (1958) British Journal of Psychology, 49(1).

Pickle, L.W., Mingle, M., Jones, G. K. and White, A. A. (1997) *Atlas of United States Mortality*, Hyattsville, MD: National Center for Health Statistics.

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

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

Wegman, E. J., D. B. Carr and Qiang Luo (1993) "Visualizing Multivariate Data," *Multivariate Analysis: Future Directions*, ed. C.R. Rao, North Holland, New York. pp. 423-466.

Wilkinson, L. (1993) "Comment on A Model for Studying display Methods of Statistical Graphics," *JCGS*, 2(4).

Dan Carr
Center for Computational Statistics
George Mason University
dcarr@galaxy.gmu.edu

Ru Sun
Institute for Computational Sciences
and Informatics
George Mason University
rsun@osf1.gmu.edu

A Word from the Editor of JCGS

By Andreas Buja

This is the second installment of an occasional column from the editor of the "Journal of Computational and Graphical Statistics." In the previous column I introduced JCGS, its history and charter, its ambitions (publish only the best) and mode of operation (fully electronic submission and review). I also reported a dramatic increase of 60% in the submission rate over 1997; meanwhile, this increase has increased to 100%. Consequently, we have seen a doubling of the number of pages in the recent June issue. The forthcoming September issue promises to be even more voluminous, partly due to a special section on massive datasets. The editorial board of associate editors had to be expanded to absorb the volume and to maintain the high quality of the review process. Being at the helm of a journal during a period of such growth is extremely satisfying and well worth the unexpected demands on my time. Equally satifying are the qualitative aspects of the submissions: We see excellent work being published, and many leading authors regularly submit their manuscripts to JCGS, a trend that had started under the stewardship of the previous editor, Bill Kennedy.

The success of JCGS and the ensuing workload have slowed down a couple of initiatives on which I was hoping to report in the present column: electronic publishing and a software adjunct. More on this soon.

An area where progress has occurred is in structuring and enlivening the content of JCGS: Future issues will occasionally include special sections and discussed papers. We make a strong start along these lines in the September issue with both a special topics section of seven articles on massive datasets, and a graphics paper by Eno and Terrell on "Scatterplots for Logistic Regression" with a comment by Dan McCaffrey and a rejoinder by the authors. The December issue will most likely be plain, but early next year we hope to have a special section on systems and languages in statistics, and a discussed paper on EM and "optimization transfer" by Lange, Hunter and Yang.

The idea of special sections is to actively solicit papers on important topics that do not regularly get published, or at least not with the desirable frequency. Such topics may be found for example in areas that have not yet jelled into well-defined mainstream research, or never will, massive datasets being a case in point. Many among us care about the challenges posed by large amounts of data, but these challenges seem to define more an area of fermentation than a well-defined stream of scholarly research. Still, it is often these open-ended problems that drive new ideas and future research priorities.

Solicitations for special sections are a special challenge for editors. How can one persuade authors to write about a particular topic at a particular time? In order to raise the odds of success, Sallie Keller-McNulty and I, co-editors of this section, went back to an important workshop on massive datasets held in 1995. We selected a few entries from the proceedings and asked the authors for updates of their articles, taking into account the lessons learnt in the last four years. Seven out of eight solicitees responded positively. The result is a display of important work in medical and satellite imaging, analysis of telecommunications network data and data mining in health care, concluded with an assessment of the situation by Peter J. Huber.

The September issue is in color, which allows the authors of the massive datasets articles to show their images in the proper medium. This also allows us to publish two data visualization articles that depend on color in essential ways. Both concern Mosaic plots, the topic also of an article by Heike Hoffman in the previous issue of this Newsletter. Michael Friendly writes about "Extending Mosaic Displays: Marginal, Partial, and Conditional Views of Categorical Data," and Martin Theus and Stephan Lauer write about "Visualizing Loglinear Models," with Mosaic plots, of course. On an amusing note: The Titanic survival data seem to become to categorical data analysis what Fisher's Iris data are to discriminant analysis. Hoffman, Friendly, Theus and Lauer all use this data for illustration.

A third visualization paper is the above mentioned discussed paper by Eno and Terrell on "Scatterplots for Logistic Regression." Discussions are the perfect way to expose novel ideas to the kind of critique that ultimately establishes their value. Eno and Terrell's idea is novel, and Dan McCaffrey in his discussion makes a start at bringing out what works and how and when.

As Bayesian computing is a major part of current statistical computing, many will be interested to know what the September issue offers in this area: "An MCMC Convergence Diagnostic Using Subsampling" by S. G. Giakoumatos, I. D. Vrontos, P. Dellaportas and D. N. Politis, and "Logarithmic Pooling of Priors Linked by a Deterministic Simulation Model" by Geof H. Givens and Paul J. Roback

Finally, we publish two of the four award winning papers of the 1998 student competition: "Approximate Conditional Inference in Logistic and Loglinear Models" by Alessandra R. Brazzale, and "Bayesian Analysis of a Two State Markov Modulated Poisson Process" by Steven L. Scott. Both papers are excellent and passed the JCGS review in no time. Two more award winning papers will hopefully be ready for publication at a later point, but we thought we should not hold up the two that passed the hurdle early on.

Andreas Buja
AT&T Labs - Research
andreas@research.att.com

NEWS CLIPPINGS AND SECTION NOTICES

Winning Student Papers!

The results of the Computing Section's 1999 Student Paper Competition are in! The requirements of the competition were that the student be the first author of a paper in the area of statistical computing, which might be original methodological research, a novel application, or a software-related project. The four winners are invited to present their papers at a special contributed session at the Joint Statistical Meetings, and the Section gives them a grant towards their attendance expenses. The competition is open to students in the fall of the year prior to the competition.

A number of good entries were received, from which the selection committee, consisting of the Council of Sections representatives of the section, selected four. These are (in alphabetical order):

Alexandre Bureau, UC Berkeley, "*An S-PLUS Implementation of Hidden Markov Models in Continuous Time*" (with James P. Hughes and Stephen Shiboski)

Ilya Gluhohvsky, Stanford, "*Image Restoration Using Modifications of Simulated Annealing*"

Peter D. Hoff, University of Wisc-Madison, "*Nonparametric Maximum Likelihood Estimation Via Mixtures*"

Muhammad Jalaluddin, University of Wisc-Madison "*An Algorithm for Robust Inference for the Cox Model with Frailties*" (with Michael R. Kosorok)

The students will be recognized at the Statistical Computing/Statistical Graphics business meeting at the 1999 Joint Meetings in Baltimore, and will make presentations based on their papers in a special contributed session (currently scheduled for Tuesday, August 10 at 2 pm). The papers will also be published in an upcoming issue of JCGS.

John M. Chambers Wins ACM Award

Initiates Student Computing Award

On Saturday May 15, the Association for Computing Machinery (ACM) presented its prestigious Software System Award to researcher John Chambers of Bell Labs. The award was presented for the design of the S System for statistical computing, which the ACM said has "forever altered how people analyze, visualize, and manipulate data ... S is an elegant, widely accepted, and enduring software system, with conceptual integrity, thanks to the insight, taste, and effort of John Chambers." And in more good news for statistical computing, Chambers announced plans to turn over his $10,000 award to the American Statistical Association to endow a new prize that will recognize outstanding student work in software for statistics.

This is the first time in its 17-year history that the Software System Award has been given for data analysis software, and the first time it has been given to a statistician. Beginning with the UNIX* System - created in 1969 by Bell Labs researchers Dennis Ritchie and Ken Thompson - the Software System Award has recognized ideas and developments that have had a major, lasting impact on computing, such as TCP/IP and the World Wide Web.

The first versions of S in the 1970s pioneered the use of data visualization and interactive statistical computing. Subsequent versions provided richly enhanced modeling capability, and user extensibility, based on its functional object-based approach.

Still more recent versions provide a powerful class/method structure, new techniques to deal with large objects, extended interfaces to other languages and files, object-based documentation compatible with HTML, and powerful interactive programming techniques. The commercial version, S-Plus, is used across many disciplines where analysts must struggle with creative ways to manage and extract useful information from data. More information about S is available at cm.bell-labs.com/stat/S.

John Chambers is one of the researchers pursuing a new joint project, *Omega*, aimed at the next generation

of statistical software. In this project, emphasis is on Java and distributed computing, with the goal of a wide range of new software in open source, benefiting and involving the whole statistical computing community. More about Omega and its activities can be found on the project's website, `www.omegahat.org`.

Looking toward the future in another way, Chambers donated the prize money from his Software System Award - all $10,000 of it - to the ASA, to endow a prize for the best student software written to support the computing used in statistics.

The purpose of the prize, Chambers said, "is to recognize contributions in the design and implementation of software that has value for the statistical community, and to raise awareness in that community of the importance of good software to those involved in statistical applications and research." In particular, Chambers hopes the existence of the prize will foster greater recognition that software design is a key component of statistical research, deserving recognition on the same level as mathematical theory and other essential elements.

Statistical software created by undergraduate or graduate students in any field will be eligible for the prize. The prize will be administered by the Statistical Computing section of the ASA. Details of the prize will appear in future issues of the Newsletter and Amstat News.

Short Courses at JSM '99

Regression Graphics: Ideas for Studying Regressions Through Graphics

Professor R. Dennis Cook
University of Minnesota

Sunday, August 8
Co-sponsored by the Graphics and Computing Sections

Professor R. Dennis Cook of the University of Minnesota will be teaching a one-day course on Sunday, August 8, prior to the Joint Statistical Meetings in Baltimore. The title of the course is *Regression Graphics: Ideas for Studying Regressions Through Graphics* and it is based on his recent book by the same title (Wiley, 1998, ISBN 0-471-19365-8). The course is co-sponsored by the Section on Statistical Graphics and the Section on Statistical Computing. Participants in the course will receive the book, software, and other materials as part of the course fee.

The focus of the course is *dimension reduction*: In regression, our goal is to learn about the conditional distribution of a response Y given a $p \times 1$ vector of predictors $\mathbf{X}$; and in regression graphics, a central goal is to try to reduce the dimension of $\mathbf{X}$ without loss of information on the conditional distribution of $Y|\mathbf{X}$ and without requiring a model. We call this *sufficient dimension reduction*, and the associated graphics are called *sufficient summary plots*. These plots can be quite useful in a regression study, particularly in guiding the initial choice of a model.

The course will promote a new context for regression that centers on dimension reduction and sufficient summary plots. The methods will be developed and illustrated using interactive computer graphics. Prerequisites are familiarity with standard regression methodology at the level of one of the major textbooks in the area, and basic knowledge of linear algebra.

The course will begin with an introduction, foundations of central reduction subspaces, and illustrative examples. We will then proceed to discuss the relationships between these ideas and standard fitting methods (least-squares, robust regression, and generalized linear models). Three-dimensional plots (3D plots) may be used to visualize and to help identify the central subspace. As the number of predictors increases, a series of 3D plots, guided by *graphical regression* methods, may be needed. Specific numerical methods such as sliced inverse regression (SIR) and sliced average variance estimation (SAVE) are used and put in the context of the examples presented earlier. We will also discover new graphical approaches to well-known problems like the identification of outliers and regression mixtures.

Professor Cook is well known as one of the principle researchers in regression graphics and diagnostics. He is the author or co-author of several books and a large number of research papers, and he is a Fellow of the ASA, IMS, and a member of the International Statistical Institute.

The web site `www.stat.umn.edu/RegGraph/` provides more information about the book, plus links to a short tutorial on regression graphics, data sets, examples, information about software, and a more detailed course outline.

Don't miss this opportunity to learn about a new, ground-breaking approach to one of the most popular statistical methodologies.

Russell V. Lenth

Statistical Shape Analysis

Ian L. Dryden and Kanti V. Mardia
University of Leeds, UK

Wednesday, August 11, 1999
Co-sponsored by the Computing and Graphics Sections

Statistical Shape Analysis involves methods for the geometrical study of random objects where location, rotation and scale information can be removed. The subject is a new and exciting area of statistics, offering many fresh challenges. There have been many advances made in the past 10 years, and there is a huge variety of applications. The course

- lays the foundations of the subject
- discusses key ideas
- discusses the very latest developments
- offers practical guidance
- gives comparisons of techniques.

and will be based around the new text *Statistical Shape Analysis* by the presenters.

The course primarily concentrates on landmark data, where key points of correspondence are located on each object. Careful consideration of the similarity invariances requires methods appropriate for non-Euclidean data analysis. In particular, multivariate statistical procedures cannot be applied directly, but can be adapted in certain instances. Various applications will be given throughout, including in biology, medicine, image analysis, genetics and agriculture.

We begin with introductory material on shape, size and coordinate systems. Planar Procrustes analysis is then discussed to highlight the main components of shape analysis. The shape space and general Procrustes methods are introduced, probability distributions for shape are described and statistical inference is discussed. Some deformation methods for shape change are also given and we also discuss shape in image analysis. Finally, various alternative procedures including landmark-free methods are critically discussed and compared.

Some computer demonstrations will be given. Further details can be found at the web site `www.amsta.leeds.ac.uk/~iand/course`.

Stochastic Optimization and the Simultaneous Perturbation Algorithm

James C. Spall
John Hopkins University

Co-sponsored by the Section on Physical and Engineering Sciences and the Computing Section

This course will introduce statistical practitioners and researchers to some of the broad issues in the field of stochastic optimization and discuss a relatively new stochastic optimization approach (simultaneous perturbation stochastic approximation-SPSA) that has attracted considerable international attention in a variety of problems in the physical and social sciences, engineering, and biomedicine. The essential features of SPSA are its efficiency for multivariate problems and its relative ease of implementation for practitioners (which follows, among other aspects, by avoiding the objective function gradient vector needed in many other methods). Applications in statistical parameter estimation, neural network training, experimental design, and simulation-based optimization will be discussed. Some comparisons with genetic algorithms, simulated annealing, and other approaches will be included. Since SPSA is relatively easy to implement, it is expected that the participants will be able to quickly put into practice many of the ideas in the course.

The instructor, Dr. James Spall, has extensive experience in both the practical and theoretical aspects of stochastic optimization. He is the author of a forthcoming book, Introduction to Stochastic Search and Optimization (Wiley), and has published extensively in the statistics, engineering, and simulation literature. Dr. Spall is employed by the John Hopkins University Applied Physics Laboratory and teaches part-time in the Johns Hopkins School of Engineering. He has given short courses at many technical conferences and holds two U.S. patents for inventions in control systems.

SECTION OFFICERS

Statistical Graphics Section - 1999

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

Edward Wegman, Chair–Elect
703-993-1680
George Mason University
ewegman@galaxy.gmu.edu

Michael M. Meyer, Past–Chair
412-268-3108
Carnegie Mellon University
MikeM@stat.cmu.edu

Deborah Swayne, Program Chair
973-360-8423
AT&T Labs – Research
dfs@research.att.com

Daniel Carr, Program Chair–Elect
703-993-1671
George Mason University
dcarr@galaxy.gmu.edu

Antony Unwin, Newsletter Editor (98-00)
49-821-598-2218
Universität Augsburg
unwin@uni-augsburg.de

Robert L. Newcomb, Secretary/Treasurer (98-99)
714-824-5366
University of California, Irvine
rnewcomb@uci.edu

Michael C. Minnotte, Publications Liaison Officer
801-797-1844
Utah State University
minnotte@math.usu.edu

Bradley A. Jones, Rep.(99-01) to Council of Sections
508-368-8458
Mathworks
brad@mathworks.com

David W. Scott, Rep.(98-00) to Council of Sections
713-527-6037
Rice University
scottdw@rice.edu

Roy E. Welsch, Rep.(97-99) to Council of Sections
617-253-6601
MIT, Sloan School of Management
rwelsch@mit.edu

Statistical Computing Section - 1999

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

Russel D. Wolfinger, Chair–Elect
919-677-8000
SAS
sasrdw@sas.com

Karen Kafadar, Past–Chair
303-556-2547
University of Colorado-Denver
kk@tiger.cudenver.edu

Mark Hansen, Program Chair
908-582-3869
Bell Laboratories
cocteau@bell-labs.com

James S. Marron, Program Chair–Elect
919-962-5604
University of North Carolina, Chapel Hill
marron@stat.unc.edu

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

Merlise Clyde, Secretary/Treasurer (98-99)
919-681-8440
Duke University
clyde@isds.duke.edu

James S. Marron, Publications Liaison Officer
919-962-5604
University of North Carolina, Chapel Hill
marron@stat.unc.edu

Leland Wilkinson, Rep.(99-01) Council of Sections
(312) 651-3270
SPSS
leland@spss.com

Terry M. Therneau, Rep.(97-99) Council of Sections
507-284-1817
Mayo Clinic
therneau@mayo.edu

Naomi S. Altman, Rep.(97-99) to Council of Sections
607-255-1638
Cornell University
naomi_altman@cornell.edu

INSIDE

Statistical COMPUTING & GRAPHICS

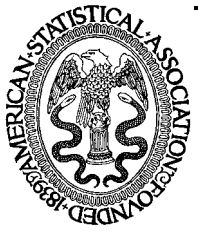
The *Statistical Computing & Statistical Graphics Newsletter* is a publication of the Statistical Computing and Statistical Graphics Sections of the ASA. All communications regarding this publication should be addressed to:

Mark Hansen
Editor, Statistical Computing Section
Statistics Research
Bell Laboratories
Murray Hill, NJ 07974
(908) 582-3869 • FAX: 582-3340
cocteau@bell-labs.com
cm.bell-labs.com/who/cocteau

Antony Unwin
Editor, Statistical Graphics Section
Mathematics Institute
University of Augsburg
86135 Augsburg, Germany
+49-821-5982218 • FAX: +49-821-5982280
unwin@uni-augsburg.de
www1.math.uni-augsburg.de/~unwin/

All communications regarding ASA membership 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@amstat.org



Nonprofit Organization
U. S. POSTAGE
PAID
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.