A Word from our 2006 Section Chairs

PAUL MURRELL
GRAPHICS

Have you ever wanted to write a book, but not known where to start?
Now is a very good time to jump in, because there is currently a very simple recipe for success: just put R in the title and you will have to beat the publishers off with a stick!

But it is not just Statistical Computing hogging the literary limelight. Judging by publications in recent years, a second rule of thumb would be this: just publish a book on Statistical Graphics!


STEPHAN R. SAIN
COMPUTING

This year has certainly gone by in a hurry and it is good to see snow again on the mountains above Denver (unless you were trying to catch a flight into or out of Denver last week!). The computing section has had an active and successful year. The JSM went very well with a strong program and a well-attended mixer (co-organized with the Section on statistical graphics). The student and Chamber’s award winners represented an impressive collection of research.

Continues on Page 2.........

Featured Article 4
Deadlines 8
Tools for Computing 9
Teaching Graphics 14
Tools for multivariate 18
News 23
Graphics, Continues from Page 1....
(There is a nasty conflict of interest, but I must also mention my own effort, “R Graphics”, because it cunningly combines both rules!).

Furthermore, “Interactive Graphics for Statistics” by Martin Theus and Matthias Schonlau is soon to appear and “Interactive and Dynamic Graphics for Data Analysis” by Dianne Cook, Deborah F. Swayne, Andreas Bujais is under development (see http://www.ggobi.org/docs/).

Combine this evidence with the fact that Hadley Wickham won this year’s Chambers Award with a statistical graphics package ggplot and it is clear that Statistical Graphics is in excellent health.

This year’s JSM reflected that state of health with great participation in five invited sessions, one topic-contributed, and two regular contributed sessions, plus a coffee roundtable and two roundtable luncheons. Many thanks to Juergen Symanzik and Simon Urbanek for their tremendous efforts in organizing these sessions.

The section also co-sponsored the Data Expo poster session at the JSM, which, thanks to amazing enthusiasm from the participants, was also very successful. The results and poster materials from this session are now available online at

http://www.amstat-online.org/sections/graphics/dataexpo/2006entries.php. Looking ahead to next year, there is likely to be plenty of activity again. Our Program Chair, Simon Urbanek, has a number of invited sessions planned; please contact him if you have ideas for topic-contributed sessions or even if you are just planning a regular-contributed talk. The section is also looking at sponsoring a number of continuing Continuing Education proposals; thanks to John Castelloe for his continuing efforts with the CE programme.

My term as chair is coming to an end, so I would like to thank all of the section officers for their hard work. Occupying a position within the executive committee of the section does sometimes get quite busy, but it is extremely satisfying to be able to contribute to the sustenance and growth of Statistical Graphics.

Computing, Continues from Page 1....
Also during the year, the section helped support a workshop on fast manifold learning, a Bioconductor conference, the useR! conference, and we continued our support for the annual Interface conference.

It is not too early to start thinking ahead to next year’s JSM in Salt Lake City. The invited program is quickly getting settled and a number of great continuing education proposals have been submitted. I also urge everybody to consider organizing or participating in the topic contributed program. This is great way to further highlight what is going on in the fields of computational statistics and statistical computing. And, the number of topic contributed sessions the section supports is supposed to improve our chances of obtaining more invited sessions. The process for putting one of these topic contributed sessions together is pretty easy and I have organized a couple of these sessions in the past. Simply choose a topic and get five of your colleagues who are probably already planning on attending the meetings to participate in the session. Then contact the program chair (Ed Wegman) with the title of your session and the speaker information. Easy! One clear benefit of the topic contributed sessions is that there are only five speakers as opposed to seven in the contributed sessions. That’s a whole five more minutes per speaker!

Starting in January, a new slate of officers takes over leadership of the section. John Monahan (North Carolina State University) takes over the chair’s role with Deborah Nolan (University of California, Berkeley) coming in as chair-elect. Next year’s program chair is Edward Wegman (George Mason University) and Wolfgang Jank (University of Maryland) will be the new program chair-elect. Finally, the section has a new awards officer as J.R. Lockwood (Rand Corporation) will be taking over for a three-year term.

I’d like to thank the current officers and volunteers for all their hard work to support the section. Tim Hesterberg, past-chair, has been an invaluable resource. John Monahan, as chair-elect, has also been a big help. David Poole has done an excellent job as secretary/treasurer as well as the current program chair, Michael Trosset. Vincent Carey, Robert Gentleman, and Juana Sanchez have served as our council of section representatives and Todd Ogden has
been our web liaison and publication officer. A number of people also helped out with judging our section’s awards this year, an incredibly important job that they did very well, despite the inherent challenges in such endeavors. Finally, I’d like to thank the newsletter editors, Di Cook and Juana Sanchez, for all their effort putting the newsletters together.

I’d like to close with some final thoughts. Over the past several months, I’ve been working with the Geophysical Statistics Project at the National Center for Atmospheric Research, and I’ve quickly gotten a new perspective about the challenges facing those of us working in applied and computational statistics and about the role of statistics in “big science.” As in many areas, data sets in the geophysical sciences are already large, but getting larger still, and more and more is being demanded from the analysis of such data sets. There are fantastic computational tools at our disposal, yet at times it feels like we are just scratching the surface of how to use them. During a recent conference talk, I was asked if we could handle the size and complexity of the data sets we were proposing to study. Without thinking too much about it, I quickly responded that we could. Even after some contemplation while sitting in the airport awaiting my flight home, I’m confident that we are up to the challenges of modern scientific data analysis although it will probably take even more cooperation with not only our scientific colleagues but our computer scientist friends. But, from my point of view, it is exactly these challenges that make what we do fun!

Lastly, there are a number of opportunities to get involved in the section. If you are interested in helping out or even if you just have some suggestions about how the section can better serve the membership and the community, please feel free to contact me or the incoming chair.

---

**Editorial Note**

Juana Sanchez

It is with great sadness that I have to announce that Di Cook, Editor of the SCGN newsletter for Statistics Graphics, has to leave this position to dedicate her time to other worthy endeavors that no doubt will significantly impact Statistical Graphics as much as all her past work. She has dedicated many years of her life to bring you the latest news and excellent research contributions that have had an impact in the development of Statistical Graphics and Computing. Her vision and her creativity have made possible the continuity of the newsletter as other Editors have been passing the torch while she stayed. Even as she steps down, she has contributed to this new issue in the article by Heike Hofman, Dianne Cook and Charles Kostelnick on how to teach Statistical Graphics to undergraduates. Di has also reviewed Naomi Robbins’ latest book for us and has, as usual, attracted other contributions. She is going to be missed a lot, but I am sure that she will continue to contribute her or her students’ research to the newsletter and will keep sharing her editorial standards and good taste with us. As she steps down, we are looking for a volunteer to take her place (see insert on page 13 in this newsletter).

This issue of the newsletter also features an article by Thomas Lumley on the R package dichromat, which helps statisticians evaluate the suitability of color schemes for dichromats, people with two-dimensional rather than three-dimensional color perception. Jouni Kerman and Andrew Gelman present the R package Umacs, Universal Markov Chain Sampler, which helps write Gibbs/Metropolis samplers to simulate from posterior distributions. They also present the complementary package rv, which helps manage and summarize the results of the simulations. Another feature article by Alexander Gribov, Antony Unwin, and Heike Hofmann, present the software Gauguin for visualization of Multivariate data and they illustrate how it works via the Data Expo NASA data set. Finally, do not miss the Announcements of the Student Paper and the Chambers Competitions. It is still time to apply.

Finally, all photos appearing here are Copyright 2006 ASA.
Red–green color distinctions have been important to humans and many primates for millions of years. Plants, and their bird and insect customers for flowers and ripe fruit, had coevolved color coding schemes, and monkeys and apes that could see colors could benefit from this arrangement. The mechanisms that developed are interesting for their own sake and in the light they shed on color coding in graphics today. Much more detail on this topic can be found in Mollon (2000).

Primitive primates can (and presumably early primates could) distinguish light and dark, and yellow and blue. Mutations in the retinal cells sensitive to a wide color band in the yellow part of the spectrum led to versions that were preferentially sensitive to red or to green light. Individuals with two different copies of the gene could distinguish red and green, bringing obvious advantages. Since the gene is on the X chromosome, males could have only one copy and so would not have these benefits. Later mutations produced a doubled gene, so that a single X chromosome could carry both color versions and both males and females could see in full color.

In a few percent of chromosomes in humans the duplicated gene is broken: either one version does not work at all or both versions are sensitive to the same color range. This is not a problem for women, who have a spare copy, but in men it produces the red–green color blindness often called ‘Daltonism’ after the first person to write about it in detail, British chemist and physicist John Dalton. Dalton noticed, for example, that most people seemed to think that pink and red were similar colors, but he thought pink and blue were similar and quite different from red. DNA analysis of one of Dalton’s eyes (left to the Royal Society in his will) shows that he was missing the green-sensitive pigment, a condition called “deuteranopia”; “protanopia”, in

Figure 1
which the red-sensitive pigment is missing, is about equally common.

Given the fairly high frequency of colorblindness or milder forms of anomalous red–green vision it is worth trying to avoid graphics that communicate information primarily through red–green distinctions. The R package dichromat (Lumley, 2003) is designed to help statisticians evaluate the suitability of color schemes for dichromats, people with two-dimensional rather than three-dimensional color perception.

There is a single primary function, dichromat(), with two arguments. The first argument is a vector of colors, and the second specifies one of the three types of dichromatic vision: “protan” (no red), “deutan” (no green), and “tritan” (no blue). The output is a modified set of colors that would look the same as the input to someone with the specified type of dichromatic vision, but which has had the color range collapsed. The data for this transformation comes from experiments by Vienot et al. (1999), and a similar tool for checking bitmap images is on the internet at http://www.vischeck.com.
1. An example: the Ishihara test plates

An interesting illustration of these transformations comes from applying them to the Ishihara plates (Ishihara, 1959) used for diagnosing color vision deficiencies.

Figure 1, where someone with normal vision sees the number 74, someone with red–green color-blindness sees 21. Scanning this image then reading the data into R, allows us to see how the test works. Figure 2 shows a scatterplot matrix of all the colors that appear on at least ten of the roughly 10^6 pixels in the image.

The coordinate system for these colors is CIE Lab, which is an orthogonal coordinate system for a roughly perceptually uniform color space, meaning that any two colors separated by the same distance are about equally easy to distinguish.

In this coordinate system, “L” runs from dark to light, “a” from green to red and “b” from yellow to blue. A red–green color-blind viewer will be able to distinguish colors only if they are well separated in the L and b dimensions. In the upper panels the plotting colors are the actual colors of the pixels, in the lower panels the plotting colors have been transformed with dichromat.

Figure 3 shows the same points in the Lab coordinate system after transformation with dichromat. The ab scatterplot panel shows that red–green distinctions have been abolished, with the effect of collapsing the red–green and blue–yellow information to a single dimension. The reason that the ab panel is not a perfect line is because of complications in translating from colors of light as used by dichromat to colors of objects under white light as described by the Lab coordinate system.

The Lab coordinate system is designed so that pairs of points separated by the same small distance have approximately the same perceptual difference in color. This means that we can use Euclidean distance in multivariate data analyses such as clustering to explore the data. As there are 17000 observations we need a clustering method that can handle large data sets, and I
used a subsampling-based partitioning-around-medioids algorithm developed by Kaufmann & Rousseuw (1990) and implemented as clara() in the R cluster package (Maechler et al. 2006).

Figure 4 plots the cluster indicators for a six-cluster solution based on the original data and a five-cluster solution after transformation. The first two columns are the foreground cluster with normal vision, the points making up the '74'; the first row is the foreground cluster with color-deficient vision, giving the '21'. We can see that collapsing the red–green distinctions moves some yellow-green points into the background and moves some pinkish points into the foreground. Looking at the original plate, we can see how removing the yellow-green points will turn the '4' into a '1'. Looking more carefully we can see the pinkish points around the outline of the '7', providing the base of the '2' and rounding off the sharp upper right corner of the '7'.

Another view of how the points are hidden comes from a scatterplot matrix colored according to cluster membership. In Figure 5 the upper panels are colored according to cluster membership for normal vision, the lower panels according to cluster membership for deuteranopia. In each case the plotting color is taken from the center of the cluster.

The key sets of colors are clearest when comparing the upper and lower 'ab' panels. In the upper panel there are three pairs of small color groups: one green, one yellow-green, and one orange. The green and yellow-green patches form the '74'. In the lower 'ab' panel the green pair have joined with the orange pair to form the '21' and the yellow-green pair has been absorbed into the background. In the upper, normal-vision graphs, the clustering is clearest in the 'La' panel: the red–green signal is by far the strongest. In the lower graphs, for deuteranopia, the clustering is clearest in the 'Lb' panel; with the red–green signal suppressed the weaker blue–yellow signal defines the clusters.

2. Choosing color schemes

Figure 6 shows a small example of the color-coded matrix displays common in bioinformatics. The lower panels uses the traditional red–green scale; the upper
For normal vision, the information is clear on either scale, and the red–green scale is probably preferable for its greater symmetry. The blue–yellow scale, however, conveys the same information in the second column (deutanopia) and the third column (protanopia) as with normal vision. On the red–green scale not only does the range of color become severely compressed, it is distorted in a non-monotone way.

The dichromat package and the Vischeck web site provide a way to check your graphics for visibility, but how can you fix a graph that is not sufficiently clear? As this example shows, simply replacing red and green with blue and yellow may often work well, but there are at least three important disadvantages. The first is that even moderately saturated blues and yellows have to differ in luminance: yellows are light colors and blues are dark colors. The second disadvantage is that it is more difficult to calibrate computers, projectors, and printers to reproduce intense blues accurately. A third disadvantage is that the retina has relatively few blue-sensitive cells, so that the spatial resolution of vision is much poorer. Blue–yellow color differences that would be easily visible in a map or image may be hard to

The ColorBrewer website, [http://www.colorbrewer.org](http://www.colorbrewer.org), is an excellent resource. It gives a selection of color schemes for maps and other large-area graphics, and also provides information on which schemes work well in a variety of difficult conditions --- including dichromatic vision, but also including reproduction on low-quality projectors and printers, and when reduced to greyscale. It is not necessary and may not be possible that a color scheme be equally good for people with normal and dichromatic vision and some compromises are inevitable. For example, a color gradient that is perceptually uniform to normal vision may be non-uniform to someone with deutanopia. However, it is possible to design most graphics so that the main information being conveyed is at least visible to people with red–green color-blindness.

**References**


**INTERFACE 2007, CALL FOR PAPERS**

The Interface 2007 meeting will take place in Philadelphia, Pennsylvania on 23-26 May 2007. The general theme of Interface 2007 will be Systems Biology. The conference is being hosted by the Center for Statistical and Information Science (Alan J. Izenman, Director), the Department of Statistics, and the Center for Science & Technology (Zoran Obradovic, Director) at Temple University. Interface 2007 will take place at the DoubleTree Hotel, Philadelphia.


This announcement is a call for papers. Submissions are welcome. For more information, contact Alan J. Izenman, Department of Statistics, Speakman Hall, 1810 North 13th Street, Temple University, Philadelphia, PA 19122-6083, (215) 204-8166, alan@temple.edu or csis@temple.edu. A website is currently under construction.

---

**The Student Paper Competition Deadline is**
Monday Dec 18, 2006, by 5:00pm EST

**The Chambers Statistical Software Deadline is**
Monday, February 26, 2007 by 5:00pm EST

See Page 23 of this Newsletter
Tools for Computing

TOOLS FOR BAYESIAN DATA ANALYSIS IN R
Jouni Kerman, Novartis Pharma AG, Switzerland
Andrew Gelman, Columbia University, USA
jouni@kerman.com

Introduction

Bayesian data analysis includes but is not limited to Bayesian inference (Gelman et al., 2003; Kerman, 2006a). Here, we take Bayesian inference to refer to posterior inference (typically, the simulation of random draws from the posterior distribution) given a fixed model and data. Bayesian data analysis takes Bayesian inference as a starting point but also includes fitting a model to different datasets, altering a model, performing inferential and predictive summaries (including prior or posterior predictive checks).

These tasks require a flexible computing environment that allows us to fit a Bayesian probability model (generating simulations from the joint posterior distribution), but also to manipulate and summarize simulations graphically and numerically.

The most general programs currently available for Bayesian inference are WinBUGS (BUGS Project, 2004) and OpenBUGS, which can be accessed from R using the packages R2WinBUGS (Sturtz et al., 2005) and BRugs. In addition, various R packages exist that directly fit particular Bayesian models (e.g. MCMCPack, Martin and Quinn (2005)), or emulate aspects of BUGS (JAGS). In this article, we describe the ongoing development of two R packages that perform important aspects of Bayesian data analysis.

Umacs

Umacs (Universal Markov chain sampler) is an R package (to be released) that facilitates the construction of the Gibbs sampler and Metropolis algorithm for Bayesian inference (Kerman, 2006b). Writing one’s own Gibbs/Metropolis sampler is sometimes necessary for large problems that cannot be fit using programs like BUGS.

Two programs implementing Gibbs samplers differ essentially just by their updating functions. Different Metropolis samplers sample from different posterior functions, but have a similar program structure. Umacs provides the necessary program structure around user-supplied Gibbs updating functions or Metropolis samplers, writing a complete, customized sampler function in R, ready to be run.

The user supplies data, parameter names, updating functions (which can be some mix of Gibbs samplers and Metropolis jumps, with the latter determined by specifying a log-posterior density function), and procedures for generating starting points. Using these inputs, Umacs generates (writes) a customized R sampler function that automatically updates, keeps track of Metropolis acceptances (and uses acceptance probabilities to tune the jumping kernels, following Gelman et al. (1995)), monitors convergence (following Gelman and Rubin (1992)), summarizes results graphically, and returns the inferences as arrays of simulations, or as simulation-based random variable objects (see rv, below).

Umacs is customizable and modular, and can be expanded to include more efficient Gibbs/Metropolis steps. Current features include adaptive Metropolis jumps for vectors and matrices of random variables (which arise, for example, in hierarchical regression models, with a different vector of regression parameters for each group). Real-time trace plots can be defined for any scalar parameters or for the convergence statistics, if desired (Figure 5).

Figure 1 illustrates how a simple Bayesian hierarchical model (Gelman et al., 2003, page 451) can be fit using Umacs: \( y_j \sim N(\theta_j, \sigma^2_j) \), \( j = 1, \ldots, J \) (\( J = 8 \)), where \( \sigma_j \) are fixed and the means \( \theta_j \) are given the prior \( t_\nu(\mu, \tau) \). In our implementation of the Gibbs sampler, \( \theta_j \) is drawn from a Gaussian distribution with a random variance component \( V_j \). The conditional distributions of \( \theta, \mu, V, \) and \( \tau \) can be calculated analytically, so we update them each by a direct (Gibbs) update. The updating functions are to be specified as R functions (here, theta.update, V.update, mu.update, etc.). The degrees-of-freedom parameter \( \nu \) is also unknown, and updated using a Metropolis algorithm. To implement this, we only need to supply a function calculating the log-
s <- Sampler(
  J = 8,
  sigma.y = c(15, 10, 16, 11, 9, 11, 10, 18),
  y = c(28, 8, -3, 7, -1, 1, 18, 12),
  theta = Gibbs(theta.update, theta.init),
  V = Gibbs(V.update, V.init),
  mu = Gibbs(mu.update, mu.init),
  tau = Gibbs(tau.update, tau.init),
  nu = SMetropolis(log.post.nu, nu.init),
  Trace("theta[1]"
)
)

Figure 1: Invoking the Umacs Sampler function to generate an R Markov chain sampler function s(…). Updating algorithms are associated with the unknown parameters ($\theta$, $V$, $\mu$, $\tau$, $\nu$). Optionally, the non-modeled constants and data (here $J$, $\sigma$, $y$) can be localized to the sampler function by defining them as parameters; the function $s$ then encapsulates a complete sampling environment that can be even moved over and run on another computer without worrying about the availability of the data variables. The "virtual updating function" Trace displays a real-time trace plot for the specified scalar variable (thus updating the graphical window which acts as a parameter).

The function produced by Sampler runs a given number of iterations and a given number of chains; if we are not satisfied with the convergence, we may resume iteration without having to restart the chains. It is also possible to add chains. The length of the burn-in period that is discarded is user-definable and we may also specify the desired number of simulations to collect, automatically performing thinning as the sampler runs.

Once the pre-specified number of iterations are done, the sampler function returns the simulations wrapped in an object which can be coerced into a plain matrix of simulations or into a list of random variable objects (see rv, below), which can be then attached to the search path.

Figure 2: Real-time trace plot of the scalar component $\theta_1$ in Umacs; different colors refer to different chains. It is possible to define any number of trace plots for any scalars in the model. A trace plot behaves conceptually just like a parameter that is updated during each iteration of the Gibbs sampler. In practice, we update the graph every 10 or 50 iterations not to slow down the sampler.

rv

rv is an R package that defines a new simulation-based random variable class in R along with various mathematical and statistical manipulations (Kerman and Gelman, 2005). The program creates an object class whose instances can be manipulated like numeric vectors and arrays. However, each element in a vector contains a hidden dimension of simulations: the rv objects can thus be thought of being approximations of random variables. That is, a random
scalar is stored internally as a vector, a random vector as a matrix, a random matrix as a three-dimensional array, and so forth. The random variable objects are useful when manipulating and summarizing simulations from a Markov chain simulation (for example those generated by Umacs). They can also be used in simulation studies (Kerman, 2005). The number of simulations stored in a random variable object is user-definable.

The rv objects are a natural extension of numeric objects in R, which are conceptually just “random variables with zero variance”—that is, constants. Arithmetic operations such as + and − and elementary functions such as exp and log work with rv objects, producing new rv objects.

These random variable objects work seamlessly with regular numeric vectors: for example, we can impute random variable \( z \) into a regular numeric vector \( y \) with a statement like \( y[\text{is.na}(y)] <- z \). This converts \( y \) automatically into a random vector (rv object) which can be manipulated much like any numeric object; for example we can write \( \text{mean}(y) \) to find the distribution of the arithmetic mean function of the (random) vector \( y \) or \( \text{sd}(y) \) to find the distribution of the sample standard deviation statistic.

The default print method of a random variable object outputs a summary of the distribution represented by the simulations for each component of the argument vector or array. Figure 3 shows an example of a summary of a random vector \( z \) with five random components.

```
> z
name     mean     sd   Min  2.5%  50%  97.5%  99.5% Max
[1] Alice  59.0  27.3  -28.66  1.66  42.9  59.1  75.6  114  163
[2] Bob    57.0  29.2  -74.14  1.98  38.3  58.2  75.9  110  202
[3] Cecil  62.6  24.1  -27.10 13.25  48.0  63.4  76.3  112  190
[4] Dave   71.7  18.7  -2.88  34.32 60.6  71.1  82.9  108  182
[5] Ellen  75.0  17.5  -4.12  38.42 64.1  75.3  86.2  108  162
```

Figure 3: The print method of an rv (random variable) object returns a summary of the mean, standard deviation, and quantiles of the simulations embedded in the vector.

Standard functions to plot graphical summaries of random variable objects are being developed. Figure 4 shows the result of a statement \( \text{plot}(x, y) \) where \( x \) are constants and \( y \) is a random vector with 10 constant components (shown as dots) and five random components (shown as intervals).

![Figure 4: A scatterplot of fifteen points (x, y) where five of the components of y are random, that is, represented by simulations and thus are drawn as intervals. Black vertical intervals represent the 50% posterior intervals and the gray ones the 95% intervals. This plot was simply obtained by a command \( \text{plot}(x, y) \) (with appropriate supplementary arguments). The light grey line is a regression line computed from the ten fixed points, included for reference.](image)

Many methods on rv objects have been written, for example \( \text{E}(y) \) returns the individual means (expectations) of the components of a random vector \( y \).

A statement \( \text{Pr}(z[1] > z[2]) \) would give an estimate of the probability of the event \( \{z_1 > z_2\} \).

Random-variable generating functions generate new rv objects by sampling from standard distributions, for example \( \text{rnorm}(n=10, \text{mean}=0, \text{sd}=1) \) would return a random vector representing 10 draws from the standard normal distribution. What makes these functions interesting is that we can give them pa-
parameters that are also random, that is, represented by simulations. If \( y \) is modeled as \( \mathcal{N}(\mu, \sigma^2) \) and the random variable objects \( \mu \) and \( \sigma \) represent draws from the joint posterior distribution of \( (\mu, \sigma) \)—we can obtain these if we fit the model with Umacs or BUGS for example—then a simple statement like `rnorm(mean=mu, sd=sigma)` would generate a random variable representing draws from the posterior predictive distribution of \( y \). A single line of code thus will in fact perform Monte Carlo integration of the joint density of \( (y^{\text{rep}}, \mu, \sigma) \), and draw from the resulting distribution \( p(y^{\text{rep}} | y) = \int \int \mathcal{N}(y^{\text{rep}} | \mu, \sigma)p(\mu, \sigma | y) \, d\mu \, d\sigma \). (We distinguish the observations \( y \) and the unobserved random variable \( y^{\text{rep}} \), which has the same conditional distribution as \( y \)).

**Summary**

Most of the work of writing a standard Gibbs/Metropolis sampler can be produced automatically; Umacs makes this possible by writing a customized sampler given only the updating functions or log-posterior functions relevant to the model. The user-defined parameters are embedded into standard looping structures and Metropolis updating routines, saving the trouble of writing the program from scratch. This saves time and makes debugging the sampler program easier.

Once posterior simulations are generated, it is awkward to work with the resulting inferences, display them graphically, generate posterior probability statements or generate predictions, since the inferences are in the form of numerical arrays of simulations and not accessible directly as random variables. The package ‘rv’ provides a new simulation-based random variable object class, which makes the job of manipulating and summarizing posterior inferences easier and provides the foundation of a “Bayesian programming environment.” Using random variable objects instead of arrays of simulations saves time and effort in writing—and understanding—program code.

We hope these packages will be useful and also will motivate future work by others, so that Bayesian inference can be performed in the interactive spirit of R.

**Acknowledgements**

We thank Tian Zheng, Shouhao Zhao, Yuejing Ding, and Donald Rubin for help with the various programs and the National Science Foundation for financial support.

**Bibliography**


WANTED: NEWSLETTER CO-EDITOR, STATS GRAPHICS

The Statistical Computing and Graphics Newsletter (SCGN) needs a new co-Editor on the Stat Graphics side. This is a great opportunity to serve the Statistical Graphics Section and the ASA in general. Co-editing it is a volunteer job with many rewards.

The Newsletter is a joint product of the Statistical Computing and Statistical Sections of the ASA, hence having two editors, one for Stats Computing and another for Graphics. There are two issues per year: one in the Fall and one in the Spring. The spring issue contains a lot of information about the upcoming ASA meetings, other meetings sponsored by the two sections, announcements of the competition awards and feature articles that anticipate future trends in Stats Computing and Graphics. The Fall issue talks about what happened in those past meetings, announces the competitions and also contains feature articles of high interest. Both the Fall and the Spring issues contain other interesting news and the Chair's columns plus some special columns, depending on availability of contributions for them.

The Editors of SCGN select contributions from different authors after extensive review and decide the final contents of the newsletter and what format the newsletter will have. They follow up on authors to guarantee a timely delivery once their article is accepted, collect news, gather columns from contributors and make sure that everything is done in a timely fashion and appropriately. All this material is then edited and entered into a newsletter semi-template (currently in Pages, a product of Apple's iWorks, but not necessarily so for ever). After the Executive Committees of both sections have approved, and the authors have proofed their pieces, the Newsletter is then posted online and Section members are notified that it is ready. Lately we have also been sending a postcard through regular mail, and will continue to do so.

This is a volunteer job with lots of room for creativity and for making the ASA sections you are part of visible to a wider group of statisticians.

If you are interested in becoming a Co-editor, please contact the Statistics Graphics Chair, Paul Murrell by email. His email address is p.murrell@auckland.ac.nz
Teaching Statistical Graphics

AN EXPERIENCE BETWEEN DUSTY ARCHIVES AND BLOGS

Heike Hofmann, Dianne Cook, Charles Kostelnick, Iowa State University

hofmann@iastate.edu
dicook@iastate.edu
chkostel@iastate.edu

1. Introduction

Three of us (two Statisticians and one English Professor) find ourselves facing a class of eight.

With this kind of teacher/student ratio we probably have one of the smallest classes on campus (the Iowa State campus, that is) and it would not have been possible if it were not for a special initiative of the College of Liberal Arts and Sciences to fund interdisciplinary teaching. The exact name of the course is “Visual communication of quantitative information”. We are offering the course on an undergraduate level with the option for graduate credit requiring additional and more independent work from students. The background of our students is diverse: six undergraduates with majors as different as advertising, technical writing, and geography and two grad students majoring in HCI and geography.

2. Syllabus

The objective of the course is to help prepare students to be active citizens in the information technology age. Students will develop critical thinking skills about how information is visually presented, and they will learn how to accurately and attractively communicate quantitative information using graphics. At the end of the course students will:

✓ know about important historical and contemporary examples,
✓ know about and how to implement the elements of graphical design,
✓ be able to evaluate visual presentations of information in the media, and
✓ be able to use the computer to generate graphics to communicate information effectively.

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Text Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>Week 1</td>
<td>What is this course about? Historical and famous examples</td>
<td>1.1, 1.2, 1.3, 1.5</td>
</tr>
<tr>
<td>Week 2</td>
<td>Communication purpose graphics</td>
<td></td>
</tr>
<tr>
<td>Week 3</td>
<td>Types of quantitative information Basic data plots</td>
<td>1.4, 3</td>
</tr>
<tr>
<td>Week 4</td>
<td>Visual conventions Graphical elements: points, lines, color</td>
<td>2.3, 4</td>
</tr>
<tr>
<td>Week 5</td>
<td>Graphical perception and misconceptions</td>
<td>6</td>
</tr>
<tr>
<td>Weeks 6,7</td>
<td>Interacting with graphics</td>
<td>5, 7</td>
</tr>
<tr>
<td>Week 8</td>
<td>Ethics, privacy, public data</td>
<td></td>
</tr>
<tr>
<td>Week 9</td>
<td>Presenting tables</td>
<td>2,2</td>
</tr>
<tr>
<td>Week 10</td>
<td>Geographic visualization, cartography</td>
<td></td>
</tr>
<tr>
<td>Week 11</td>
<td>Text visualization</td>
<td>10</td>
</tr>
<tr>
<td>Week 12</td>
<td>Network visualization</td>
<td>8</td>
</tr>
<tr>
<td>Week 13</td>
<td>Scientific visualization</td>
<td></td>
</tr>
<tr>
<td>Week 14</td>
<td>Contemporary Examples</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Tentative Syllabus of Course Materials

3. Resources

3.1. Books

There are many very attractive books on statistical graphics, all of which we liked for different reasons, but there was not one single book we wanted to use as a textbook.

For lack of the perfect book, we found ourselves drawing from many different sources. Here our favorites:


This is the book we picked for a textbook, and on which we based our tentative schedule of materials (see table 1). Unfortunately, the second edition went just out of print, when we started teaching. This caused a bit of confusion and delay until everybody had a copy...
in hand.
Edward R. Tufte (2001): “The Visual Display of Quantitative Information” - probably the book - the classical reference and best known book in the area. For a textbook, particularly on an undergraduate level, there are too few details and no exercises to work through. While reading excerpts, our students also complained a bit about the avuncular style of writing.
Naomi Robbins (2004): “Creating More Effective Graphs” did not appear until after the course started, but we'll keep it in mind for the next time.
Colin Ware (2004): “Information Visualization: Perception for Design” - a good reference for a perspective from the InfoVis community. In book form, this provides the most recent summary of perceptual cognition results, including some of the fundamental Gestalt theory.

3.2 Online Resources
The Statistical Atlases of the United States from 1870, 1880, and 1890 based on the ninth, tenth, and eleventh census, which appear online on websites offered by the Geography and Map Division of the US Library of Congress (http://memory.loc.gov/ammem/browse/).

Figure 1. Revenues and expenditures of the fiscal years of the United States of America between 1789 and 1890.

These atlases not only provide insight into American history but they also provide a rich resource of beautiful and astonishingly modern visualization techniques. Figure 1 shows a juxtaposition of revenues and expenditures of the United States budget for the years 1789 to 1890. Both revenues and expenditures are split according to their principal components. On the expenditure side the higher expenses of the army for the War of 1812-15 between the United States and Great Britain, the Mexican-American War of 1846-48, and the American Civil War of 1861-65 can clearly be seen.

Michael Friendly & Daniel J. Denis: Milestones in, the History of Thematic Cartography, Statistical Graphics, and Data Visualization
By now this is another classic for statistical graphics at http://www.math.yorku.ca/SC. The website gives a detailed chronology of innovations with the highlights of historical charts and maps.

Michael Bach: Optical Illusions & Visual Phenomena
an award-winning site with a huge collection of examples in perceptual (mis-)conceptions and detailed explanations of why and how. The site makes great use of interactive capabilities of the WWW by providing movies and dynamic gifs for perceptual issues in motion: http://www.michaelbach.de/ot/. Figure 2 shows the strong (and beautiful) rotation of the “wheels” occurs in relation to eye movements. On steady fixation, the effect disappears.

Figure 2. The rotating snake: strong rotation of the “wheels” occurs in relation to eye movements. On steady fixation, the effect vanishes.

Info Aesthetics is one of the top BLOGs with the newest examples in information visualization. As a
reference it might not be suitable because of the varied quality of the submissions, but it is a great source for novel ideas and nifty visualizations. Figure 3 shows a stand-alone data visualization application by digitectonics.com that monitors a stock portfolio.

**Figure 3.** Flower garden: visualization of stocks; number of petals, color, direction, and height are used to encode properties of the stock.

Different flowers represent the real-time performance of selected stocks via the color, height, & radian of animated blossom flowers. A flower grows from the bottom (ground) & stops at the height, reflecting its share price: the higher the stock price, the higher the ‘flower stem’. As soon as it reaches the top, it begins to blossom fan-wise to the degree that reflects the percentage of price change. The color (green or red) & direction (upward or downward) of a blossom indicate a particular stock’s status of ascent or descent in price compared to its previous trading day. A flying bee will show up around a flower if there is recent news of that particular stock.

**4. Assignments & Projects**

**Discussing good and bad graphics**

Being able to critically assess presented information does not seem to be a huge problem for somebody in academia. But it is a major step for a freshman from noddingly accepting facts to questioning them or even suggesting improvements in their presentation. Getting everybody in the course to critically read graphics was one of the major goals of our class. The first assignment was a major reality check for us, as we mainly faced answers that could be summarized by “I like the graphic, the data is clearly presented, there is nothing obviously wrong with the display.”

**Meeting the Newspaper Guys:**

This assignment entailed designing charts for an article in the *ISU Daily* on student housing.

Overall, this was a very informative but strenuous experience. Due to the very tight deadline of one weekend, our students were exposed to considerable pressure. With the result being published in the *ISU Daily* it was well worth the effort, and we were left with the satisfying feeling of having achieved something “real”. Figure 4 shows a comparison of on- and off-campus housing rates in two separate barcharts with a check mark plot showing available amenities in on-campus housing.

**Figure 4.** Charts our students designed comparing housing rates, published in the *ISU Daily* on Mar 1st, 2006.

**Making a poster**

Here we used the data and set-up of the *InfoVis challenge 2006*, which dealt with a 1% sample of the US Census of 2000, i.e. we have data, and more of it than everybody involved ever wished for ...

**5. Conclusions**

The course was highly successful, partly because it was very rewarding to see the final presentations by the students. We will definitely offer the course again in the future. What will be different? Next time, we will use assignments earlier on to involve students more
actively from the start of the class. One problem crystallizing at the end of the course turned out to be technical issues: lack of computing/ software skills which resulted in a good deal of frustration, where students did know exactly what they wanted to do, but did not know how to do it with the software. To alleviate this, we were planning on having more tutorials at the beginning of the next course. Back in 1962 John W. Tukey discussed the difficulties of teaching data analysis to students. One of the main problems he identified was the unwillingness of teachers to say “I don’t know”. Interestingly, students do not like hearing this phrase from a teacher either, as came up during the final evaluation of the course. However, facing the unknown, vague, or unexpected should be part of an academic education and lead to a scientific and questioning mindset. Teaching statistical graphics is not easy and is not done in many universities. As a tool of communication, however, graphics are essential in our everyday lives, and we believe that students in virtually any discipline are well served by a course like ours.

References
Spence, R. (2001), Information Visualization, Addison-Wesley.

DATA EXPO 2006 WINNERS

A highlight of the joint statistical meetings in Seattle this August 2006 was the data expo, held for the first time in many years. You must visit the expo website to get a good feeling for the amount of work that went into it and to see all the posters presented, which Paul Murrell skillfully uploaded to the web site for a spectacular display of talent and promising work. The web site is


There were 14 poster entries presented in a topic contributed data expo poster session on Mo 08/07/2006, 10:30 AM to 12:20 PM. You can take a close look at these posters on the web site. The FIRST PRIZE: “Visualizing Several Abnormal Climate Changes in Central America from January 1995–December 2000” by Sang-Hoon Cho, University of Wisconsin-Madison; Hyonho Chun, University of Wisconsin-Madison; Deepayan Sarkar, State University of Wisconsin. The first prize, which consists of $1000 cash plus a set of NASA books. We see them below with their poster and receiving the award during the Stats Computing and Graphics Reception.
Tools for Multivariate Data Visualization

ABOUT GLYPHS AND SMALL MULTIPLES: GAUGUIN AND THE EXPO

Alexander Gribov, Antony Unwin, Heike Hofmann
1 Augsburg University, 2 Iowa State University
alex967600@freenet.de
unwin@math.uni-augsburg.de
hofmann@iastate.edu

1. Introduction

Glyphs are defined as geometric shapes scaled by the values of multivariate data. Each glyph represents one high-dimensional data point (or sometimes the average of a group of data points). Their best-known representatives are probably Chernoff faces (Chernoff, 1973), which can also be found in extended versions (Flury and Riedwyl, 1981; Bruckner, 1978; Huff and Black, 1978). Various studies have assessed Chernoff faces with respect to their information richness and visualization capabilities (Lee et al., 2003; Healey, 1996; Chernoff and Rizvi, 1975; Morris et al., 1999) - surprisingly not always unfavorably.

![Figure 1. Three simple examples for star glyphs. Each glyph represents ten numbers between 0 and 10 as shown in the titles](image)

Besides faces, other glyphs exist - star glyphs are widely used: a star glyph of a p dimensional point has p axes, i.e. (half- ) lines coming out of the origin at uniformly separated angles. Values are drawn as endpoints of each of these axes. All of those endpoints are connected to form a polygon (Siegel et al., 1972; Ward, 1994), see figure 1. For technical purposes we can also think of a star glyph as a parallel coordinate plot in polar coordinates. In that way glyphs fit the description of small multiples, introduced by Tufte: Small multiples are “illustration of postage-stamp size [...] indexed by category or label, sequenced over time like the frames of a movie, or ordered by a quantitative variable not used in the single image itself. Information slices are positioned within the eyespan, so that viewers make comparisons at a glance—uninterrupted visual reasoning. Constancy of design puts the emphasis on changes in data, not changes in data frames” (Tufte, 1990), page 67. (Insisting that an ordering variable is not used in the individual images fits in with Tufte’s general principles, but is not a requirement. Sometimes a little information redundancy can be helpful.)

2. The GAUGUIN Software

GAUGUIN (Grouping And Using Glyphs Uncovering Individual Nuances) is a project for the interactive visual exploration of multivariate data sets. It supports a variety of methods for displaying flat-form data and hierarchically clustered data. The prime aim of the project is to add interactive capabilities to glyph representations. Many factors influence the interpretation of glyph visualizations, including the form of glyph chosen, which variables are included, the axis ordering within each glyph, glyph size, and the ordering of glyphs in the display. Being able to vary these flexibly and smoothly is essential to get the most information from the data.

![Figure 2. Four different types of glyphs as implemented in GAUGUIN. From upper right to lower bottom there are filled and unfilled line glyphs and filled and unfilled star glyphs.](image)
GAUGUIN offers four different glyph shapes (see figure 2). The number of data elements which can be displayed simultaneously is limited, because each glyph requires a minimum amount of screen space to be viewed. Hierarchical glyphs can be drawn for groups or clusters of cases. They are composed of a highlighted case representing the group and a band around it showing the variability of the members of that group. GAUGUIN also provides scatterplots and tableplots, and via Rserve (Urbanek, 2003) is able to use R to calculate MDS views and clusters for the data. All GAUGUIN displays are linked interactively and can be directly queried. More information can be found on the project’s website.

3. Data Expo 2006 - An Example
The data for this example come from the Data Exposition 2006 sponsored jointly by ASA’s sections on Statistical Graphics, Statistical Computing, and Statistics and the Environment. Measurements on ozone, pressure, temperature, surface temperature and cloud development at low, medium and high altitude were taken on a grid of 24 by 24 points over Central America between a latitude of 55.5 W and 114.5 W and longitude between 36.5 N and 21.5 S (see figure 3). All variables except elevation are reported at 72 points in time; once per month from Jan 1995 to Dec 2000. Elevation is reported for Jan 1998. All figures in this section (if not indicated otherwise) have been done using GAUGUIN. The advantage of using R for figures 5 and 6 in the paper is the access to axes and titles in the presentation form of the charts.

Figure 3. Google’s hybrid map of the investigated area (left). Elevation data on a topographical color scale is used to give the picture on the right, the 24 x 24 points of observations are marked by “+”s. The four large black dots show the locations of the values shown in figure 6.

A simple first example: Pressure
Monthly measurements for pressure are very stable, as the monthly averages of pressure are only slightly influenced by local occurrences of high and low pressure. Other than that, pressure depends almost solely on the altitude of a location. For a location at an altitude of h meters above sea level, the approximate average pressure is given as 1013 (h/1500)/100 mbar. At sea level, this gives a pressure of approximately 1000 mbar.

Figure 4. Pressure data for each location as glyphs. Glyphs are drawn according to their spatial location as given by Latitude and Longitude. Coloring stems from a hierarchical clustering using Ward’s method with 5 clusters. With increasing altitude glyphs tend to become smaller. High altitude locations show larger variability. A large “jump” in pressure occurs about half way through the recorded time (between May and June 1998) hinting at a change in the way measurements were taken. On the right is a summary of all five clusters in pressure values. Two clusters show no changes in pressure, one shows a small jump, while the remaining two have a large jump. The clusters with a jump show more variability.

Figure 4 shows the pressure measurements of each location as a glyph. Glyphs are ordered according to the location’s Latitude and Longitude. The different colors are given by a hierarchical clustering of the pressure data (Ward’s method with 5 clusters). A geographical pattern appears in the clusters. As expected, the overall glyph size shrinks with an increase in altitude (i.e. pressure goes down from 1000 mbar at sea level, to approx. 928 mbar at 500 m above sea and 861 mbar at 1000 m above sea level). What might have been expected is the larger variability in pressure measurements at higher locations (mountains...
are notorious for their weather instability - this might have an effect on monthly averages). Completely unexpected is the large jump in pressure values in many locations half way through the recording period. Between May and June 1998 pressure values change drastically in 41 locations and change by an average of about 50 mbar in another 45 locations (see right hand side of figure 4 for a summary).

![Diagram](image)

**Figure 5.** Average monthly temperatures at a single location between January 1995 and December 2000 (R chart)

**Figure 6.** Star plots for four locations in Central America. From left to right the locations are in the Rocky Mountains, the Pacific Ocean close to the equator, and twice in the Andes. The upper row shows star plots with 72 axes. The lower row has twelve axes, one for each month and the six years are drawn on top of one another (R chart)

**Temperature: small multiples of small multiples**

Figure 5 shows a star glyph with six years of monthly temperature data. The oldest data are darkest, more recent measurements are brighter in color. The flower-like shape is produced by the strong seasonal element in the data, with highest temperatures in July and August and lowest temperatures in January. The scale is the same for every one of the 72 axes. Minimum and maximum are set to 269.0 K and 310.0 K. Figures 5 and 6 have been drawn in expanded form in R to display their underlying structure with all the axes marked and some labeling added. Figure 6 shows average monthly temperatures at four locations in Central America. The data of all locations are shown twice: a star glyph with 72 axes is drawn in the top row, while the stars in the bottom row only have 12 axes - one for each month, so the data cover six cycles. The location on the left is in the Rocky Mountains of North America. The dominant feature in temperature is the strong seasonality with low temperatures in January and high temperatures in July and August. The second plot shows a location in the Pacific close to the equator. A slight seasonality is visible, but mainly temperatures are very steady. The temperature during the end of 1998 remains high until Fall of 1998, when it drops back to normal. The third plot shows temperatures at a location in the Andes of South America. No strong seasonal effects are visible, but the overall average temperature increases steadily. The fourth plot shows a location in South America - the seasonality is reversed from the Northern hemisphere with highest temperatures reached in January. The row of plots at the bottom shows glyphs of the same locations with temperature cycles of 12 months. More recent years are colored lighter. Seasonality is now coded in relation of the lines to the origin (black dot in the middle). The increase in temperature at the third location can be seen by the spiral-like shape of the glyph with predominantly darker colors in the inner and lighter colors at the outer locations of the glyph.

Figure 7 shows glyphs of temperature measurements for each location ordered spatially. Even though glyphs usually do not provide good global summaries of trends or relationships (Lee et al., 2003), the data in this example can be summarized nicely: the flower shapes in the top rows indicate strong seasonal variations in temperatures, the bottom rows show similar flowers - but are rotated by 30 degrees. This translates to the (hopefully familiar concept of ) warm temperatures in the southern hemisphere during the cold season in the northern hemisphere and vice versa. Temperatures closer to the equator are more level with only slight seasonal effects. Interesting, but a bit hard to spot, is the Pacific area just south of the equator: the flowers
indicate some seasonality, but the petals in the north east corner of the flower cannot be separated easily. This means that temperatures did not go down substantially between these two summers - the clear sign of an El Nino event.

![Figure 7](image)

**Figure 7.** Glyphs of monthly average temperature measurements over six years ordered by Latitude and Longitude.

Adding colors emphasizes similarities drastically (Ware, 2004). Figure 8 shows a set of eight glyph plots of the same temperature values. The colors discriminate between 2 clusters (top left) and 9 clusters (bottom right). Accompanying cluster views show glyphs of the variability within clusters. With increasing number of clusters the within cluster variability decreases. For six and more clusters cluster sizes become fairly small.

![Figure 8](image)

**Figure 8.** Small multiples of small multiples: eight temperature glyph plots are colored according to clusterings with increasing numbers of clusters (from k=2, top left to k=9, bottom right). Above and below the star glyph charts are summary plots of the clusters, showing color and variability within each cluster. With an increasing number of clusters the variability within clusters decreases.

![Figure 9](image)

**Figure 9.** Glyph dendrogram corresponding to the hierarchical clustering in the previous figure.
4. Summary
All graphic displays benefit from being made interactive. GAUGUIN includes querying, zooming, a choice of glyph forms, selection and reordering of variables, case selection and linking, all in an interactive form. Applying these tools to the Expo dataset led to the results reported in this paper. Although the interactive flexibility and power cannot be shown here in a printed form, they were important for the analysis process. It was particularly interesting how effective glyphs could be, even in this relatively large dataset.

References

Don’t forget....

56th Session of the ISI
INTERNATIONAL STATISTICAL INSTITUTE
22 - 29 AUG, Lisboa 2007
Book Reviews

“CREATING MORE EFFECTIVE GRAPHCS”
NAOMI ROBBINS
WILEY 2005

This is an elegant, efficient guide to common-sense plotting of data. It is written in an easy to follow format, with no-nonsense plain language. Poor plot examples are clearly marked by a Stop sign bearing the words “Not recommended”.

Everyone thinks that they know how to plot data. It’s a “G” rated activity. If it’s so easy why is it that the plots statisticians make are almost always abominable? Is it the responsibility of the software producer to make default plots adhere to the now well-studied aesthetics and perceptual guidelines, so that any dummy has a good chance of producing appropriate data plots? Or should statisticians be educated in the fundamentals of plotting data?

Naomi Robbins’ book is the perfect solution to educating statisticians. The chapters in the book follow this sequence:

★ What is an effective graph?
★ What are the ubiquitous problems prevailing in currently populist plots?
★ Tasks that we need to perform to decode a plot, and which ones we humans do well.
★ More effective graphs than those in common use.
★ Graphical principles.

The book also shows before and after examples, demonstrating what the reader should have learned from the previous chapters. A question and answer chapter addresses common questions about graphs.

Every statistician should have a copy on their shelves. Undergraduate statistics majors and graduate students in statistics should be encouraged to use this book in their studies on statistics

Reviewer: Dianne Cook

News

ANNUAL COMPETITIONS
CALL FOR ENTRIES
JR Lockwood,
Awards Officer, 2007
Statistical Computing Section

Student paper competition 2007

The Statistical Computing and Statistical Graphics Sections of the ASA are co-sponsoring a student paper competition on the topics of Statistical Computing and Statistical Graphics. Students are encouraged to submit a paper in one of these areas, which might be original methodological research, some novel computing or graphical application in statistics, or any other suitable contribution (for example, a software-related project). The selected winners will present their papers in a topic-contributed session at the 2007 Joint Statistical Meetings. The Sections will pay registration fees for the winners as well as a substantial allowance for transportation to the meetings and lodging (which in most cases covers these expenses completely).

Anyone who is a student (graduate or undergraduate) on or after September 1, 2006 is eligible to participate. An entry must include an abstract, a six page manuscript (including figures, tables and references), a blinded version of the manuscript (with no authors and no references that easily lead to identifying the authors), a C.V., and a letter from a faculty member familiar with the student’s work. The applicant must be the first author of the paper. The faculty letter must include a verification of the applicant's student status and, in the case of joint authorship, should indicate what fraction of the contribution is attributable to the applicant. We prefer that electronic submissions of papers be in Postscript or PDF. All materials must be in English.

All application materials MUST BE RECEIVED by 5:00 PM EST, Monday, December 18, 2006 at the address below. They will be reviewed by the Student Paper Competition Award committee of the Statistical Computing and Graphics Sections. The selection criteria used by the committee will include innovation.
and significance of the contribution. Award announcements will be made in late January, 2007.

Additional important information on the competition can be accessed on the website of the Statistical Computing Section, www.statcomputing.org. A current pointer to the website is available from the ASA website at www.amstat.org. Inquiries and application materials should be emailed or mailed to:

Student Paper Competition,
c/o J.R. Lockwood
The RAND Corporation,
4570 Fifth Avenue, Suite 600
Pittsburgh, PA 15213
lockwood@rand.org

John M. Chambers Statistical Software Award, 2007

The Statistical Computing Section of the American Statistical Association announces the competition for the John M. Chambers Statistical Software Award. In 1998 the Association for Computing Machinery presented its Software System Award to John Chambers for the design and development of S. Dr. Chambers generously donated his award to the Statistical Computing Section to endow an annual prize for statistical software written by an undergraduate or graduate student. The prize carries with it a cash award of $1000, plus a substantial allowance for travel to the annual Joint Statistical Meetings where the award will be presented.

Teams of up to 3 people can participate in the competition, with the cash award being split among team members. The travel allowance will be given to just one individual in the team, who will be presented the award at JSM. To be eligible, the team must have designed and implemented a piece of statistical software. The individual within the team indicated to receive the travel allowance must have begun the development while a student, and must either currently be a student, or have completed all requirements for her/his last degree after January 1, 2004. To apply for the award, teams must provide the following materials:

Current CV’s of all team members.

A letter from a faculty mentor at the academic institution of the individual indicated to receive the travel award. The letter should confirm that the individual had substantial participation in the development of the software, certify her/his student status when the software began to be developed (and either the current student status or the date of degree completion), and briefly discuss the importance of the software to statistical practice.

A brief, one to two page description of the software, summarizing what it does, how it does it, and why it is an important contribution. If the team member competing for the travel allowance has continued developing the software after finishing her/his studies, the description should indicate what was developed when the individual was a student and what has been added since.

Access to the software by the award committee for their use on inputs of their choosing. Access to the software can consist of an executable file, Web-based access, macro code, or other appropriate form. Access should be accompanied by enough information to allow the judges to effectively use and evaluate the software (including its design considerations.) This information can be provided in a variety of ways, including but not limited to a user manual (paper or electronic), a paper, a URL, online help to the system, and source code. In particular, the entrant must be prepared to provide complete source code for inspection by the committee if requested.

All materials must be in English. We prefer that electronic text be submitted in Postscript or PDF. The entries will be judged on a variety of dimensions, including the importance and relevance for statistical practice of the tasks performed by the software, ease of use, clarity of description, elegance and availability for use by the statistical community. Preference will be given to those entries that are grounded in software design rather than calculation. The decision of the award committee is final.

All application materials must be received by 5:00pm EST, Monday, February 26, 2007 at the address below.
The winner will be announced in May and the award will be given at the 2007 Joint Statistical Meetings.

Information on the competition can also be accessed on the website of the Statistical Computing Section (www.statcomputing.org or see the ASA website, www.amstat.org for a pointer), including the names and contributions of previous winners. Inquiries and application materials should be emailed or mailed to:

Chambers Software Award
c/o J.R. Lockwood
The RAND Corporation
4570 Fifth Avenue, Suite 600
Pittsburgh, PA 15213
lockwood@rand.org

ANNUAL COMPETITIONS

2006 WINNERS

J.R. Lockwood, Awards Officer, 2007
Statistical Computing Section

The Statistical Computing Section of ASA sponsors three annual competitions aimed at promoting the development and dissemination of novel statistical computing methods and tools: the Student Paper competition (jointly with the Statistics Graphics Section), the John M. Chambers Award, and the Best Contributed Paper competition. Winners of all three awards are selected prior to the Joint Statistical Meetings (JSM), being officially announced at the Monday night business meeting of the Statistical Computing and Statistical Graphics Sections at JSM.

The Student Paper competition is open to all who are registered as a student (undergraduate or graduate) on or after September 1st of the previous year when the results are announced. Details on submission requirements are provided in the competition’s announcement, which goes out in mid to late September, at the Statistical Computing website at http://www.statcomputing.org and also in the News section of this newsletter. The four winners of the Student Paper competition are selected by a panel of judges formed by the Council of Sections Representatives (COS-REPs) of the Statistical Computing and Statistical Graphics Sections, who work hard to get the results announced by the last week of January. As part of the award, the winners receive a plaque, have their JSM registration covered by the sponsoring sections and are reimbursed up to US$ 1,000 for their travel and housing expenses to attend the meetings. The winning papers are presented at a special Topics Contributed session at JSM, which typically takes place on Tuesday. The winners of the 2006 Student Paper competition, presented in alphabetical order, were:

- **Youjuan Li**, University of Michigan-Ann Arbor (advisor: Ji Zhu) “Efficient Computation and Variable Selection for the L1-norm Quantile Regression”

- **Fan Lu**, University of Wisconsin-Madison (advisor: Grace Wahba) “Kernel Regularization and Dimension Reduction”

- **Rebecca Nugent**, University of Washington-Seattle (advisor: Werner Stuetzle) “Clustering with Confidence”

- **Philip Reiss**, Columbia University (advisor: Todd Ogden) “An Algorithm for Regression of Scalars on Images”
in this newsletter, News section. The prize includes a plaque, a cash award of US$ 1,000, plus a US$ 1,000 allowance for travel and hotel expenses to attend JSM (with registration fee covered by the section.) The winner of the 2006 John M Chambers Award was:

- Hadley Wickham, Iowa State University (advisors: Di Cook and Heike Hofmann) “ggplot and reshape: Practical Tools for Organizing, Summarizing, and Displaying Data” (http://had.co.nz/jca2006)

Finally, the Best Contributed Paper award is determined on the basis of the evaluations filled out by the attendees of the Contributed and Topics Contributed sessions of JSM which have the Statistical Computing Section as first sponsor. All presenters in those sessions are automatically entered in the competition. The prize includes a US$ 100 cash award and a plaque. The winner of the 2005 Best Contributed Paper Award is

- Heather Turner, Research Fellow at the Department of Statistics, University of Warwick, UK, (jointly with David Firth, from the same department) for the paper “Multiplicative Interaction Models in R” in the session “Algorithms and Software”.

Last, but not least, after three years of dedicated service that ensured the success of these competitions, Jose Pinheiro passed the responsibilities to me (J.R. Lockwood from the RAND Corporation) who will serve a three-year term that began in September 2006. I look forward to my term of service and hope that I get to know many of you over the coming years.
Statistical Computing Section Officers 2006

Stephan R. Sain, Chair
ssain@math.cudenver.edu
(303)556-8463
John F. Monahan, Chair-Elect
monahan@stat.ncsu.edu
(919)515-1917
Tim Hesterberg, Past-Chair
timh@insightful.com
(206)802-2319
Michael Trosset, Program Chair
trosset@math.wm.edu
(757)221-2040
Ed Wegman, Program Chair-Elect
ewegman@gmu.edu
(703)993-1691
David J. Poole, Secretary/Treasurer
poole@research.att.com
(973)360-7337
Vincent Carey, COS Rep. 05-07
stycz@channing.harvard.edu
(617)325-2265
Robert Gentleman, COS Rep. 04-06
gentlem@hsph.harvard.edu
(617)632-5250
Juana Sanchez, COS Rep. 06-08
and Newsletter Editor
jsanchez@stat.ucla.edu
(310)825-1318
Thomas F. Devlin, Electronic Communication Liaison
devlin@mozart.montclair.edu
(973)655-7244
J.R. Lockwood, Awards Officer
lockwood@rand.org
412-683-2300-Ext 4941
R. Todd Ogden, Publications Officer
ogden@cpmc.columbia.edu
212-543-6715
John J. Miller, Continuing Education Liaison
jmiller@gmu.edu
(703)993-1690

Statistical Graphics Section Officers 2006

Paul R. Murrell, Chair
paul@stat.auckland.ac.nz
(649)373-7599 x85392
Jeffrey L. Solka, Chair-Elect
jeffrey.solka@navy.mil
(540)653-1982
Mario Peruggia, Past Chair
peruggia@stat.ohio-state.edu
(614)292-0963
Juergen Symanzik, Program Chair
juergen.symanzik@usu.edu
(435)797-0696
Simon Urbanek, Program Chair-Elect
urbanek@research.att.com
(973)360-7056
John Castelloe, Secretary-Treasurer
john.castelloe@sas.com
(919)677-8000
Daniel B. Carr, COS Rep 05-07
dcarr@gmu.edu
(703)993-1671
Edward J. Wegman, COS Rep 05-07
ewegman@galaxy.gmu.edu
(703)993-1680
Naomi B. Robbins, COS Rep 04-06
naomi@nbr-graphs.com
(973)694-2686
Dianne Cook, Newsletter Editor
dicook@iastate.edu
(515)294-8865
Linda Williams Pickle, Publications Officer
picklel@mail.nih.gov
(301)402-9344
Monica D. Clark, ASA Staff Liaison
monica@amstat.org
(703)684-1221

The Statistical Computing & Statistical Graphics Newsletter is a publication of the Statistical Computing and Statistical Graphics Sections of the ASA. Until a new Co-editor for the Statistical Graphics Section comes in to replace Di Cook, all communications regarding the publication should be addressed to:

Juana Sanchez, Editor Statistical Computing Section. Department of Statistics University of California, 8125 MS Building, Los Angeles, CA 90095 (310)825-1218 jsanchez@stat.ucla.edu www.stat.ucla.edu/~jsanchez

All communications regarding ASA membership and the Statistical Computing and Statistical Graphics Section, 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