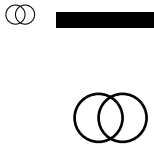
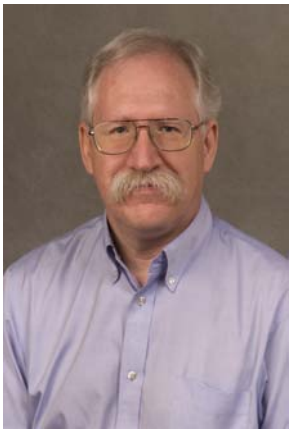
Volume 21, June 2010



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

Statistical COMPUTING & GRAPHICS

A Word from our 2010 Section Chairs



LUKE TIERNEY
COMPUTING

These are interesting times for computing. With processor speeds no longer increasing significantly, parallel computing in various forms is receiving a lot of attention as the way forward for improving software performance. Current processors feature multiple cores, and it has become common to see questions on mailing lists asking "how can we get our software to use all available cores."

Continued on page 2 . . .

SIMON URBANEK
GRAPHICS

The future has always been exciting in graphics as new technologies have allowed us to visualize and understand data in new ways, better and faster. From fast computers that allowed us to make graphics interactive to new technologies such as the Web that have allowed us to make visualizations ubiquitous. It is always exciting to find ways how to leverage the newest–

Continued on page 2 . . .

Contents of this volume:

Computing Chair
Continued from page 1.

The answer to this question is easy, but the answer to the real question, "how can we get our programs to take advantage of multiple cores and run faster than on a single core," is not, and will be a challenge for those writing statistical software for a number of years to come.

Another approach to parallel computing that has received a lot of attention recently within the general scientific computing community as well as within the statistical computing field is the use of computing capabilities available in graphics processors, or GPUs. This is leading to the curious phenomenon of racks of computers, each with a powerful graphics processor, and not a monitor in sight. A number of papers are starting to appear on the great performance improvements that are possible using GPUs at least for certain special problems. GPU computing is more challenging in a number of respects than multi-core parallel computing, and it is too early to tell whether this will be an approach with long term viability and more general applicability, but it is an interesting area well worth keeping an eye on.

A third interesting area to follow is cloud computing. This phrase is used to describe a number of different things, but one variant is to allow users to rent access to one or more virtual computers accessed through the Internet that can be used to carry out a computation much as one would on a cluster but without the need to own and manage the cluster hardware. This approach is very attractive for some applications, but many issues important to statisticians, including data safety issues, are still open and need to be kept in mind when considering this approach. But once again there are a number of statistical computing projects exploring this approach. As I said at the outset, these are interesting times and we will see where it all leads.

Our Program Chair Thomas Lumley has put together an excellent program for JSM 2010 in Vancouver, including five invited sessions sponsored by the Section, eight more jointly sponsored invited sessions, five topic contributed sessions, 13 contributed sessions, and one contributed poster session. Several of these sessions, as well as a topic contributed session on grid computing, highlight parallel computing issues.

Other notable meetings this summer include the Interface meeting in June in Seattle and the UseR! 2010 meeting in July in Gaithersburgh — the second UseR! meeting to be held in North America. Both meetings highlight issues with the analysis of very large data sets, and the UseR! 2010 meeting in particular features a number of exciting sessions and presentations on parallel computing in statistics.

It is not too early to start thinking about JSM 2011 in Miami Beach. If you have ideas for invited sessions or topic contributed sessions our Program Chair Elect David Poole would like to hear about them. Also consider offering a Continuing Education course or Technology Workshop — these are always very much appreciated by our membership.

Luke Tierney
University of Iowa

Graphics Chair
Continued from page 1.

developments in computing to our advantage be it in the form of technological advances pushing the barriers of data visualization or in the form of emerging trends such as social networks and collaborative visualization.

Undoubtedly, the race to higher speeds has been pushed further especially in computer graphics allowing complex rendering of large data on commodity computers - which leaves us with the challenge how to use it for the best for sensible data analysis. There is also the excitement of designing interactive graphics with the rich tactile human-computer interaction interfaces enabled by the emerging multi-touch devices.

However, I think the new excitement this time is not limited to new technologies, devices or paradigms – it is the immense growth of the available data sources both in size and variety. The data flood is coming, fueled by the omnipresence of personal devices, gathering loads of data from all the embedded sensors - location, acceleration, compass, temperatures, and many others. The decrease of the relative cost of data acquisition, the massive spread of smart devices as well as the decreasing cost of data storage has given way to an almost torrential accumulation of data that just

screams to be visualized and analyzed - a great opportunity for us.

Despite the technological advances this is not simply a question of scalability. It means an entire paradigm shift as we can no longer repeatedly rely on the 'iris' dataset to demonstrate our visualization techniques. We have to find new ways to provide various insights into the emerging massive and streaming data, going beyond incremental improvements. This gives us the unique opportunity to think out of the box and try new visions in graphics.

Our JSM program is strong this year thanks to our Program Chair Heike Hofmann and session organizers. It reflects the new directions of emerging data – from the 'Quantified Self' invited session on massive personal data, Data Expo session as a follow up to the large data visualization challenge, recent advances and future direction in R graphics, visualizing high-dimensional data up to several cutting-edge contributed and poster sessions. We have also very interesting co-sponsored sessions with focus on visualization. Finally, we are most pleased to announce a special session designated for the presentation of The Statistical Computing and Graphics Award.

Always looking into the future we are already gearing up for the topics to be presented at the next year's JSM in Miami Beach so please feel free to send ideas and suggestions for sessions to our Program Chair-elect Webster West.

Simon Urbanek
AT&T Research

Highlights from the Joint Statistical Meetings

Stat Computing program

The Statistical Computing and Statistical Graphics sections join to present the inaugural Statistical Computing and Graphics Award, to Robert Gentleman and Ross Ihaka, for the creation of R; the session will include presentations by both recipients. Other highlights of the Computing invited program include a session on the Netflix Prize, organized by Chris Volinsky; and a session on models for complex networks, introducing the SAMSI research program for the coming year. Topic contributed sessions in Computing (and Graphics) include the annual student paper competition. We have also co-sponsored a number of sessions organized by other sections, such as a session on statistical issues in data from Big Science projects. For complete details of all Computing sessions, see the online program. Last, but not least, do not forget to attend the Computing and Graphics mixer on the Monday evening.

Thomas Lumley
University of Washington

Editor's Note

Nicholas Lewin-Koh (Computing)
Andreas Krause (Graphics)

This first issue of Volume 21 of our newsletter contains again some interesting articles.

Antony Unwin's paper on "Getting into hot water over hot graphics" is the result of an initial request to summarize the "hot" (or "cool") topics in the area of graphics. It summarizes neatly how graphics and research on graphics has evolved over the past few decades, providing thought-provoking comments from an insider.

If you wish to discuss this article, please drop us a note. We would be delighted if we could spark an interesting exchange of thoughts and perceptions as diverse as the interpretation of a graph.

Charles Roosen and Richard Pugh write about "Creative Techniques for Exploratory Data Graphics in PK/PD." You do not need to be in the field of pharmacokinetics/ pharmacodynamics to get inspiration for creativity when it comes to analyzing data.

The regular columns, "A word from our session chairs," feature insights from our new section chairs, Luke Tierney (Computing) and Simon Urbanek (Graphics).

We look forward to the upcoming Joint Statistical Meetings (JSM) in Vancouver and focus on the part organized by our sections.

All of these articles all provide interesting reading and we hope you enjoy them. If you have comments, please email the editors.

We are always looking for interesting contributions. Please contact us if you have short articles, software, graphs, or anything that you think might be of interest to the ASA Sections for Statistical Computing and Graphics.

Nicholas and Andreas

Change in Editorship

With this issue we welcome Martin Theus as the new editor for the statistical graphics section. Martin succeeds Andreas Krause who has been the graphics section editor since 2007.

Martin Theus is Senior Project Manager in the Analysis Center of the Business Intelligence Unit of Telefónica o2-Germany. His research and application areas are data visualization and data mining as well as exploratory data analysis. Martin has worked in industry and research in both, Germany and the United States. He received his Ph. D. in Statistics from the University of Augsburg in 1996. Martin is author of the data analysis software Mondrian and co-author of the book "Graphics of Large Datasets."

We are looking forward to interesting new discussions and contributions, welcome Martin!

At the same time that we welcome Martin, we say goodbye to Andreas. Andreas has faithfully

edited the newsletter for the last 3 years, and we wish him a heartfelt thanks for his service. Andreas has out lasted two consecutive editors from the computing section, sothat is one point for graphics. Thanks to hsi dilligent work transferring editorship has been an easy process. We wish him all the best in his current endeavors.

Getting into hot water over hot graphics

Antony Unwin
Institute of Mathematics
University of Augsburg, Germany
unwin@math.uni-augsburg.de

Background — what is hot?

When the editors asked me to write a short note on the hot topics in graphics research, I was very pleased. It's an excellent idea for someone to be asked to put forward their views contentiously and provocatively and then to encourage others to comment, enhance, and improve.

However, the scale of the task gradually dawned on me. Data graphics as whole is a hot topic because more and more researchers from very different areas publish graphics. This means there is no way that one person is able to keep abreast of all the developments that are going on. Any attempt at a summary is bound to leave out more than it includes. Of course, in that frame of mind no one would ever write anything, so I started to think about the term "hot topics" again.

What is a "hot topic" anyway? (Presumably it is cool to be working on a hot topic.) We might say it is a research topic that is currently getting a lot of attention and where quick, possibly easy, progress is expected. For instance, dynamic graphics was a hot topic for a while in the 1980s. More recently heatmaps have been a hot topic for visualising microarray data. Geographic mashups are a hot field in information visualization.

Where might "hot topics" in graphics be found? Journals tend to lag behind, especially given their long refereeing process, so conferences are probably the place to look, particularly as attractive and motivating presentations can work up interest and attention for a topic. However, if a topic is to stay "hot", there should be some good publications.

With that in mind I took a look at the last couple of years of the Journal of Computational and Graphical Statistics (JCGS). It will surprise no one reading this newsletter that JCGS has not published many papers on graphics. Between 10 and 15% of the papers (depending on how you define whether a paper was on graphics or not) concerned graphics. If there had not been two issues with subsections devoted to graphics this figure would have been much less. This is not the place to discuss why JCGS is actually JCgS, so let's move on.

What about other topics?

Actually, current "hot" topics are not so important, when there are still so many opportunities to make progress on fundamental research topics in graphics. Progress in these areas could make a real difference, they are not just temporarily fashionable. For me these topics are:

1) The need for a formal structure for data graphics. Bertin and Wilkinson are excellent books and they need to be built on, an essential requirement for a successful theory. Hadley Wickham's R package, ggplot2 (`http://had.co.nz/ggplot2/`), which is based on Lee Wilkinson's "Grammar of Graphics" is a very nice piece of work.

However, it is as yet only a part implementation, and, for example, mosaicplots and treemaps have still to be implemented in ggplot2. How this is achieved will help us understand the grammar better and should help us move on. Formalising interactive graphics is another part of this process and it would be a real step forward if someone could show how to do that.

2) Finding good ways for teaching the design and use of graphics. Many graphics texts give good guidance on how to draw particular graphics (and plenty of sensible guidance on why some published graphics are awful), but there is little advice on general strategies for choosing graphics and even less on how to decode and interpret graphics.

The statistical principle of thinking hard about

where data have come from and whether any general conclusions can be drawn from them often seems to be neglected when looking at graphics. This is a leftover from the days when graphics were regarded dismissively as descriptive statistics, not worth the attention of serious mathematicians. (Perhaps it is a good thing to keep serious mathematicians away from analysing data, but it's not a good thing when they have influence over what is taught to students.)

If we had a formal structure (see above), it might be easier to structure teaching. There is nevertheless plenty of scope for more innovation in teaching, especially given the ready availability of powerful hardware with fast and flexible display facilities.

3) Gaining a better understanding of how graphics are perceived, especially interactive and dynamic graphics. There has been some research in this field and there should be a lot more. The problem is tricky because there are so many different factors involved, though isn't that just the kind of problem where the statistical approach should help? Perhaps eyetracking research will make a difference, and good visualization of the data from eyetracking experiments could contribute a lot to making progress. Statisticians should get involved here and not leave it all to the computer scientists.

4) Improving the quality of presentation graphics in academic journals and the media. Despite the good work of Tufte, Cleveland and others, standards do not seem to improve. Perhaps this is just a basic law of nature: in any field the dross probably outnumbers the good and it certainly outnumbers the excellent. The poor quality of published presentation graphics could be due to users adopting the defaults that graphics packages offer them without thought, even if the package could be made to produce something much better.

I will refrain from mentioning Excel in this context and just mention that some of the defaults in R, in particular the open circles for points in scatterplots, seem to me, to put it mildly, suboptimal. The availability of online graphics toolkits may change that if their interfaces make users more aware of alternatives and make it easy for them to try them out. (And before someone points out how easy it is to change defaults and explore alternatives in many packages, R included, let me suggest that overriding defaults has to be *really* easy for users to consider it.)

5) Integrating graphics and modelling. Graphics generate ideas, which should be tested. Determining combinations of statistical models and graphics which complement each other effectively would be a very positive step in this direction. It always puzzles me that so little is done on this. There has always been some work on which graphics can be used to complement statistical analyses (and the newish SAS ODS graphics make a big point of this), but there is hardly anything done on which analyses should complement graphics. There seem to be big opportunities here to do interesting and valuable research. Scatterplots shed a great deal of light on the interpretation of correlation coefficients. We need a lot more than correlation coefficients to complement scatterplots.

All these topics are important, though not necessarily "hot" or new.

There are other topics I would like to include, even though they are not specific to graphics, not "hot", and possibly not even tepid. One is the issue of desktop organisation. Modern computing power enables us to produce many graphics and many analyses quickly. Unfortunately the software we have available does not manage and organise all these graphics and analyses well. We have surely all looked at desktops cluttered with windows wondering how we got into such a mess. A second issue is Graphical User Interfaces.

These are by no means as good or as supportive as they could be, though they seem to be making progress. The advent of smartphones and popular tablet computers could lead to productive developments here. They do seem to suggest a change in design philosophy, away from the developer-centered view ("You're lucky to have the software at all, just learn to use the interface I provide.") to a more user-centered view ("This interface may restrict your control, but it's intuitive to use and allows you to do everything that an average user needs.").

Looking for hot topics — getting warmer?

New developments can arise from applications (there has been a lot of work in the last few years on visualizing microarray and related data) or from new software. New visualisations of applications sometimes generate a lot of attention.

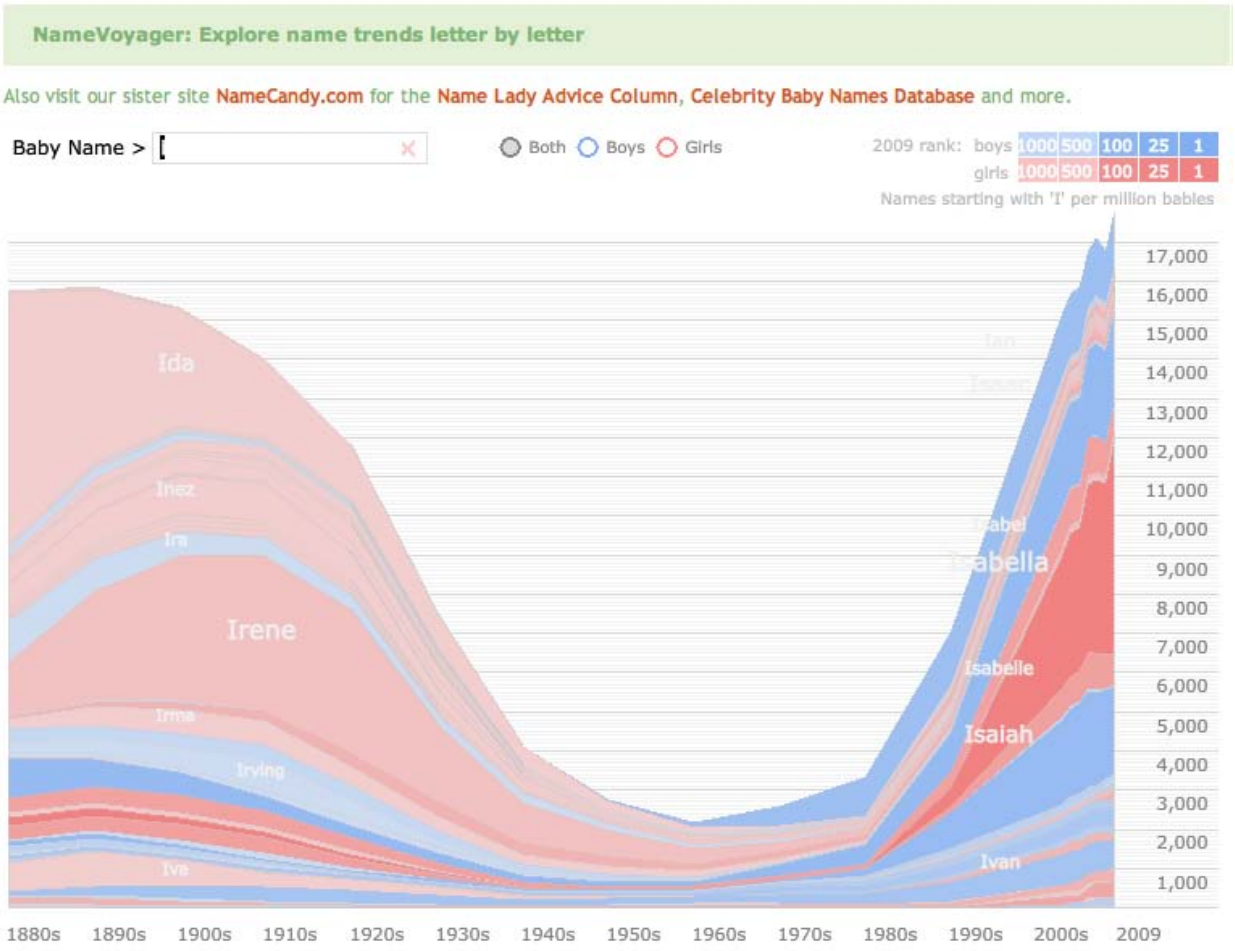


Figure 1: A screenshot from Name Voyager where names beginning with the letter I have been selected. Note the dip in popularity and the fact that the names after the dip are different from the names before the dip.

Martin Wattenberg's Name Voyager (`http://www.babynamewizard.com/voyager`) is an excellent example, see Figure 1.

The combination of a clever animation, an interesting database, and a smooth implementation works very well and shows that implementation is just as important as the idea itself. At the moment this is an isolated graphic type and it is not clear what more it might lead to. Additional applications are around and the idea is one to be watched. This is often the case with graphics. Someone finds an effective way of visualizing a particular dataset, sometimes with only limited association with previous work, and then developments lead to useful generalisations.

Generalisation is essential, if graphics are to be incorporated in a formal theory. The alternative view of graphics as just a collection of different plot types is quite prevalent. The group at Stanford, who are responsible for the research software Protovis, recently published an article called "A Tour through the Visualization Zoo" (`http://queue.acm.org/detail.cfm?id=1805128`), which presents visualisation in this way. That is an unsatisfactory approach (and is not improved by their including stem and leaf plots as one of their graphics).

However, their software looks interesting and new software, taking advantage of progress in computer power both in hardware and software, is definitely a hot topic. One visualization idea which has got a lot of attention is Martin Krzywin-

ski's Circos (mkweb.bcgsc.ca/circos/), using circular design to make multivariate displays. An example by a student of mine, Simon Bley, using some of the data from last year's JSM Data Challenge on flight delays (`http://stat-computing.org/dataexpo/2009/`) is shown in Figure 2. There are, as always, several people and groups working on new developments and it will be fascinating to see which ones are generally adopted.

Websites — web excitement is cold comfort or the promise of a warm future?

Two TED talks (Technology, Entertainment, Design: `http://www.ted.com/`) have drawn attention to graphics. Hans Roslin's Gapminder talk is splendid, using important data and a clever animated visualization (plus some sword swallowing) to highlight development trends across countries over time. The other talk was by Tim Berners-Lee on open data, i.e. making data available to the public. This is obviously a step that statisticians should welcome, since it encourages people to carry out statistical analyses.

That is the theory. In practice the analyses carried out are not necessarily statistical and the first example Berners-Lee shows of bicycle accidents would benefit considerably from some statistical input (cf. the second topic in my list above). The Guardian newspaper in England has set up a datablog to draw more attention to some datasets and encourages its readers to contribute graphical displays. A recent example for presenting the deficits of OECD countries as submitted by a reader is shown in Figure 3. Whether you like such graphics or not, more attention is being paid to graphics and more people are getting involved. This has to be a good thing and is "hot".

Given the many websites which now exist for presenting data, we might expect increasing numbers of innovative ideas and promising new approaches. That may well be true, except they are being swamped by extensive use of non-innovative ideas and distinctly unpromising old approaches. Finding good graphics in this situation is like trying to find good poetry by reading all the poems put up on the web. And good graphics, like good poetry, are partly a matter of taste.

The R Graph Gallery (`http://addictedtor.free.fr/graphiques/`), hosted in France, offers users the opportunity to rank the over 150 graphics submitted. (The fact that at the beginning of June 2010 the top graphic scored 80.27 out of 100 based on 86 votes while the bottom graphic scored 0.22 out of 100 based on 12082 votes should alert us to possible problems with the voting system.) Artists are doubtless sometimes distressed at the work produced by amateurs, clothes designers are probably not best pleased by how most people dress, why should statisticians expect standards in our field to be of a different caliber?

Strength in numbers

Turning away again from "hot" topics, let me suggest another cooler one, though one I think is cool in both senses of the word. Traditionally graphics were single graphics and some superb individual ones were drawn which combined huge amounts of information in one display (Minard's chart of Napoleon's invasion of Russia being the classic example).

Nowadays there is no reason to restrict ourselves to single graphics, we can draw sets of graphics, each conveying a different piece of information and contributing to the overall picture. There has been some discussion of this approach. Forrest Young introduced the more restricted idea of spreadplots. Eick and Carr wrote about perspectives in their JCGS article in 2000.

What should we recommend? Are a set of simple individual plots, possibly linked, better than one comprehensive image, or is a single well-designed image more effective through its intrinsic coherence? This question is closely linked to the issue of having a theory for graphics and it would be great to see some serious discussion. There are a lot of attractive possibilities to compare and contrast.

Cool judgement

As readers interested in graphics (and if you have got this far I presume you are interested), we all know the expression "A picture is worth a thousand words." There is a related quotation of Woodrow Wilson's: ÒOne cool judgment is worth a thousand hasty counsels. The thing to do is to supply light and not heat.Ó So following Wilson's advice

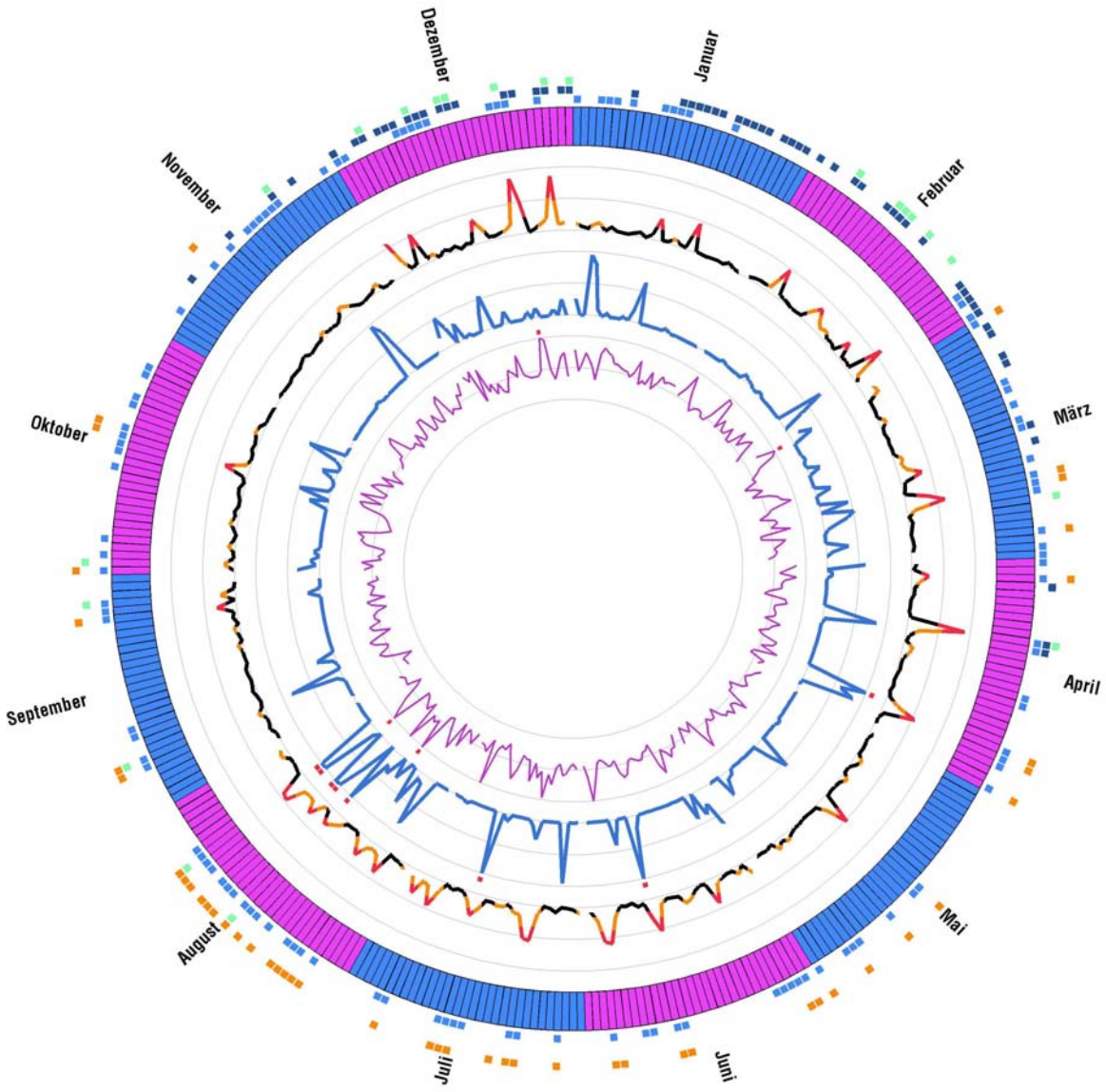


Figure 2: Weather influences on flight delays at Chicago airport in 2007.

I would recommend that we don't chase hot topics, we concentrate on thinking about developing fundamental theory and tools based on our own cool judgement. Graphical research has been in the shade for too long, we need to shed some light on it, but in a cool, considered way and not in a hot and hasty way.

Note: The graphics accompanying this article have not been explained in detail, as they are used only for illustration (as graphics so often are) and are not part of the main argument. The usual health warning applies: published graphics may imply incorrect conclusions which damage your reasoning. Make sure you understand what data have been used, how they were collected, and how they are represented, before drawing any substantive conclusions.

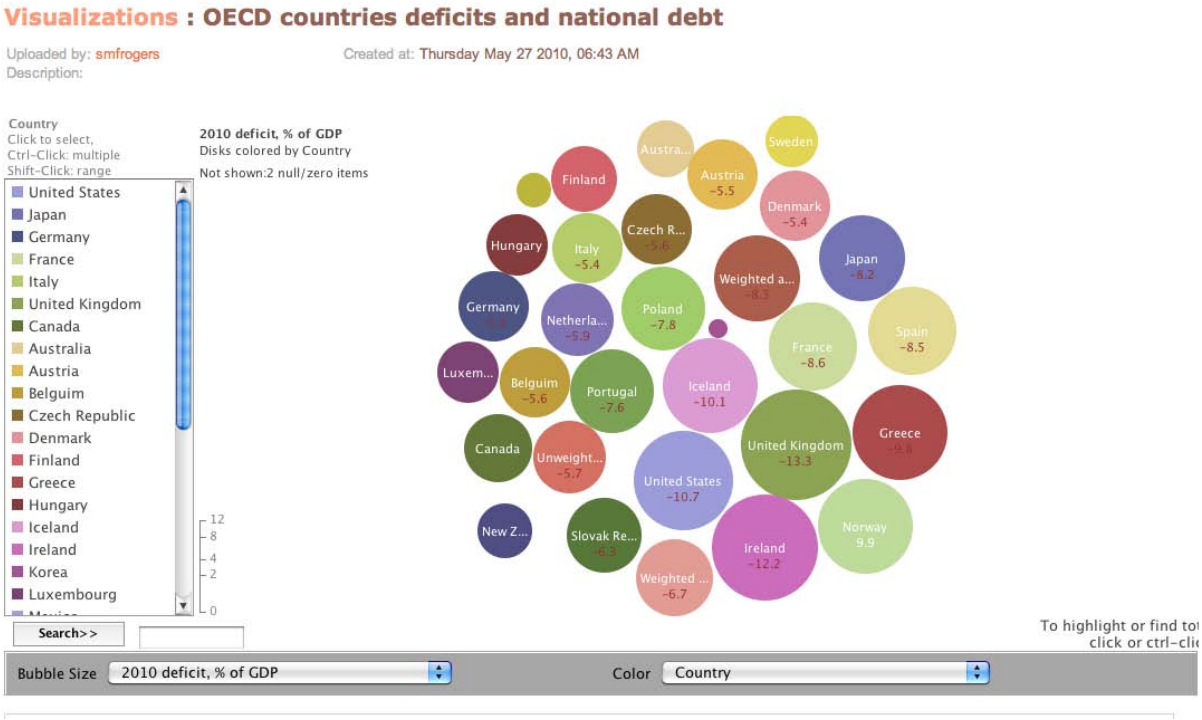


Figure 3: A suggestion for displaying the deficits of OECD countries submitted by a Guardian reader.

What's On(line) in Computing and Graphics?

Andreas Krause

The Elements of Statistical Learning

One of the best selling books ever in statistics was made available online.

"The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Hastie, Tibshirani, and Friedman (Springer-Verlag, New York) is now available for download in PDF format from `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

Data, data everywhere

The Economist, supposedly one of the most influential news sources, published a special report on managing information titled "Data, data everywhere" in its February 27, 2010, edition.

The leader, "The data deluge," introduces the topic of data generation and collection in vast quantities, and gives examples of successful analyses as well as advantages (detection of credit card fraud, fraudulent insurance claims, mobile phone use) as well as bad news (stolen data, lost data, identity theft).

The special report, "Data, data, everywhere" includes articles such as "Clicking for gold" and "Needle in a haystack," describing how internet companies profit from *data* on the web.

An interesting example is the principle of recursive learning from the data. The example given is spell-checking: Microsoft supposedly spent millions of dollars over 20 years to develop a robust spell-checker where Google got the data for free: its program is based on the misspellings that users type into a search window and then "correct" it by clicking on the right result. Translation and voice recognition are current topics of research based on a similar principle using statistical inference.

Available to subscribers only at `http://www.economist.com/node/15579717?story_id=15579717`

The entire special report can be bought as PDF for $4.95 at `http://www.economist.com/node/15560366`

R Bloggers

The web site "R bloggers" offers articles about R at `http://www.r-bloggers.com/`.

Some recent articles include "Getting started with Sweave," "Advanced graphics in R," "Visualization of regression coefficients," and

"ggplot2 version of figures in 'Lattice: Multivariate Data Visualization with R.'"

Statistics at Google

Google's business is built around data. (see also the Economist articles referenced in this section). Daryl Pregibon, a prominent statistician, describes life and work at Google in a featured article in the ASA Newsletter. The newsletter and the article (pp. 3-5) are available online at `http://content.yudu.com/A17l78/Yudumay09/resources/index.htm`

Google Research Publications

Google publishes much of its own research. The Google research home page, `http://research.google.com/`, and the publication page, `http://research.google.com/pubs/papers.html`, feature a variety of highly interesting articles, categorized into (among others) Algorithms and Theory, Artificial Intelligence and Data Mining, Information Retrieval, and Human-Computer Interaction and Visualization.

For Todays Graduate, Just One Word: Statistics

The New York Times published an article that paints a rosy picture for statisticians.

Hal Varian, chief economist at Google, predicts that the sexy job in the next ten years will be statistics.

The article is accessible online at http://www.nytimes.com/2009/08/06/technology/06stats.html

We Have Met the Enemy and He Is PowerPoint

The New York Times published an impressive example of how visualization techniques can be used to overdo it. A PowerPoint slide that was meant to illustrate a complex military situation was in fact overdone and that complex itself that a general remarked that "When we understand that slide, we'll have won the war."

http://www.nytimes.com/2010/04/27/world/27powerpoint.html

Section Video Library Online

Our sections have made available a great selection of videos. To quote from the web page, "The graphics video library is an archive of great technical and historic interest which captures much of the history of dynamic graphics for data analysis over the past 30 years. The collection is currently composed of 38 videos, listed below in chronological order."

The videos include J.B. Kruskal's 1962 (!) video on multidimensional scaling as well as the legendary PRIM-9 (Picturing, Rotation, Isolation, Masking) by J.W. Tukey, J.H. Friedman, and M.A. Fisherkeller from 1973.

http://stat-graphics.org/movies/

66th Deming Conference

The 66th Deming Conference will be held on December 6-10, 2010 in Atlantic City, NJ.

This conference has an applied biostatistics theme with an emphasis on biopharmaceutical applications. It is an unusual conference in that it focuses on half-day and two day tutorials only, given by authors of recently published books.

For the first three days, there will be two parallel half-day tutorial sessions for a total of 12 tutorial sessions (December 6, 7, 8).

Then the conference will continue with two 2-day short courses on December 9 and December 10.

Books used for the tutorial sessions and for the short courses will be sold at appreciable discounted prices from publishers. An exhibition of books on applied statistics will be held during the conference.

The conference home page is at http://www.demingconference.com/ and the program is already online.

A Tour through the Visualization Zoo

Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky have put together an interesting collection of visualization examples. Alongside, they introduce some of the principles of (good) visualization.

Their article, subtitled "A survey of powerful visualization techniques, from the obvious to the obscure," is a thought-provoking contribution to the discussion on visualization and graphics. Various examples provide inspiration about what is possible to creatively visualize data.

The article is online at http://queue.acm.org/detail.cfm?id=1805128.

Creative Techniques for Exploratory Data Graphics in PK/PD

Charles Roosen and Richard Pugh
Mango Solutions
Basel, Switzerland, and Chippenham, Wiltshire, UK
roosen@mango-solutions.com

Introduction

As timescales for analytical turnaround shorten, we are often required to make more decisions more quickly based on more information. One field where this certainly holds true is the study of pharmacokinetics and pharmacodynamics (or PK/PD) in the early phases of drug development.

Here, the primary focus is to explain the manner in which a compound enters and exits the blood stream using mathematical-statistical models. In PK/PD, the typical analysis involves modeling concentration of an administered drug over time. These models are a critical component in dose finding and safety assessments during the drug development process.

In PK/PD, the primary aims of the exploratory data analysis (EDA) process are to establish the quality of the data on which analyses are to be performed, and to identify structures within the data which could lead to more appropriate choices within the modeling process.

This paper presents some useful custom graphs applied to PK/PD data. The intent is not just to present specific types of plots, but also to demonstrate general principles of creativity for efficient exploratory data analysis. The techniques are of particular interest for data involving a series of measurements on individuals over time. All plots are created with the lattice package [1] in R [2].

Data Quality

Throughout this paper, we have used a popular simulated dataset distributed with the Xpose package [3] (available from `http://xpose.sourceforge.net/` or from the authors by request). This data has been selected due to a commonly occurring challenge with PK/PD data. It consists of measurements over time for numerous individuals (Subjects), which can make trends difficult to spot due to the large amount of information available on the same data range.

Many data quality issues are simple to spot, particularly incorrectly formatted values and data values that are significantly out of range. However, incorrect values that are still within the overall data range (such as transposed values) are more difficult to identify. In this case, we must rely on the structural information we anticipate within the data in order to spot outlying values.

If an incorrectly coded value is still within the anticipated data range, basic univariate analyses will not be able to identify the issue. In this case, some errors may be impossible to spot without quality control (QC) steps against the data source. We can, however, use basic graphical methods to spot observations that do not seem to follow similar trends. These observations may be erroneous, or may be highly informative.

In this section we will concentrate on 2 key variables: the observed concentration variable to be modeled and Time after Dose, taken as the primary independent variable.

A simple graph of observed concentration versus Time after Dose produces a graph in which the data groupings significantly overlap. Figure 1 displays all of the values in a single plot with grey lines connecting the observations for each Subject and a black line displaying a loess [4] smooth. In this plot it is hard to differentiate between Subjects or assess which Subjects have potentially aberrant measurements.

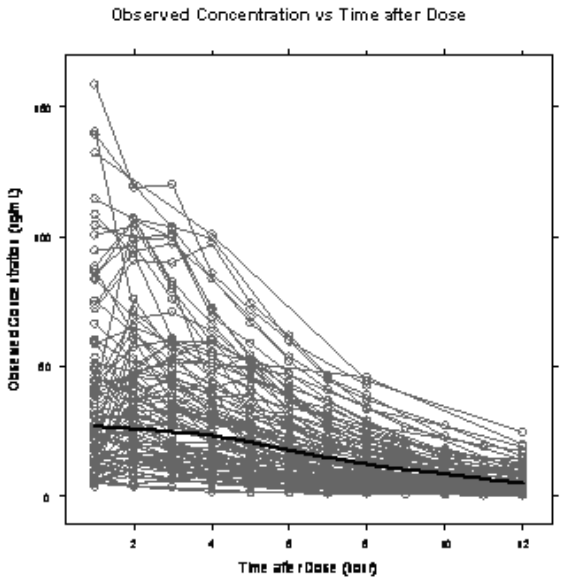


Figure 1: Observed Concentration versus Time after Dose

In the field of PK/PD, we typically understand the ways in which the direction of trend of the concentration curve should change, based on the dosing mechanism employed. Typically after the drug is administered the concentration will increase sharply to a maximum concentration, and then smoothly decrease as the drug is metabolized and eliminated.

As such, we can often spot erroneous data in a dependent versus independent relationship by simply analyzing the sign of the gradient at each step within a graph of observed concentration versus time after dose. The gradient is the slope of the line connecting two observations, so a switch in sign indicates a switch in the direction of trend.

Figure 2 graphs observed concentration versus time after dose, split by the number of changes in the sign of the gradient during the time period studied. In this plot, an upward facing green triangle is used for a value that is *greater* than the previous value and downward facing red triangle for a value that is *less* than the previous value. A change in the direction of trend occurs at locations where the plotting symbol changes.

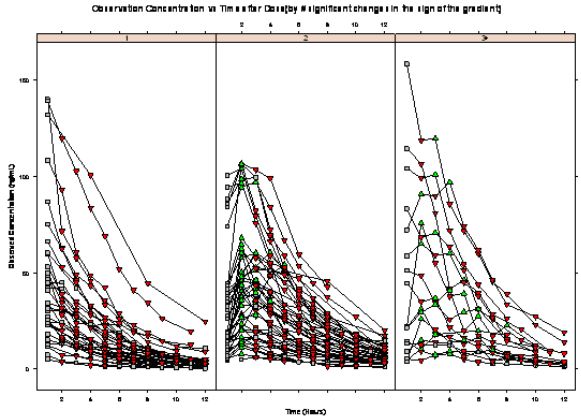


Figure 2: Observed Concentration versus Time after Dose, split by the number of times the sign of the gradient changes direction

As you can see here, the majority of the data has at most 1 change in the sign of the gradient. Selecting subjects where there are more shifts in gradient sign will often expose data groups which may require further analysis. Figure 3 reproduces the above graphic for selected subjects with a high number of gradient sign changes. We expect that these short-term increases are due to measurement or reporting inaccuracies.

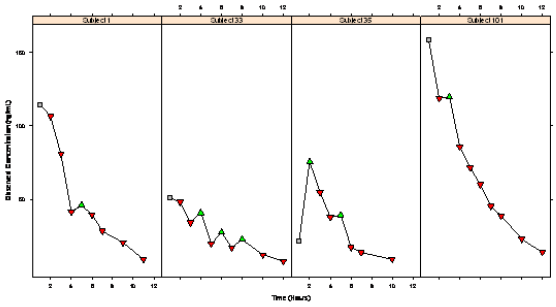


Figure 3: Observed Concentration versus Time after Dose, split by the number of times the sign of the gradients changes direction (including only selected subjects)

Another way to look for potentially erroneous data that could be contained within the acceptable range of the data is to scale the data against a model-agnostic smoother (such as a loess smoother). In order to perform this scaling, the fol-

lowing steps are used:

- Log the concentration data
- Fit a smooth line to the logged data
- Calculate differences from this smooth line
- Calculate the mean difference from smooth line for each subject, and subtract this from the differences (effectively centering each subjects data)
- For each time point, subtract the mean of the data and divide by the standard deviation
- Plot the calculated values vs time and look for outlying values

Figure 4 presents the scaled data generated versus Time after Dose. In this graphic, the subject identifiers are shown in dark blue, with subject groups linked by light grey lines. The change of colour here focuses the reading on the subject values themselves (darker colour) as opposed to the trend of each subject (lighter colour). Horizontal grid lines have been added to allow ease of reading.

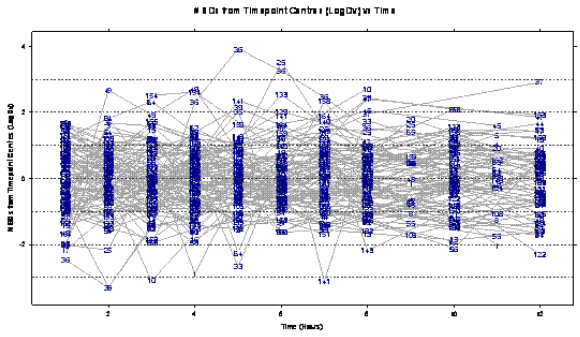


Figure 4: Scaled Standard Deviations from Timepoint Centres versus Time after Dose, with Subject identified

From this analysis, we can identify subjects that may warrant further analysis based on large absolute differences from 0. For example, the following subjects have at least 1 value that is more than 3 standard deviations from the mean timepoint centre in the above plot: 10, 25, 36, 39, and 141. Creating a plot of observed concentration versus time after dose for these subjects produces the graph shown in Figure 5.

In this instance, we want to illustrate how these subjects compare to the general trend of the data. If we were to include all subjects in this graph, the visualization would look too busy, and the focal message could be confused.

As such, in order to represent the general trend of the data, we have included polygons of colour representing the 0, 25, 75 and 100 percentiles of the total data set on each plot; these are represented as light and dark blue polygons (0 and 100 percentiles drawn in light colour, overlapped with dark polygons representing the 25th to 75th percentiles). In addition, the centre of the data is included as a (black) median line. To contrast against this background, the individual subject data has been graphed as bright red points linked by a line.

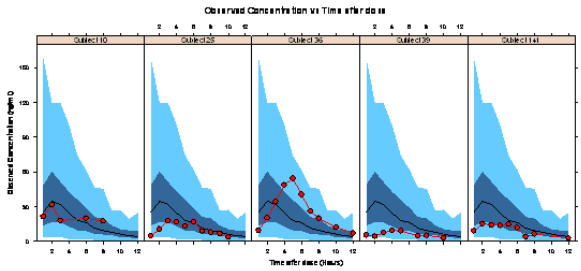


Figure 5: Observed Concentration vs Time after Dose including only those subjects flagged in the above analysis

From this analysis, subjects 10 and 36 certainly warrant further investigation. In particular, we should look at the observation at time point 3 for subject 10 (which appears to be comparatively low), and the reasons that the concentrations peak comparatively late for subject 36 when compared with the general trend of the data (since most concentrations peak at 2 hours).

Exposing Potential Model Structures

When investigating the use of covariates within our model steps, we can partition the data viewed by levels of candidate covariates. When performing this step, we should ensure the plots of each subcategory are graphed on the same axis scale. Also, when the dependant variable is known, each plot

should be aligned with the Y axis on the same level so comparisons can be readily made.

When there are a large number of subjects to be viewed, it often makes sense to graph select percentiles of the data. This makes it easier to spot overall trends. In Figure 6, the 0, 25, 50, 75, and 100 percentiles are displayed for each covariate group. As with the last graphic, the 0 to 100 range is in light blue, 25 to 75 in darker blue, and a black line shows the median.

To enhance the comparative usage of the graphics, the range of the overall data can also be plotted as a shadow (shown in grey on the following plots).

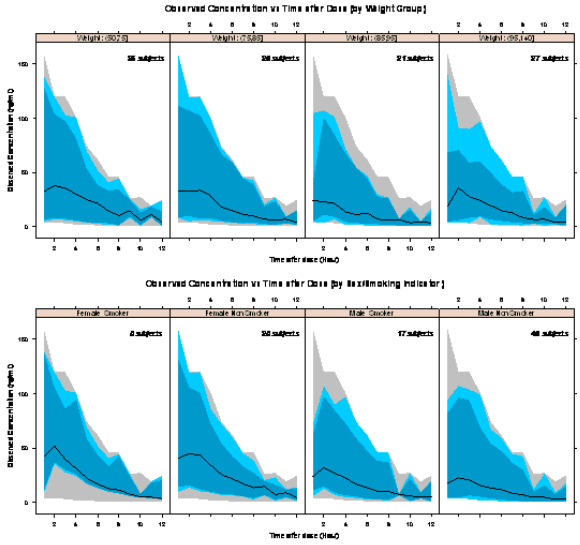


Figure 6: Observed Concentration vs Time after Dose plots split by a variety of covariate groupings, with the overall data range included as a grey shadow on the plot

The first graphic shown illustrates little or no change in the overall range of concentrations for subjects in the first two weight groups (50 to 85), with a marked decrease in concentrations beyond that point. The second graphic shows higher concentrations for female subjects, and a later peak in concentrations for male subjects. The graph also suggests a slight increase in concentration values for smokers.

Summary

We began with a straightforward plot of concentration over time by individual (Figure 1). This essentially gave us a tangle of overlaid points and lines where it was hard to distinguish the trend for individual Subjects.

To identify series with potentially aberrant values, we then created separate plots for series with 1, 2, and 3 or more changes in direction of trend (Figure 2). We also examined the series for specific individuals with a large number of changes (Figure 2). This provided a way to examine within series inconsistencies which likely reflect measurement or recording errors.

Taking a different approach, we next used a series of steps to normalize the values within each time point (Figure 4). This provided a way to examine between series inconsistencies in which a series has values that are extreme compared to the other series.

To examine how the series with extreme values compare with the other series in general, we created separate plots for individual series overlaid over shaded regions showing the quartiles of the data as a whole (Figure 5).

Finally, we created separate plots of quartiles for different covariate levels to assess whether the concentration varies based on covariates (Figure 6).

By combining points, lines, text, and shaded regions in creative ways, a wide variety of exploratory graphics can be produced to help gain insight into data.

Student Paper Competition Winners

Fei Chen, Avaya Labs
Awards Chair, Statistical Computing and Statistical Graphics Sections

As in previous years, the Statistical Computing and Graphics Sections sponsored their annual student paper competition. The requirements 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 winners are invited to present their papers at a special contributed session at the Joint Statistical Meetings, and the Sections pay their expenses to attend.

This year a large number of excellent entries were received, from which the selection committee has chosen four winners (in alphabetical order):

- **H**an Liu
  "Multivariate Dyadic Regression Trees for Sparse Learning Problems"
  Advisors John Lafferty and Larry Wasserman, Statistics and Machine Learning Program, Carnegie Mellon University

- **S**eo Yo Park
  "Multicategory Composite Least Squares Classifiers"
  Advisor Yufeng Liu, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill

- **A**li Shojaie
  "Discovering Graphical Granger Causality Using the Truncating Lasso Penalty"
  Advisor George Michailidis, Department of Statistics, University of Michigan

- **Y**ing Sun
  "Functional Boxplots for Complex Data Visualization"
  Advisors Jeff Hart and Marc G. Genton, Department of Statistics, Texas A&M University

The students will be recognized at the Statistical Computing/ Statistical Graphics business meeting at JSM 2010. Congratulations to the winners and many thanks to the judges for their hard work in making this year's competition a success!

UseR! 2010

Nicholas Lewin-Koh, Genentech

Starting July 20, useR! 2010 (`http://user2010.org/`) will take place in Gaithersburg, Maryland on the campus of the National Institute of Technology and Standards (NIST). UseR! represents a relatively new player on the block for the statistical community. While conferences centered around programming languages are not new, for instance PyCon `http://www.pycon.org/`, or the International Lisp Conference. UseR! is focused on innovations around the R statistical programming language, `www.R-project.org`. Attendance at UseR! has grown over the last few years, what is exciting is the cross section of people attending. The attendees are a cross section of industry, academia and government, representing fields from finance to medicine. This year's conference boasts invited lectures from Richard Stallman of the Free Software Foundation as well as our current chair of the computing section, Luke Tierney. Also featured at the meeting will be a panel discussion on "The Challenges of Bringing R into Commercial Environments", this is a conundrum many of us working in regulated industries have dealt with, where the release cycle of open source software can run counter to the conservative needs of businesses who must practice version control and deal with validation of computer systems. Besides these featured events there will be many exciting talks and posters. Once again this should prove a meeting worth attending.

Section Officers

Statistical Computing Section Officers 2010

Luke Tierney, Chair
luke@stat.iowa.edu
(319) 335-3386

Jose C. Pinheiro, Past Chair
jose.pinheiro@novartis.com
(862) 778-8879

Richard M. Heiberger, Chair-Elect
rmh@temple.edu
(215) 808-1808

Usha S. Govindarajulu, Secretary/Treasurer
usha@alum.bu.edu
(617) 525-1237

Montserrat Fuentes, COMP/COS Representative
fuentes@stat.ncsu.edu
(919) 515-1921

Thomas Lumley, Program Chair
tlumley@u.washington.edu
(206) 543-1044

David J. Poole, Program Chair-Elect
poole@research.att.com
(973) 360-7337

Barbara A Bailey, Publications Officer
babailey@sciences.sdsu.edu
(619) 594-4170

Jane Lea Harvill, Computing Section Representative
Jane_Harvill@baylor.edu
(254) 710-1517

Donna F. Stroup (see right)
Monica D. Clark (see right)

Nicholas Lewin-Koh, Newsletter Editor
lewin-koh.nicholas@gene.com

Statistical Graphics Section Officers 2009

Simon Urbanek, Chair-Elect
urbanek@research.att.com
(973) 360-7056

Antony Unwin, Past-Chair
unwin@math.uni-augsburg.de
+49-821-598-2218

Juergen Symanzik, Chair-Elect
symanzik@math.usu.edu
(435) 797-0696

Rick Wicklin, Secretary/ Treasurer
Rick.Wicklin@sas.com
(919) 531-6629

Peter Craigmile, GRPH COS Rep 08-10
pfc@stat.osu.edu|
(614) 688-3634

Mark Greenwood, GRPH COS Rep 10-12
greenwood@math.montana.edu
(406) 994-1962

Heike Hofmann, Program Chair
hofmann@iastate.edu
(515) 294-8948

Webster West, Program Chair-Elect
websterwest@yahoo.com
(803) 351-5087

Donna F. Stroup, Council of Sections
donnafstroup@dataforsolutions.com
(404) 218-0841

Monica D. Clark, ASA Staff Liaison
monica@amstat.org
(703) 684-1221

Andreas Krause, Newsletter Editor
andreas.krause@actelion.com

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 the publication should be addressed to:

Nicholas Lewin-Koh
Editor Statistical Computing Section
lewin-koh.nicholas@gene.com

Andreas Krause
Editor Statistical Graphics Section
andreas.krause@actelion.com

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
TEL (703) 684-1221
FAX (703) 684-2036
asainfo@amstat.org