A Word from our 2008 Section Chairs

DAN ROPE
GRAPHICS

Have you heard? Did you see? Racing across the ‘Net and gaining attention like a wet snowball bounding down a Colorado mountain gains momentum; it is the latest and greatest technological advance—and data visualization potentially makes it simpler, more powerful and downright attractive. Today’s regular stream of technology breakthroughs certainly keep things exciting for those that live to innovate. Many such advances are out there now and many more are coming...

DEBORAH NOLAN
COMPUTING

The buzz this summer is about the future of statistical computing. Alan Karr has organized a session on “Analytical Computing Platforms for the Future” at the International Symposium on Business and Industrial Statistics in Prague; Luke Tierney has organized the session “Advances in Statistical Computing and Graphics” at the World Congress in Probability and Statistics in Singapore;

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

Editorial Note 3
Feature Article - Fisher 4
Scientific Article - Zhu 7
Technology, Commerce 9
JSM Program 11
News & Conferences 14
Puzzles 16

PAGE 1
Ross Ihaka and Duncan Temple Lang will speak on “Back to the Future: Lisp as a Base for a Statistical Computing System” at the International Conference on Computational Statistics in Portugal; and “The Future of Statistical Computing” by Lee Wilkinson is forthcoming in Technometrics. It is not just hype; researchers in statistical computing are working to develop new systems that will lead statistics into modern computational realms, such as those brought by parallel computing, massive data sets, streaming data, and emerging technologies.

The program Wolfgang Jank has put together for the Section on Computational Statistics at the upcoming JSM in Denver highlights some of these computational challenges. For example, three sessions address problems from large-scale, complex applications in online advertising, social networks, and text mining. Another challenge, one that matters greatly to me, relates to the session, “Designing Courses on Statistical Computing”, organized by David Banks and sponsored by our section. This session grew out of an NSF-funded workshop on “Integrating Computing into the Statistics Curricula”, held at Berkeley last year and organized by Mark Hansen, Duncan Temple Lang, and myself. Although many agree that there should be more computing in the statistics curriculum and that statistics students need to be more computationally capable and literate, it can be difficult to determine how the curriculum should change because computing has many dimensions. Last summer’s workshop brought together faculty who have been addressing these questions. This summer and next, two follow-up workshops will be held to bring together faculty who are interested in changing their undergraduate statistics programs in innovative ways to incorporate more computing into existing courses and to add new courses with a focus on computing to the curriculum. Computing is an increasingly important element of statistical research and practice. It is an essential tool in a statistician’s daily work, it shapes the way we think about statistics, and it broadens our concept of statistical science. It’s high time we think seriously about modernizing the statistics curriculum to keep pace with these changes. If you want to chat about this topic, join me at the round-table luncheon, “Broadening the Undergraduate Statistics Curriculum”, on Wednesday at the JSM. Also coming up in Denver, is the continuing education course sponsored by our section, “Computational Statistics: Methods for Monte Carlo Integration and Optimization”, given by Jennifer A. Hoeting and Geof H. Givens. And of course, there is our traditional mixer with the Statistical Graphics Section on Monday evening, which promises to be a fun event with some terrific door prizes. Visit the online program to see all that we have in store http://www.amstat.org/meetings/jsm/2008/.

On a final note, our Program Chair-Elect, Robert McCulloch, will be putting together the program for 2009 soon. If you have ideas for invited sessions, send them along to Rob.

- Deborah Nolan

But wait, there's more happening on the “new” Web. You see, a funny thing happens when you put a couple hundred million people together on an interactive network. They socialize. They interact. They produce data—a whole lot of data! What techniques can we use to analyze data of this scale—or, even, of this nature? The excitement increases when you include a geo-spatial aspect. This is certainly an interesting area of study both for statistical computing as well as graphics and there are a few sessions this year covering these topics.

But, so far we are talking only about the here and now. What is coming next?

On the horizon, clouds are gathering. But, these are not rain clouds. Rather, these are computing clouds loaded with enough parallel processing power to make even Star Trek characters envious—and best of all, you will pay for it the same way that you pay for your telephone bill. What will the next generation of applications in statistical computing and graphics look like using processing power of this magnitude?

Moving from the obscenely large to the elegantly small, we are witnessing tremendous advances in mo-

 PAGE 2
bile devices—now fully programmable by anyone. iPhone software development is open to eager thinkers at-large and Google's Android operating system for mobile devices is on the way. Here, standardization forms the basis for other disciplines, such as ours, to innovate. It is guaranteed that some of these applications will display data—even large amounts of data, graphically. How can such a small display area remain effective? What sorts of interactive graphics features will arise from new user interface metaphors such as multi-touch?

Indeed, these are exciting times and I truly believe statistical graphics has an important role to play. As usual, this is a good time of year to begin brainstorming session ideas for JSM 2009. Please be sure to pass along your thoughts to our program-chair elect, Steven MacEachern.

No pressure or anything.

- Dan Rope

Monday, August 4th, 7:30 PM
Statistical Computing and Graphics Business Meeting and Mixer
HY Centennial Ballroom C
We expect to see you there.

EDITORIAL NOTE
What's Inside this Issue

Michael O’Connell and Andreas Krause
moconnell@insightful.com
andreas.krause@actelion.com

In this issue of the newsletter, we have two wonderful statistical graphics articles. Mark Fisher describes statistical graphics in the wind industry. This is a rapidly growing industry - there is a national goal to produce 20% of US energy from wind by 2030. Mu Zhu shows us how to do a trilinear plot. Both articles feature excellent statistical graphics.

We need more articles for the newsletter. Please contact us if you have any short articles, excerpts, software or graphs. This is a good forum for short articles on statistical computing or graphics. You can reference your newsletter article and still publish your work elsewhere.

David Hunter and Wolfgang Jank give us a guide to JSM 2008. We have a great program this year including presentations from many hot new technologies and companies such as Swivel, ManyEyes, Facebook and Google. The program chairs have done a great job.

We introduce a new section - Technology and Commerce Corner - including an interview with the head of the New York Times graphics department, Steve Duennes, and some contributions from young entrepreneurs in the field. The Steve Duennes article is a must read.

The awards for this year are announced by JR Lockwood, including the Chambers Award.

In the conference roundup we have a review of the Graybill VII conference and the International Chinese Statistical Association (ICSA) conference.

Finally, we have the solution to the puzzle from the last issue - this was a popular puzzle with many entries that were close to the solution. And we have a new puzzle! So please get your entries in on the new puzzle before our next issue.

See you in Denver!!

Michael and Andreas
Featured Article

STATISTICAL GRAPHICS IN THE WIND INDUSTRY

Mark A. Fisher
Airstreams
www.air-streams.com
markfisher@air-streams.com

1. Introduction

Wind energy is a rapidly growing form of renewable energy. There are currently only a few thousand wind energy jobs in the United States producing about 1% of the nation's electrical power. And there is a national goal of reaching 20% of our power generated by 2030. This is expected to dramatically increase the number of jobs in the wind industry [1].

Utility sized wind farms will have between 60 and 120 wind turbines each rated at 1 to 3 MW of power. Each of these wind turbines is controlled by a SCADA (Supervisory Control And Data Acquisition) system. Note that part of the acronym is Data Acquisition. The data is typically being recorded as averages every ten minutes (however, with the advent of cheap data storage, higher rates are sometimes used). Thus there are 144 records for each turbine each day, and each record will have 100 or more fields being measured. These fields measure values such as electrical power, wind speed, temperatures of components, and other operational parameters. Each record will identify the wind turbine and the timestamp. In addition, there are other event based forms of data which are recorded when the turbine changes from one state to another. While these are typically more rare than the regular ten-minute data, this can also generate a large amount of data.

The role of the statistician or data analyst in this environment is to create summaries of this data to help the technicians optimize the performance of the site. Turbines with big problems have already shown up in the most basic measures used, however subtle issues can be lost without a constant examination of the data. For example, power can be lost if the wind turbine is not pointed directly into the wind. The site technicians need to have a quick way of determining which turbines need additional attention. Over the past few years several graphical techniques have been developed using R and S-Plus to help find these turbines.

2. Event Based Data

When a wind turbine changes state a record is created indicating what the new state of the turbine is. One of the states given the most attention is “faulted.” The turbine is designed to shutdown when certain problematic conditions occur. For example, there might be a high wind fault where the turbine shuts down to protect itself from wind speeds that are too high. Other faults types will cover situations where the electrical grid is problematic, or the generator is overheating, or some mechanical problem is detected. (An internet search can find videos of what can happen when these safety systems fail. The results can be dramatic.) However, the desire is to have the turbine produce as much power as possible, so constant investigation of turbine faults is important.

Figure 1 shows one graphical display used to look for patterns in faulting. The vertical lines represent certain external events, in this case grid failures. The turbine events have faults coded in black and they consist of a point marker and a segment showing the length of the whole event. The yellow and red ones represent repairs and maintenance events. The right hand scale provides information of the number of faults and the total duration of the faults. Wind information is added since storms may cause multiple turbines to fault simultaneously. All of this except the wind direction information is just using basic graphing features of R. The wind direction is calculated using the circular package [2], then plotted using arrows.

The circular package is also used to look at time of day patterns for faulting. The plot in Figure 2 can illustrate whether there are any diurnal patterns to fault occurrence [3]. There are at least a couple factors that can cause diurnal patterns in faults. First most site personnel are around during the day, so this could lead to faster resets, providing more opportunity for faults. Second there is typically a diurnal pattern to wind speeds (and sometimes directions). There are also diurnal patterns to electricity usage, which might have
an impact on the grid behavior, and hence faulting. In Figure 2 the dots represent fault events, the green ones are during normal work hours, while the blue ones are outside normal work hours. The density provides a way for the reader to quickly see which time periods seem to have a greater concentration of faults. Finally, the red rose diagram inside the plot provides estimates of the energy lost for each time period.

3. Sensor Data

The event data can highlight some problems with the turbine however, problems with turbine performance and some components may show up in the 10-minute data recorded by the SCADA system. For large components, such as the gearbox or generator, replacement will require a crane, making these components some of the most expensive to replace. Trying to determine which turbines might have problems developing with these components is one way to improve site performance.

By making appropriate summaries (averaging over time or for a given power level) values for all the turbines at a site can be created, and outliers can be spotted. For this the ellipse [4] package is helpful. Looking at two variables, such as generator and gearbox temperatures, simultaneously, turbines where overheating might be occurring may pop out as outliers.

The colors in Figure 3 represent turbines with generators produced by different manufacturers. At this particular site, the wind turbines had two different kinds of generators installed. So separate ellipses are needed since different models of generator may have different normal operating temperatures. The outer curve represents a 95% confidence ellipse, while the inner curve represents a 90% confidence ellipse. Several of these turbines (W37, W38, W47, W48) are crossing the boundary for the 95% confidence ellipse. These then would receive additional attention and be tracked for the next report.

Distributions for each variable are plotted on the top and right sides. The bimodal Generator Temperature corresponds to the known issue of having generators manufactured by two different companies. Noting that the Gearbox Temperature appears to have a trimodal (at least starting to be trimodal) and there is only one type of gearbox at the site, the turbines running hotter gearbox temperatures should receive additional investigation. High gearbox temperatures sometimes is
associated with excessive wear and could be an indication of a problem developing.

While techniques like the confidence ellipse can help find turbines that are behaving differently than normal, this requires sites with enough wind turbines to make such statistical techniques meaningful. For sites where there are a small number of turbines other techniques which compare a turbine to itself can be used.

Each turbine has a specific power curve that it has been designed to meet. This power curve is basically a function taking wind speed and gives a power the turbine is supposed to produce at that wind speed. In the reality of ten minute averages for both wind speed and power, there is a certain scatter around the function (see Figure 4). The power curve is a complex S-shaped curve and it can be difficult to quantify deviations from the curve.

Therefore how well the turbine is meeting the power curve is now quantified by two numbers. First, the correlation, which will indicate a deviation from linear. This can mean the turbine is having problems with pitch, or perhaps an anemometer issue, which affects how the turbine responds to different wind speeds. Second, the slope itself, which gives whether the turbine is performing above or below the power curve. Typically if it appears to be performing above the power curve, this indicates that there is a problem with the anemometer. If it is performing below the power curve, there are several possible problems, one of which is that the turbine isn't pointed directly into the wind, which causes a power loss.

4. Conclusions

Wind energy is a rapidly growing part of the energy sector. Until recently the major concerns were whether the turbine was running or not. As larger companies have become involved, better measurement and monitoring of turbine performance has been getting more attention. While analysts remain key in understanding the complex data needed for even the most basic reports, more attention has begun to be
given to how to improve the performance of the wind farms.

It has been only in the last few years that statistical techniques have begun to be applied in the wind industry. So there are plenty of opportunities to try analyses that no one has ever tried before. Then the results of these analyses make a difference.

5. References


The tri-linear plot is an effective way to display trinomial data. Allen (2002) provides a number of very interesting examples. Table 1 contains a subset of the data from Allen (2002; Table 1), which summarizes the percentage of respondents to a 1996 general social survey who thought the U.S. government should spend more or less with regard to five specific issues. Figure 1 shows the corresponding tri-linear plot, which makes it visually obvious that the majority of the respondents want the government to increase spending on education and decrease spending on foreign aid and welfare.

<table>
<thead>
<tr>
<th>Issue</th>
<th>More</th>
<th>Same</th>
<th>Less</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Security</td>
<td>52</td>
<td>40</td>
<td>8</td>
</tr>
<tr>
<td>Education</td>
<td>76</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td>Foreign Aid</td>
<td>4</td>
<td>19</td>
<td>77</td>
</tr>
<tr>
<td>Welfare</td>
<td>16</td>
<td>27</td>
<td>58</td>
</tr>
<tr>
<td>Mass Transportation</td>
<td>36</td>
<td>53</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 1. Percent of respondent opinions regarding U.S. government spending. Source: Allen (2002; Table 1).
2. Constructing the tri-linear plot

What is the best way to draw such a tri-linear plot? A common answer is to compute the coordinates for the tri-linear plot directly using elementary geometry. For example, suppose $x_1$, $x_2$, and $x_3$ are counts of how many people in a given state voted for George Bush, Bill Clinton and Ross Perot in the 1992 election. To represent this state by a point in $P$ inside an equilateral triangle $\Delta ABC$ (Figure 2), simply let $a$ be the distance from $P$ to the side of the triangle facing the vertex $A$, define $b$ and $c$ likewise, and try to find the point $P$ that satisfies the relationship

$$\frac{a}{x_1} = \frac{b}{x_2} = \frac{c}{x_3}.$$ 

Though elementary in principle, such a calculation can be algebraically tedious for most people.

I describe an alternative method that is much more appealing conceptually. Although the tri-linear plot itself is a 2-dimensional plot, it is actually much simpler to think in terms of 3-dimensional geometry. Let $x_i = (p_{i1}, p_{i2}, p_{i3})^T$ be a trinomial observation, where $p_{ik} \geq 0$ is the probability (or proportion) of observation $i$ “belonging to” category $k$. The constraint

$$\sum_{k=1}^{3} p_{ik} = 1 \quad (i)$$

can be easily enforced by normalization. The value $p_{ik}$ can also be thought of as the “affinity score” between observation $i$ and category $k$.

The three vertices of the tri-linear plot correspond to the three pure categories; they are the points $(1,0,0)$, $(0,1,0)$ and $(0,0,1)$. The constraint (i) implies all observations $i = 1, 2, ..., n$ lie on the unique plane determined by the three vertices. Moreover, because of the implicit constraint $p_{ik} \geq 0$, they all lie inside the triangle enclosed by the vertices; see Figure 3.
This means that the coordinates of the data themselves form a tri-linear plot automatically, except that such a plot resides in 3-dimensional space. Therefore, to produce a tri-linear plot, the only calculation needed is a simple projection of everything into a 2-dimensional subspace, which automatically yields the proper coordinates.

In particular, it is easy to see that a convenient choice of the projection directions are unit vectors $a_1$ and $a_2$ in $\mathbb{R}^3$, where

$$a_1 \propto (1, -1, 0)^T \quad \text{and} \quad a_2 \propto \left(\frac{-1}{2}, \frac{-1}{2}, 1\right)^T.$$

Pretty cool, isn't it?

3. References


4. S-PLUS/R Code

```r
trilinplot<-function(data, box = F, abbrv.name=F,
vertex=dimnames(data)[[2]],
item=dimnames(data)[[1]], ...) {

# Produces trilinear plot.
# copyright 2002 by
# Mu Zhu
# University of Waterloo
# Waterloo, Ontario, N2L 3G1
# Canada
#
# references:
# Allen, T. (2002), "Using and

# Interpreting the Trilinear Plot",
# ---------------------------------------------

if (is.null(vertex)) vertex<-c("A", "B", "C")
if (is.null(item)) item<-1:nrow(data)
if (abbrv.name) item <- abbreviate(item)
# offset to make text position appear
# prettier
eta <- 0.08

# vertex position
anchor <- diag(3)
# projection directions onto 2-d
alpha1 <- c(1, -1, 0)/sqrt(2)
alpha2 <- c(-0.5, -0.5, 1)/sqrt(1.5)
# plot the triangle
vert.x <- anchor %*% alpha1
vert.y <- anchor %*% alpha2
plot(vert.x, vert.y, type="p", xlab="", ylab="",
 xlim = range(vert.x)+2*c(-eta, eta),
 ylim = range(vert.y)+2*c(-eta, eta),
 axes=F, ...)
if (box) box()
polygon(vert.x, vert.y)
offset.x <- c(eta, -eta, 0)
offset.y <- c(-eta,-eta,eta)
text(vert.x+offset.x, vert.y+offset.y,
 as.character(vertex))

# make sure each observation contains
# proper percentage
rowsum <- as.vector(as.matrix(data) %*%
 rep(1,3))
data <- data/rowsum

# item position
pts.x <- as.matrix(data) %*% alpha1
pts.y <- as.matrix(data) %*% alpha2
# plot items onto triangle
n <- nrow(data)
points(pts.x, pts.y)
offset.x <- eta/2*sample(c(-1, 1), size=n,
 replace=T)
offset.y <- eta/2*sample(c(-1, 1), size=n,
 replace=T)
text(pts.x+offset.x, pts.y+offset.y,
 as.character(item), cex=0.62)
}

Code is also available on our website:
http://stat-computing.org
Technology and Commerce Corner

This issue introduces a new section on Technology and Commerce. We hope to encourage all kinds of folks, especially entrepreneurs, who are working with current computational and graphical technologies, to tell us about their work.

This issue, we are excited to include an interview with the New York Times graphics editor, Steve Duennes on our website.
http://blog.stat-computing.org/2008/05/talk-to-newsroom-interview-with-steve.html

Steve provides us all kinds of fascinating angles on graphics. The article includes the best examples of New York Times print and online graphics from the 30-person graphics team themselves. Steve talks about influences of folks like Ed Tufte, John Grimwade, Nigel Holmes, Ben Shneiderman, Martin Wattenberg and artists like Sol LeWitt, Saul Steinberg and Mark Lombardi. There is a link to the fabulous old short film, Powers of Ten, by Charles and Ray Eames. Some of the graphs are truly fabulous, like Barry Bonds’ home run record.

In a similar vein, the new book, Design and Science: The Life and Work of Will Burtin (Lund Humphries), seems like a good read from the perspective of historical scientific graphics. Burtin was art director for Fortune magazine in the 40’s and design consultant for Kodak, IBM, Union Carbide and Upjohn in the 50’s. He did much great work at Upjohn including graphics quantifying effects of antibiotics on bacteria and walk through installations of cells and the brain.

We are also excited to get a couple of contributions from young bloggers and website visualization entrepreneurs.

Please pass the word that we are encouraging entrepreneurs to tell us about their work on computing and graphics. We hope that this section can grow!

Enjoy the interview with Steve Duennes from the New York Times, the article below from Nathan Yau on visualization and blogging, and the note and graph below from Vincent Granville. More entrepreneurs and commercial folks please!!

- Michael O’Connell

LEARNING ABOUT DATA VISUALIZATION THROUGH BLOGGING

Nathan Yau
FlowingData

I am a proud statistics graduate student from the University of California at Los Angeles, but I live in Buffalo, New York. My focus is in data visualization. About a year ago, my wife finished medical school and matched to a three-year emergency medicine program in Buffalo while I was only two years into my doctorate. Luckily, I was able to arrange things so that I could complete my degree from the other side of the country. However the question still remained - "How am I going to learn enough about data visualization to earn a PhD far from my adviser, research group, and dissertation committee?" The answer has been a lot of conference calls and emails and quarterly visits back to UCLA, but something unexpected has helped me progress - blogging.

I started my statistics and data visualization blog, FlowingData, when I moved. I originally intended it to be a place to document literature, and more than anything I wanted somewhere I could write my thoughts and think problems through. I could talk to my colleagues when I was at school; in Buffalo I am somewhat isolated, so I did this for about a month, and the blog was serving its purpose. Shortly after though, my internship with The New York Times graphics department began (which I had accepted before I left California), and this is when I really started to learn by blogging.

Documenting My Experiences

Interning at The New York Times was an incredibly exciting and educational experience. My job was to work with reporters and create statistical graphics that
supplemented stories. Every time one my graphics was published, I would post about it. I would write about what went into a graphic like the design principles behind it, the searches for data, fact-checking, data sources, and all of the revisions a graphic went through before going into the paper. For example, there a graphic about the shortage of lifeguards and drownings due to negligent staffing. My first version was dry and lacked a lot of context. I fixed this by providing excerpts on specifics of certain drownings, which provided a more complete story. Documenting such experiences helped me think and forced me extract all the details that I might not have paid much attention to, which in turn helped me improve my skills as a data-minded graphic designer.

In Good Company

After my internship at The Times, there were maybe only three hundred people subscribed to my blog’s RSS feed. I did not mind, because it was more for me than anyone else, so I continued documenting literature as well as any data visualization projects, digital and physical, that I found interesting. Today, in just under a year of blogging, there are over 2,300 subscribers and about six hundred visitors per day. Needless to say, FlowingData has grown into a bit more than a personal blog and provides me a way to teach others about what I do.

It is definitely fun knowing that so many people are reading; however, the best part about blogging is that it allows me to connect with others interested in data visualization. Before I started FlowingData, it always felt like there were so few in the field; it was rare to even find someone who knew about visualization. After blogging though, I am comforted by the fact that the data visualization community is active and growing quickly - not only in statistics, but computer science, art and design, geography, economics, math, and others. There have even been a few occasions when I asked FlowingData readers for help finding literature or ideas on solving a programming problem. There was always someone who knew the answer. So even though I am far away from my statistics department, I am still in good company.

Starting Your Own Blog

Naturally, I encourage others to start their own blogs, because it keeps you thinking and lets you connect with others who are working on and trying to solve the same problems. There are several free blogging platforms available of which Wordpress and Blogger are the most widely used. You can also start a blog under your own domain name for as little as five dollars per month.

People often tell me that they want to start a blog but do not have the time to do all the work. I tell them that once they fit it into their daily schedule, it is actually really easy and that the effort you put in is worth the satisfaction you will get out of blogging. For me, writing for FlowingData has never felt like work.

Relevant Links

- FlowingData, http://flowingdata.com
- Wordpress, http://wordpress.com
- Blogger, http://blogger.com

If you have any comments or questions about blogging or FlowingData, please feel free to email me at nathan@flowingdata.com.

MY CURRENT FAVORITE GRAPH

Vincent Granville

http://api.ning.com/files/oYQF70x3HC4Opvyqo*zMUaenVtCLnCe4KKArKBmCrGU_/vgo3.PNG

The graph shows click scores, from pay-per-click advertising campaigns on search engines and ad networks. The square size represents the number of observations (clicks) in each segment, a segment being a subset of clicks sharing similar attributes. The color represents some measure of goodness of fit: a red square means that the click, given the quality level (represented by the score), is under-performing in terms of conversion rate. X-axis is the score: high score means higher chance for the click to convert in a sale. So, the score represents the odds, for a given click, to result in a conversion, in a way similar to a FICO score.
This graph is featured on the new analytic social networking site, the Analytic Bridge. You can see other graphs on the site at http://www.analyticbridge.com/photo

The click-score graph from the Analytic Bridge site.

The Computing program at JSM

Computing at Denver: Get ready for an exciting program

by Wolfgang Jank, Program Chair

There will be an exciting program at JSM 2008 and our topics fill the entire spectrum between curriculum design, methodological development, and analysis of large and complex real-world problems. In the following I would like to give you a very brief overview (and also a teaser!) of what will happen in Denver.

The invited program consists of 6 extremely entertaining sessions. The first invited session is on Designing Courses on Statistical Computing (Sunday 8/3, 2pm), and it covers the topics Computing in Nonintroductory Stats Classes, Computational Graphics, and Computing in the Graduate Curriculum.

The second invited session on Sunday 8/3, 2pm is on Advances in Functional Data Analysis (FDA) and it will cover topics such as Functional Regression in Marketing, Forecasting with FDA in eCommerce, Modeling Sparse Generalized Longitudinal Observations, and Structural Tests on Functional Variables.

Monday and Tuesday will see two invited sessions that I'd like to especially highlight since they feature prominent speakers from the eCommerce industry, in particular from Google, Yahoo and Facebook. The first session (Monday 8/4, 2pm) covers the Analysis of Massive Social Networks with topics such as Cost-effective Outbreak Detection in Networks, Data Analysis at Facebook, and Driving Marketing Intelligence from Online Discussion. The second session (Tuesday 8/5, 10:30am) is on Statistical Challenges in Online Advertising and Search and covers topics such as Understanding Online Advertisers, Placing Online Advertisements Based on Context, and Social and Semantic Structures in Web Search.

We continue on Wednesday 8/6 (10:30am) with a session on The Analysis of Wikipedia Data, featuring talks such as Wikipedia as a Model for ASA Journals, Quality Assurance in Information Submissions, and Can Statistical Learning Sustain Wikipedia’s Model of Growth?

Thursday 8/7 (10:30am) will see our last invited session on Global Maximization in EM-Type Algorithms. This session will feature talks such as Adaptive Data Weighting Strategies for Locating the Global Maximum in EM-type Algorithms, Global Optimization with Model Reference Adaptive Search and EM, and A Probabilistic Analysis of EM for Mixture in Spherical Gaussians.

We also have several phenomenal topic contributed sessions which span the analysis of large data sets, multivariate outlier detection, machine learning, computing with massive data, analyses of large nonrandomized data sets, the bootstrap for complex problems, least angle regression and the winner of the 2008 statistical computing and graphics student paper competition.
Between that, we also have many, many excellent contributed sessions which cover all areas of computational statistics. I hope that, by now, you are as excited about this year's program as I am, and I hope that I will see many of you in the sessions sponsored by Stat Comp. If you have any questions, please don’t hesitate to contact me at wjank@rhsmith.umd.edu.

See you very soon,
Wolfgang

---

The Graphics program at JSM

Section on Statistical Graphics: JSM 2008 preview

by David Hunter, program chair

Here is a brief synopsis of the JSM events sponsored by the section on statistical graphics. (NB: This list does not include events that the section is co-sponsoring; for a full list, consult the online program at http://www.amstat.org/meetings/jsm/2008/onlineprogram.)

Some of these events – namely, the continuing education courses and roundtables with lunch on Monday and Tuesday – are fee-only events; for information about prices, select “Registration” from the “Conference Information” pulldown menu at the JSM 2008 website.

• **Sunday, Aug. 3:** The invited sessions get underway early with *Visualizing Large Datasets* from 4:00 – 5:50pm. The speakers in this session, representing both industry and academia, will discuss cutting-edge techniques for dealing with streaming data, applying simple models to large dataset visualization, and diagnosing scatterplots (via “scagnostics”) in high-dimensional space.

• **Monday, Aug. 4:** The second of four invited sessions, *Political Science, Statistical Science, and Graphics*, will run from 8:30 – 10:20am.

From a broadly applicable new R package for graphics to a discussion of new developments in cartograms to a discussion of visualization of a massive dataset derived from political speeches in legislatures, this session should lead to lively discussions not only about statistics but about geography and politics as well! Also on Monday, Anthony Unwin from Augsburg and Heike Hofmann from Iowa State will lead a continuing education course all day (8:30am – 5:00pm) entitled *Graphics of Large Data Sets*, and Diane Lambert from Google, Inc. will host a roundtable with lunch (12:30 – 1:50pm) entitled *Data Visualization for the Masses*. The CE course will be based on the book of the same name, and participants are welcome to bring laptops and should have knowledge of standard statistical graphics and experience carrying out data analysis. The roundtable will enable participants to discuss in more depth some of the same themes that will be presented in Wednesday’s invited session on the emergence of social data analysis.

• **Tuesday, Aug. 5:** Today’s invited session, running from 8:30 – 10:20am, is called *Statistics-Geography Mashups on the Web* and comprises four presentations that show how geography and statistics have been closely linked on freely accessible web pages. In addition, Tuesday features two fee-based events with self-descriptive titles: A continuing education course, *Fundamental Statistics Concepts in Presenting Data: Principles for Constructing Better Graphics*, led by Rafe Donahue from Vanderbilt from 8:00am – 12:00; and a roundtable with lunch, *Exploratory Data Analysis (EDA): Graphical Methods, Software, Applications, and Recent Developments*, led by Juergen Symanzik of Utah State from 12:30 – 1:50pm. Finally, Michael O’Connell from Insightful Corporation has organized a topic-contributed session, co-sponsored by the biopharmaceutical section, on *Statistical Graphics for Analysis and Reporting of Clinical Data* from 10:30am – 12:20pm.

• **Wednesday, Aug. 6:** The last of the four invited sessions sponsored by statistical graphics is *The emergence of social data analysis and its impact on the field of Statistics*, from 2:00 – 3:50pm. In this session, representatives from two so-called “web 2.0” sites (Many Eyes and Stat-
Crunch) will discuss the idea of “social data analysis” that allows users to upload/analyze data, publish their results and discuss their findings with other users. Finally, there is a contributed paper session from 8:30 – 10:20am entitled *Graphics for Regression, Classification, and Dimension Reduction*.

My thanks to Web West, Anton Westveld, Heike Hofmann, and Juergen Symanzik for organizing our lineup of invited sessions; to Juergen, Steve MacEachern, and Diane Lambert for organizing our roundtable sessions; to our three CE course instructors; and to Michael O’Connell for organizing a topic-contributed session.

---

**Graphics of Large Datasets**

**GRAPHICS SHORT-COURSE AT JSM**

On Monday, Anthony Unwin from Augsburg and Heike Hofmann from Iowa State will lead a continuing education course all day (8:30am – 5:00pm) entitled *Graphics of Large Data Sets*. This is an update on a course that ran very successfully last year, but this year will be even better! Instead of predominantly "slide and talk" (what is the modern equivalent of the phrase "chalk and talk"?), there will be group discussion sessions using the participants' own laptops. In this case "learning by doing" will involve analyzing datasets together with graphical methods and that's always a lot of fun as well as being very instructive.

Here is the abstract for the class:

> Graphics are great for exploring data, but how can they be used for looking at the large datasets that are commonplace today? Large datasets bring new complications and require different emphases and approaches. This course discusses how to look at ways of visualizing large datasets, whether large in numbers of cases or large in numbers of variables or large in both. It is based on the book Graphics of Large Datasets Unwin AR, Theus, M., Hofmann, H. pub: Springer (2006).

For reviews of the book see: stats.math.uni-augsburg.de/GOLD/Reviews.html.

Data visualization is useful for data cleaning, exploring data, identifying trends and clusters, spotting local patterns, evaluating modeling output, and presenting results. It is essential for exploratory data analysis and data mining. Data analysts, statisticians, computer scientists, indeed anyone who has to explore a large dataset of their own, should benefit from attending this course.

Participants should have a knowledge of standard statistical graphics and experience of carrying out data analysis. Participants are welcome to bring their own laptops to explore graphical analyses for themselves, especially in the discussion sessions.

Either the software Mondrian (which can be downloaded from stats.math.uni-augsburg.de/Mondrian/) or, if you use R, the R package iPlots should be installed.

---

**News**

**ANNUAL COMPETITIONS**

**2008 WINNERS**

**JR Lockwood,**

Awards Officer, 2008

Statistical Computing Section

The Statistical Computing Section of ASA sponsors two 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), and the John M. Chambers Award. Winners of both awards are selected prior to the Joint Statistical Meetings, 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
announced, which went out in September and is available at the Statistical Computing website at http://www.statcomputing.org.

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 2008 Student Paper competition, presented in alphabetical order, were:

- Ming-Hung Kao (advisor John Stufken) "Multi-objective Optimal Experimental Designs for Event-Related fMRI Studies"
- Ernest Kwan (advisor Michael Friendly) "Tableplot: A New Display for Factor Analysis"
- Adam Rothman (advisor Liza Levina) "Sparse Permutation Invariant Covariance Estimation"
- Michael Wu (advisor Xihong Lin) "Two-Group Classification Using Sparse Linear Discriminant Analysis"

The John M. Chambers Award is endowed by Dr. Chambers' generous donation of the prestigious Software System Award of the Association for Computing Machinery presented to him in 1998 for the design and development of the S language. The competition is open to small teams of developers (which must include at least one student or recent graduate) that have designed and implemented a piece of statistical software, with the winner being selected by a panel of three judges, indicated by the section's awards officer. Further details on the requirements for submission and eligibility criteria are provided in the competition's announcement, which is distributed in early October, and at the Statistics Computing website (see above). 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 2008 John M. Chambers Award was:

Alejandro Jara Vallejos (Catholic University of Leuven).

Alejandro received this year's award for his submission "DPpackage," an R package for Bayesian semiparametric and nonparametric models.

I want to thank the judges of both the Student Paper Competition and the John M. Chambers Award for their dedication and efforts to see that the competitions were run fairly and on time. Both competitions had a record number of high-quality submissions and the judges did an incredible job! Congratulations to all of this year's winners and I look forward to next year's competitions.

Meeting Roundup

Graybill VII

by Fred Balch

The 7th Graybill Conference, Named in honor of Professor Franklin A. Graybill was held in Ft Collins Colorado June 11th-13th.


Additionally two short courses were held. Matt Austin and Michael O'Connell gave a short course on "Statistical Graphics for Clinical Data Analysis", and Scott Evans, Lingling Li, L.J. Wei, and Marvin Zelen gave a short course on "Hot Topics in Clinical Trials". A number of additional tutorials in specialized pharmaceuti- cal statistical areas were given by leaders in their fields. The conference was sponsored by American Statistical Association and the Biopharmaceutical Section of the

PAGE 15
ASA and funded in part by Amarex Clinical Research, AMGEN, Eisai Medical Research, Inc., Gilead Colorado, Inc. and Takeda Pharmaceuticals N. America, Inc. The website for the conference is http://www.stat.colostate.edu/graybillconference/ for further details.

ICSA 2008 Symposium

by Jun Zhao

The 17th annual ICSA Applied Statistics Symposium was held June 5-7, 2008 at Embassy Suites Piscataway, NJ. The symposium delivered six excellent short courses, four keynote speeches and fifty-four parallel scientific sessions. There were also several special events including one mixer night, one career event, and one banquet night with a special presentation. The symposium set platforms for all attendees to share their cutting-edge scientific research and application techniques, to meet with old and new friends, and to learn statistical knowledge in both methodology and application. A total of 332 statisticians participated in the conference.

Four distinguished keynote speakers delivered excellent speeches in two keynote sessions in the morning of day two and three (June 5th and 6th) including Dr. Robert O’Neill from the FDA, Prof. Jianqing Fan from Princeton University, Prof. David DeMets from University of Wisconsin, and Prof. Zhiliang Ying from Columbia University.

54 scientific sessions were parallel allocated to 11 time periods on day two to day four (June 5th - 7th). The speakers, lots of them are prominent members of our profession, were from government, academia and industry. The speakers presented scientific research and application techniques on Statistics, discussed statistical methodology and application in drug development, financial and other areas, and shared experiences such as career development.

The student award participators submitted high quality research papers for competition. Four students emerged and won awards and travel grants. The winners were Sijian Wang (who won the J.P. Hsu Memorial Scholarship) from the University of Michigan, Michael Wu from Harvard University, Min Yuan from University of Science and Technology of China, and Megan Othus from Harvard University.

Puzzles

We had quite a bit of interest in the puzzle from the last issue of the newsletter. However, no-one was able to do much better than 2sqrt(2). Here is the question again and the solution as provided by Ihab Girgis from Johnson and Johnson at the S-PLUS Insightful Impact Meeting.

Following this is a new puzzle, contributed again by Stephen Kaluzny. Please submit your answers to Michael O’Connell (moconnell@insightful.com) for publication in the next newsletter. We are also interested in hearing from anyone who has any interesting puzzles to share in this section of the newsletter.

Question 1
Given a square piece of property of unit side you wish to build fences so that it is impossible to see through the property, i.e. there is no sightline connecting two points outside the property and passing through the property that does not intersect a fence. The fences do not have to be connected and several fences can come together at a point. The fences can be placed in the interior of the property, they aren’t restricted to the boundary. What is the minimum total length of fencing required and how is it arranged. For example you could place fencing along all four sides. This would have total length 4 but is not the best possible. Hint: You can do better than 2sqrt(2)

Answer 1
Some thought about the problem leads one to consider a solution as shown in fenceSolution1.png. You need to find the value of (x0, y0) that minimize L1 + L2 + L3. The distance L4 is fixed at sqrt(2)/2. So the total fence length would be L1 + L2 + L3 + L4.
If you set \((x_0, y_0) = (0.5, 0.5)\) you get \(L_1 = L_2 = L_3 = L_4\) and a length of \(2 \times \sqrt{2}\) but the hint says you can do better than that.

\[
\begin{align*}
\text{Possible Solution to the Fence Puzzle} \\
\text{Find } (x_0, y_0) \text{ to minimize length } = L_1 + L_2 + L_3
\end{align*}
\]

sqrt(2)/2 \approx 2.638959

And the three fence lines join at the point \((0.2113249, 0.2113249)\).

**Question 2**

A banana plantation is located next to a desert. The plantation owner has 3000 bananas that he wants to transport to the market by camel, across a 1000 kilometer stretch of desert. The owner has only one camel, which carries a maximum of 1000 bananas at any moment in time, and the camel eats one banana every kilometer it travels.

What is the largest number of bananas that can be delivered at the market?

Hint: You can cache bananas in the desert.

We can use `nlminb` in S-PLUS to find \((x_0, y_0)\) that minimizes \(L_1 + L_2 + L_3\). By symmetry we can assume \(x_0 = y_0\) so we want to the minimized this univariate function:

```
"fenceLen3" <- function(x0) {
  y0 <- x0
  dist3 <- c((x0 - 0)^2 + (y0 - 0)^2, (x0 - 1)^2 + (y0 - 0)^2, (x0 - 0)^2 + (y0 - 1)^2)
  sum(sqrt(dist3))
}
```

The call:

```
nlminb(start = 0.5, objective = fenceLen3, lower=0)
```

includes this partial output:

```
$parameters:
 [1] 0.2113249

$objective:
 [1] 1.931852

$message:
 [1] "RELATIVE FUNCTION CONVERGENCE"
```

So the total fence distance required is 1.931852 +
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:

Michael O’Connell, Editor, Statistical Computing Section
moconnell@insightful.com

Andreas Krause, Editor, Statistical Graphics Section
andreas.krause@actelion.com
http://www.elmo.ch

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