Volume 23, July 2013

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

Statistical COMPUTING & GRAPHICS

A Word from our 2013 Section Chairs

Montserrat Fuentes
COMPUTING

Webster West
GRAPHICS

Welcome to the Statistical Computing and Statistical Graphics Newsletter. We hope you enjoy this way to keep you informed of the exciting news of our sections.

I would like to take this opportunity to announce that Phil Chalmers, from the Psychology Department of York University, Canada, is the winner of the 2013 John M. Chambers Statistical Software Award sponsored by the Section on Statistical

Continued on page 2 . . .

This year's JSM in Montreal is shaping up to be a very good one for the Graphics Section and for our partners in the Computing Section. The Graphics Section will be sponsoring three invited sessions and three contributed sessions as well as cosponsoring more than 20 more sessions. Some of the graphics highlights include sessions on visualizing structure in complex data and visualizing life in the United States. Given the ever increasing popularity

Continued on page 2 . . .

Contents of this volume:

From Montse Fuentes, Computing Chair
(continued from page 1)

Computing. His paper is titled "Unidimensional and Multidimensional Item Response Modeling with the mirt Package." The award will be presented at the Section's business meeting/mixer on Monday evening.

The Statistical Computing Section is also sponsoring student paper awards that will be recognized at JSM during our business meeting. There are three computing papers: "Are you Normal, The problem of confounded residual structures in hierarchical models" by Adam Loy from Iowa State University; "Fast and stable multiple smoothing parameter selection in smoothing spline analysis of variance models with large samples" by Nathaniel Helwig from University of Illinois at Urbana-Champaign; and "Time-varying networks estimation and dynamic model selection" by Xinxin Shu from University of Illinois at Urbana-Champaign.

The Statistical Computing Section jointly with the Statistical Graphics section and the Knight Foundation are sponsoring the 2013 Data Expo. The goal of this initiative is to recognize through a poster competition the best way to generate graphical summaries of important features of a data set. The competition will take place at JSM and there will be cash prizes awarded to the best posters, as judged by a panel of experts. The cash awards for the top three entries are $1,500, $1,000, and $500. In addition, after JSM we will organize a special journal issue, tentatively in Computational Statistics and Data Analysis, and the best poster entries will receive an invitation to publish their work in a journal article in this special issue. Please, do not miss our 2013 Data Expo at JSM!

Finally, I would like to also announce the results of the ASA election for the Statistical Computing Section and congratulate the selected officers:

- Chair-Elect, 2014: David van Dyk
- Program Chair-Elect, 2014: Feng Liang
- Secretary/Treasurer, 2014-2015: Tim Hesterberg
- Publications Officer, 2014-2016: Usha Govindarajulu
- Council of Sections Representative, 2014-2016: John Monahan

We appreciate you taking a few minutes of your valuable time with us and please stay in touch.

Montse Fuentes, Chair
Section on Statistical Computing

From Webster West, Graphics Chair
(continued from page 1)

of big data problems, it is also very timely that we will have an invited session on visualizing big data interactively.

This year the Data Expo will feature data from the "Soul of the Community" project sponsored by the Knight Foundation in cooperation with Gallup. The data set contains information from 43,000 people over three years in 26 communities across the United States. Stop by the Data Expo to find out why some communities thrive and others do not and to see what makes people commit to their community for the long term. Many thanks to Kary Myers for putting together the graphics program this year and to Heike Hoffman for organizing the Data Expo.

It is also already time to start thinking about the 2014 JSM to be held August 2nd through the 7th in Boston. If you have ideas for sessions, please contact our program chair-elect Michael Kane (`michael.kane@yale.edu`). We can always use more submissions from graphics. Also, keep your eyes out for next years Data Expo, the details of which will be announced this Fall.

I would like to take this opportunity to extend my congratulations to Abbass Sharif from Utah State University. His paper entitled, "Multivariate Visual Data Mining Tools for Functional Actigraphy Data," was selected as the sole graphics winner of this years student paper competition cosponsored by the graphics and computing sections. This year we had 28 submissions, a big increase from previous years. Many thanks to Fei Chen and to the judges for their hard work on this competition. If you are a student or you have students that are doing work with a strong graphical component please keep next year's competition in mind. The announcement will be coming this Fall.

The recent ASA election will bring in a new group of folks to serve the graphics section in the coming years. The Chair-Elect, for 2014 is Naomi Robbins, the Program Chair-Elect for 2014 is Matt

Shotwell, the Council of Sections Rep. for 2014-2016 is Rebecca Nugent and the Publications Officer for 2014-2015 is Anushka Anand. We should all thank these individuals for their willingness to serve our section in these capacities.

We are always looking for ways to improve our section and to better serve our members. If you have ideas, please feel free to put them forward. The joint mixer on Monday evening at JSM is an excellent opportunity to express your ideas and to get together with fellow graphics and computing members. I hope to see all of you there.

Webster West, Chair
Section on Statistical Graphics

Preview: JSM 2013

The Statistical Computing/Graphics Business Meeting and Mixer

Monday, August 5, 2013, 6:00-8:00 PM in W-Fortifications (located in Le Westin Montréal according to the Online Program). We look forward to seeing you there!

Stat Computing program

The 2013 Joint Statistical Meetings in Montreal are just around the corner. We have an exciting program this year, with 6 Invited Sessions – the maximum possible for a section of our size.

- "Strategies for Large Scale Numerical and Statistical Computing" organized by Michael Kane (Yale University), Sunday 4:00 PM. Speakers: Bryan Lewis, George Ostrouchov, Luke Tierney, Norm Matloff.
- "Large Scale Statistical Computing: Methodologies, Tools, and Applications" organized by Landon H Sego (Pacific Northwest National Laboratory), Monday 8:30 AM. Speakers: Bill Cleveland, Saptarshi Guha, Ryan Hafen.
- "The Secret Weapon of the Dark Knight Against the Joker: Statistical Methods for Big and Massive Data Sets" organized by Xingye Qiao (SUNY Binghamton), Tuesday 10:30 AM. Speakers: Bin Yu, David Dunson, David Madigan, Yoshua Bengio
- "Large Scale Inference" organized by Loki Natarajan (University of California San Diego), Wednesday 10:30 AM. Speakers: Brad Efron, Laura Lazzeroni, Karen Messer (Discussant).
- "Computational Statistics in the Atmospheric and Oceanic Sciences" organized by Michael L Stein (University of Chicago), Wednesday 2:00 PM. Speakers: Alan Gelfand, Michael Stein, Dorit Hammerling, John Lindstrom.
- "Recent Advances in Bayesian Computation" organized by Dawn B Woodard (Cornell University), Thursday 8:30 AM. Speakers: Faming Liang, Yuguo Chen, Dawn Woodard, John Paisley.

There is an additional Invited Session submitted to us that was picked up by the Section on Nonparametrics that I expect will be of interest to many of our members:

- "Taming Big Data with Matrix and Tensor Decomposition Methods" organized by George Luta (Georgetown University), Tuesday 2:00 PM. Speakers: Eric Lock, Vadim Zipunnikov, Jianhua Huang, Andrzej Cichocki.

We also have 6 Topic Contributed Sessions. One of these features the winners of our Student Paper Competition (Computing and Graphics), held on Tuesday morning at 8:30. Please come out and support our budding members! Other sessions include "Challenges in using Markov chain Monte Carlo in Modern Applications," "New Robust Methods in Biostatistics," "Recent Advances in Likelihood-based Inference in Mixed Models using Data Cloning," "Bayesian Computations: Challenges, Solutions and Implementations," and "Adaptive Monte Carlo Methods for Bayesian Computation."

Finally, we have a total of 9 Contributed Sessions, grouped into sessions in areas such as

Big Data, Bayesian Modeling, Machine Learning, MCMC, and Software.

As Program Chair 2013 for Computing I would like to thank: Chris Volinsky, Program Chair 2012, for helping me get started in this role; everyone who submitted proposals throughout the process, particularly under the pressure of various timelines; our Session Chair volunteers for their upcoming efforts in Montreal; Rick Peterson for helping to keep the Section running smoothly; and Naomi Friedman, Donna Arrington, and the ASA Staff for their excellent organization of the process.

John W. Emerson, Computing Program Chair 2013
Yale University

Stat Graphics program

As program chair for Statistical Graphics for JSM 2013, I'm pleased at the terrific set of topics featured in our allocation of 3 invited sessions in Montreal:

- "Visualizing Big Data Interactively" organized by Kary Myers (Los Alamos National Laboratory), Monday 2:00 PM. Speakers: Simon Urbanek, AT&T Labs; Leanna House, Virginia Tech; Timo Bremer, Lawrence Livermore National Laboratory.
- "Painting a Picture of Life in the United States" organized by Heike Hoffman (Iowa State University), Tuesday 2:00 PM. Speakers: Howard Hogan, Eric Newburger, and Michael Ratcliffe, U.S. Census Bureau; Richard Heiberger, Temple University; Heike Hoffman, Iowa State University.
- "Visualization of Structure in Complex Data" organized by David Collins (Los Alamos National Laboratory), Wednesday 8:30 AM. Speakers: Aparna Huzurbazar and David Collins, Los Alamos National Laboratory; Ilya Shpitser, Harvard School of Public Health; Steve Marron, The University of North Carolina

Another highlight will be **2013 Data Expo**, organized by Heike Hoffman. This year's event will showcase data from the John S. and James L. Knight Foundation's Soul of the Community project. Competitors will present visualizations of this data set in a topic-contributed poster session on Monday at 2:00 PM. The session features 17 teams and $3000 in prizes. Stop by to see the results, and check out `http://streaming.stat.iastate.edu/dataexpo/2013/` for more information.

Our section is also sponsoring a contributed session covering a range of graphics topics, to be held Wednesday at 10:30 AM. And we're cosponsoring several sessions from other sections featuring interesting visualization and graphics topics, including invited sessions on "Analytics and Data Visualization in Professional Sports", "Graphical Approaches for Survey Data", and "Seeing in and Beyond R". Use the search tool at `http://www.amstat.org/meetings/jsm/2013/onlineprogram/` to get a complete list of sessions cosponsored by Statistical Graphics.

Kary Myers, Graphics Program Chair 2013
Los Alamos National Laboratory

Statistical Graphics Video Competition

"What did you do this summer?" The GGobi Foundation encourages you to submit a video showing how you have used interactive graphics this summer. First prize is $500 and second prize is $100; the winners will be announced at the Statistical Computing and Graphics mixer at JSM 2013, Monday August 5, at 7 PM. The submission deadline is August 1, 2013. More information is available on the web: `http://streaming.stat.iastate.edu/~dicook/video-competition/`.

Dianne Cook
Iowa State University

2013 Data Expo

The Sections on Statistical Computing and Graphics are running the 2013 Data Expo entitled "Soul of the Community." The competition is underway (abstracts declaring participation) were submitted in February, and posters will be on display and judged at JSM 2013. The competition invites graphical summaries of important features of data provided by the Knight Foundation in cooperation with Gallup on 43,000 people over three years in 26 communities across the US. Prizes of $1,500, $1,000,

and $500 will be awarded. More information on the competition is available on the web: http://streaming.stat.iastate.edu/dataexpo/2013/.

Heike Hofmann
Iowa State University

Continuing Education at JSM 2013

Murray Stokely will be teaching a course titled "Practical Software Engineering for Statisticians." The course will be Mon, 8/5/2013, 1:00 PM - 5:00 PM at Le Westin Montréal Palais room.

Statisticians are increasingly being employed alongside software engineers to make sense of the large amounts of data collected in modern e-commerce, internet, retail, and advertising companies. This course introduces a number of best practices in writing statistical software that is taught to computer scientists, but which is seldom part of a statistics degree. Revision control tools, unit testing, code modularity, structure, and readability, and the basics of computer architecture and performance will be covered. A few examples of real R code written in a commercial environment will be shared and discussed to illustrate some of the problems of moving from working alone or in a small group in an academic setting into a team in a large commercial setting. Some basic familiarity with programming is required. The course is language-agnostic, but R will be used in some examples.

Registration information is available online: http://www.amstat.org/meetings/jsm/2013/registration.cfm

Journal News

Journal of Computational and Graphical Statistics

The purpose of the Journal of Computational and Graphical Statistics (JCGS) is to improve and extend the use of computational and visualization methods in statistics and data analysis. The journal publishes articles that make original contributions to one or both of the fields of computational statistics and data visualization. Currently we are particularly interested in articles that address various computational and/or visualization issues for big data, broadly interpreted as complex or massive data sets.

JCGS kicked off the International Year of Statistics early with an issue devoted to networks and network models, Volume 22, issue 4, December 2012. In the 2013 JCGS volume, we will feature a special issue on Advances in MCMC (23-3, September 2013). Also check out our March 2013 issue which features a discussion article on InfoVis. Finally, JCGS will again host an invited session at JSM, "JCGS Selections: Lassoing the Year of Statistics into an Elastic Net."

Thomas Lee and Richard Levine, jcgs@ucdavis.edu

The R Journal

Volume 5, Issue 1 of the R Journal is available at http://journal.r-project.org/archive/2013-1/. As well as a new one column style, it also contained a record 20 articles, which I've roughly categorised below.

Visualisation:

- "**osmar**: OpenStreetMap and R" by Manuel J. A. Eugster and Thomas Schlesinger
- "Let Graphics Tell the Story - Datasets in R" by Antony Unwin, Heike Hofmann and Dianne Cook
- "**ggmap**: Spatial Visualization with **ggplot2**" by David Kahle and Hadley Wickham

Text mining:

- "**RTextTools**: A Supervised Learning Package for Text Classification" by Timothy P. Jurka, Loren Collingwood, Amber E. Boydstun, Emiliano Grossman and Wouter van Atteveldt
- "**RcmdrPlugin.temis**, a Graphical Integrated Text Mining Solution in R" by Milan Bouchet-Valat and Gilles Bastin

Modelling:

- "Multiple Factor Analysis for Contingency Tables in the **FactoMineR** Package" by Belchin Kostov, Mónica Bécue-Bertaut and François Husson
- "Hypothesis Tests for Multivariate Linear Models Using the **car** Package" by John Fox, Michael Friendly and Sanford Weisberg
- "Estimating Spatial Probit Models in R" by Stefan Wilhelm and Miguel Godinho de Matos
- "Fast Pure R Implementation of GEE: Application of the **Matrix** Package" by Lee S. McDaniel, Nicholas C. Henderson and Paul J. Rathouz
- "**ftsa**: An R Package for Analyzing Functional Time Series" by Han Lin Shang

The R ecosystem:

- "Statistical Software from a Blind Person's Perspective" by A. Jonathan R. Godfrey
- "Possible Directions for Improving Dependency Versioning in R" by Jeroen Ooms
- "Translating Probability Distributions: From R to BUGS and Back Again" by David LeBauer, Michael Dietze, and Ben Bolker

And a range of papers applying R to diverse subject areas:

- "**PIN**: Measuring Asymmetric Information in Financial Markets with R" by Paolo Zagaglia
- "Generalized Simulated Annealing for Global Optimization: The **GenSA** Package" by Yang Xiang, Sylvain Gubian, Brian Suomela and Julia Hoeng
- "**QCA**: A Package for Qualitative Comparative Analysis" by Alrik Thiem and Adrian Duşa
- "An Introduction to the **EcoTroph** R Package: Analyzing Aquatic Ecosystem Trophic Networks" by Mathieu Colléter, Jérôme Guitton and Didier Gascuel
- "**stellaR**: A Package to Manage Stellar Evolution Tracks and Isochrones" by Matteo Dell'Omodarme and Giada Valle
- "**mpoly**: Multivariate Polynomials in R" by David Kahle
- "The **beadarrayFilter**: An R Package to Filter Beads", by Anyiawung Chiara Forcheh, Geert Verbeke, Adetayo Kasim, Dan Lin, Ziv Shkedy, Willem Talloen, Hinrich W.H. Goehlmann, and Lieven Clement.

Hadley Wickham
The R Journal

The Foundation for Open Access Statistics

The Foundation for Open Access Statistics (FOAS) FOAS promotes open access publishing (without costs to the reader and author), open source software (GPL licensed), and reproducibility of published results. Currently the only project we financially support is the

- Journal of Statistical Software `http://www.jstatsoft.org`,

but we have a number of affliated projects we support.

- UseR! 2014 `http://user2014.stat.ucla.edu`
- Journal of Environmental Statistics `http://www.jenvstat.org`
- OpenIntro `http://www.openintro.org`
- Spatial Demography `http://spatialdemography.org`
- RKWard `http://rkward.sourceforge.net`
- Project MOSAIC `http://mosaic-web.org`

Individuals can help spread the mission of FOAS by becoming a member, subscribing to our mailing list, adding yourself to our Facebook group, and, of course, send us your tax-deductible donation to support our projects. It will also help if you announce your membership on your personal webpage and on your professional C.V. Feel free to use our logos and banners. Open Access, Open Source, and Reproducbility Projects that are of interest to statisticians can request by email to become FOAS-affiliated projects.

For more information, visit the FOAS website: `http://www.foastat.org/`.

Jan de Leeuw
FOAS and The Journal of Statistical Software

Announcements

Pre-JSM Conference: Statistical Science in Society

In June 2012, the Statistical Society of Canada approved the formation of a separate entity, the Canadian Statistical Sciences Institute/Institut canadien des sciences statistiques (CANSSI/INCASS), which will work with the Fields Institute, the Centre de Recherches Mathématiques, and the Pacific Institute for the Mathematical Sciences, to promote and sponsor research in the statistical sciences across Canada. CANSSI was officially launched in November 2012 and the process of identifying directions for research in Canada over the next few years has begun. The University of Waterloo, with support from the Fields Institute, is hosting a conference to celebrate the International Year of Statistics and the launch of CANSSI.

All are welcome to attend the Statistical Science in Society conference at Waterloo to celebrate the International Year of Statistics and the launch of the Canadian Statistical Sciences Institute (CANSSI), just before the JSM opens in Montreal. The dates are Wednesday July 31 to Friday August 2, 2013.

The conference features outstanding speakers in sessions on social networks, statistics in health, modelling risk, exploring complex models and data, modelling dependence and statistical learning. Registration is open until July 15. For the program and registration details, see: `http://math.uwaterloo.ca/statistics-and-actuarial-science/canssi2013`

Mu Zhu
University of Waterloo

Planning for JSM 2014

Graphics and Computing have always been at the core of statistical practice. In the last few years terms like "data science" and "big data" have been absorbed into the mainstream, meaning computing and graphics are seeing a resurgence in importance. As members of the statistical and graphical computing community, who been fully engaged in this exciting and relevant area of research, we should realize that this is a tremendous opportunity for us to share our accomplishments with a much larger audience, both in the statistics community and beyond. With this in mind, we would like to encourage statistical researchers with interest in visualization and computing to submit proposals for sessions at next year's JSM in Boston Massachusetts. The computing section is particularly interested in scalable methods, optimization, algorithms, programming languages, and interesting applications. The graphics section is interested in proposals in the area of interactive visualization over the web, graphical exploration of large data sets, and new directions in visualization. Both sections are particularly interested in talks that integrate emerging technologies. We look forward to seeing your proposals and attending your sessions at next year's JSM.

Michael Kane (`michael.kane@yale.edu`*) and Nicholas Lewin-Koh (*`lewin-koh.nicholas@gene.com`*)*

2014 Conference on Statistical Practice

The ASA announces the Conference on Statistical Practice Innovations and Best Practices for the Applied Statistician, February 20-22, 2014, in Tampa Florida. Statistical Practice 2014 brings together hundreds of statistical practitioners – including data analysts, researchers, and scientists – who engage in the application of statistics to solve real-world problems. The conference will provide an opportunity to learn about the latest statistical methodologies and best practices in statistical design, analysis, programming, and consulting. In addition to concurrent sessions revolving around four statistical themes, the conference will offer short courses on Thursday, February 20 and tutorials on Saturday, February 22. There will also be plenty of time to network and socialize in the exhibit hall. Registration will open on October 1, 2013. For more information about the Conference on Statistical Practice please visit `http://www.amstat.org/meetings/csp/2014/index.cfm`.

Cheryl Behrens
American Statistical Association

From Revolution Analytics

Revolution Analytics has been busy celebrating the International Year of Statistics with the opening of our London office serving the EMEA market (http://tinyurl.com/pugty83) and the opening of our joint Center of Excellence in Singapore with Dell and Intel (http://tinyurl.com/mzk3upq). Our flagship product, Revolution R Enterprise software (http://www.revolutionanalytics.com/products/revolution-enterprise.php), builds on the power of R with performance, scalability and enterprise readiness. It has recently been updated to 6.2 and we invite you to find out about the new features by viewing our webinar recording (http://www.revolutionanalytics.com/news-events/free-webinars/2013/revolution-r-enterprise-6.2/). Revolution Analytics provides a free version of our software for Academic users. You can sign up and download your copy here: http://info.revolutionanalytics.com/free-academic.html.

We are very pleased to be a Gold Sponsor of JSM. Revolution Analytics is actively recruiting R course instructors and we are also hiring! We hope you will stop by our booth on the exhibition floor and say hello!

Terry Christiani (terry@revolutionanalytics.com)
Revolution Analytics

Computing News from SAS

This article describes the following items that might be of interest to members of the Graphics and Computing sections who use SAS® software:

- An overview of the 12.1 release of SAS analytical products, which were released in September 2012
- A selection of computational and graphical papers from the SAS Global Forum 2013 conference
- A list of SAS courses that will be presented at JSM in Montreal
- A description of the forthcoming SAS 9.4 software and SAS/STAT® 12.3, which include new high-performance analytical procedures

If you have time to read only one paper from SAS Global Forum 2013, read "Current Directions in SAS/STAT Software Development" by Maura Stokes (http://support.sas.com/resources/papers/proceedings13/432-2013.pdf), which summarizes how SAS/STAT software has changed in recent years and how it is evolving for the future.

To stay up to date with SAS activities throughout the year, subscribe to the SAS Statistics and Operations Research Newsletter (http://support.sas.com/community/newsletters/stats/index.html). For a fun SAS-oriented blog that highlights statistical computation, programming, and graphics, subscribe to Rick Wicklin's blog, The DO Loop (http://blogs.sas.com/content/iml/).

The 12.1 Release of SAS Analytical Products

SAS analytical products now have their own release numbering scheme that is independent of Base SAS. There is a ton of new functionality in the 12.1 release (http://support.sas.com/rnd/app/analytics/12.1/new.html), but a few features that are worth highlighting include

- Quantile regression for censored data (the QUANTLIFE procedure)
- Model selection for quantile regression (the QUANTSELECT procedure)
- Finite mixture models (the FMM procedure)
- Bayesian modeling of missing data (the MCMC procedure)
- Improved support for mosaic plots and heat maps

For more information about SAS/STAT 12.1, see the paper "Look Out: After SAS/STAT 9.3 Comes

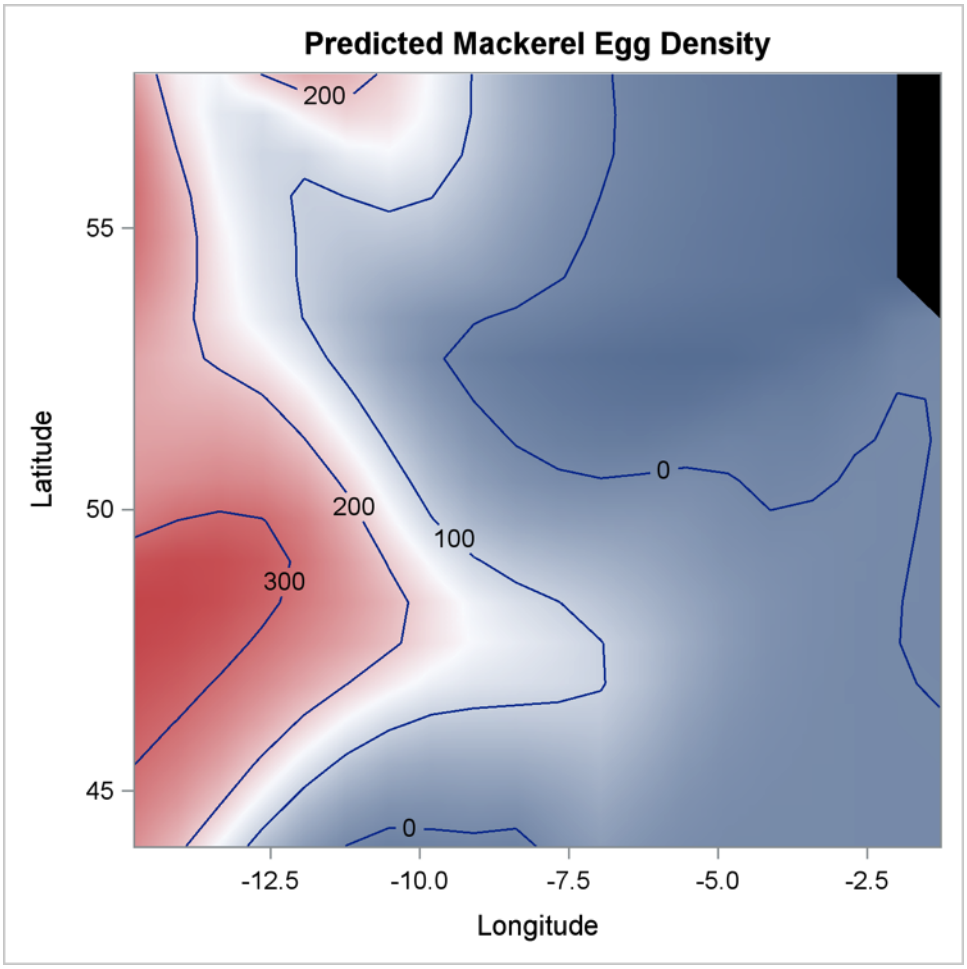

Figure 1: Contour Plot of Predicted Values from a Nonparametric Poisson Regression Model as Fitted by the ADAPTIVEREG Procedure

SAS/STAT 12.1!" (`http://support.sas.com/resources/papers/proceedings12/313-2012.pdf`).

Computational and Graphical Papers from the SAS Global Forum 2013

SAS Global Forum 2013 featured more than 500 papers that were presented in 22 categories. Readers of this newsletter might enjoy browsing papers in the following categories:

- SAS and Big Data (`http://support.sas.com/resources/papers/proceedings13/#SASDat`). In particular, I recommend:
  - "High-Performance Statistical Modeling" (`http://support.sas.com/resources/papers/proceedings13/401-2013.pdf`) by Bob Rodriguez and Robert Cohen, which describes new high-performance analytical procedures that are included in SAS/STAT 12.3
  - "Hadoop and SAS" (`http://support.sas.com/resources/papers/proceedings13/402-2013.pdf`) by Paul Kent, which contains the slides from Kent's presentation
  - "Uncovering Patterns in Textual Data with SAS Visual Analytics and SAS Text

Analytics" (http://support.sas.com/resources/papers/proceedings13/403-2013.pdf) by Dan Zaratsian, Mary Osborne, and Justin Plumley, which presents a humorous case study about visualizing unstructured data in the form of 4.85 million tweets during Super Bowl XLVII

- Statistics and Data Analysis (http://support.sas.com/resources/papers/proceedings13/#Stats). In particular, I recommend:
  - "Creating and Customizing the Kaplan-Meier Survival Plot in PROC LIFETEST" (http://support.sas.com/resources/papers/proceedings13/427-2013.pdf) by Warren Kuhfeld and Ying So, which shows how to modify graph templates that are used by SAS procedures
  - "Having an EFFECT: More General Linear Modeling and Analysis with the New EFFECT Statement in SAS/STAT Software" (http://support.sas.com/resources/papers/proceedings13/437-2013.pdf) by Phil Gibbs, et al., which describes how to use the new EFFECT statement to fit models that have nonparametric regression effects, crossover and carryover effects, and complicated inheritance effects
  - "Ordinal Response Modeling with the LOGISTIC Procedure" (http://support.sas.com/resources/papers/proceedings13/446-2013.pdf) by Bob Derr, which describes how you can use the LOGISTIC procedure to model ordinal responses

In addition, the paper "Make a Good Graph" (http://support.sas.com/resources/papers/proceedings13/361-2013.pdf) by Sanjay Matange reviews best practices for producing effective graphics for business and statistical analyses.

Activities at the 2013 Joint Statistical Meetings

The following tutorials will be presented at the 2013 Joint Statistical Meetings, Aug. 3-8 in Montreal:

- Model Selection with SAS/STAT Software, by Funda Güneş
- Creating Statistical Graphics in SAS, by Warren Kuhfeld
- Structural Equation Modeling Using the CALIS Procedure in SAS/STAT Software, by Yiu-Fai Yung
- SAS Procedures for Analyzing Survey Data, by Pushpal Mukhopadhyay
- Practical Bayesian Computation, by Fang Chen
- Techniques for Simulating Data in SAS, by Rick Wicklin, which is based on his new book, **Simulating Data with SAS**. (https://support.sas.com/pubscat/bookdetails.jsp?pc=65378)

Preview of SAS 9.4

You can get a preview of SAS 9.4 by reading "Current Directions in SAS/STAT Software Development" (http://support.sas.com/resources/papers/proceedings13/432-2013.pdf) by Maura Stokes. Many SAS/STAT procedures have been multithreaded since SAS 9 was released in 2004. SAS 9.4 (which includes SAS/STAT 12.3 software) includes new procedures designed specifically to operate on data that are stored in databases such as Teradata, Greenplum, and Hadoop. These procedures can use multiple parallel-processing techniques across a grid of servers and will be used by large corporations.

The big news for analysts at smaller institutions is that these procedures will be available to all SAS/STAT customers for use in single-machine mode. There are no additional fees associated with using these new procedures in single-machine mode. These procedures include new functionality such as model selection for generalized linear models (the HPGENSELECT procedure) and decision tree models (the HPSPLIT procedure). For a complete list of the new high-performance analytical procedures and a performance comparison with traditional SAS/STAT procedures, see the paper "High-Performance Statistical Modeling" (http://support.sas.com/resources/papers/proceedings13/401-2013.pdf) by Bob Rodriguez and Robert Cohen.

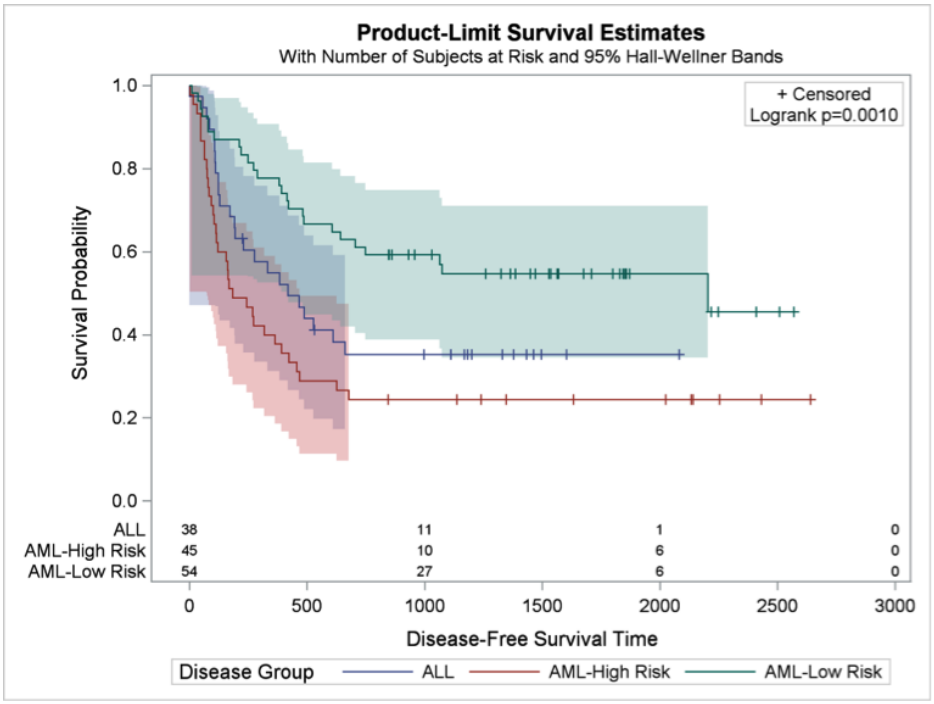

Figure 2: Customized Kaplan-Meier Survival Plot from Kuhfeld and So (2013)

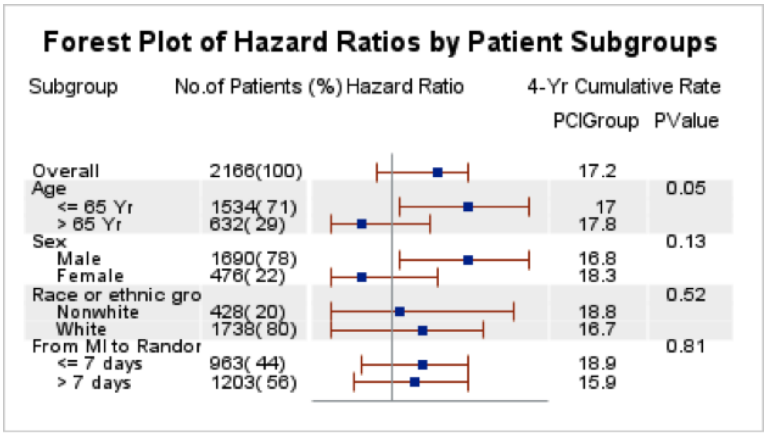

Figure 3: Forest Plot of Hazard Ratios, Which Uses the New AXISTABLE Statement to Display the Textual Columns

On the graphics side, SAS 9.4 contains many enhancements to the statistical graphics procedures (`http://support.sas.com/rnd/datavisualization/papers/sgf2013/Handouts/SGF2013_SGHandout.pdf`) such as PROC SGPLOT and PROC SGPANEL. There are also many enhancements to the Graph Template Language (`http://support.sas.com/rnd/datavisualization/papers/sgf2013/Handouts/SGF2013_GTLHandout.pdf`), which underlies all of the ODS graphics that SAS software produces.

Rick Wicklin (`rick.wicklin@sas.com`)
SAS

News from R Studio

RStudio is committed to helping facilitate the adoption of R and improving the user experience within the R ecosystem. In the past year, we have released tools falling into two categories: data analysis and code/package development. Shiny provides a framework for interactive visualization and data analysis. The evolving RStudio IDE offers enhanced capabilities for code development and package authoring. Below, we discuss some of the highlights of these new contributions. Everything is freely available under open source licenses and can be downloaded from `www.rstudio.com`.

In August of 2012 we joined forces with Hadley Wickham, author of some of the more widely used R packages (ggplot2 and plyr). Hadley is the Chief Scientist of RStudio and is continuing to work on solving a wide variety of development and data analysis problems.

Shiny

Shiny is a package that provides a framework for interactive visualization and data analysis. It is available as an R package; an open source server is provided for web deployment. The example in Figure 1 is a Shiny application built by Joshua Katz, a PhD student at NC State University working with Brian Reich. The application displays heat maps of the results from a survey on the similarities and differences in language around the United States.

Building such an app is straightforward. With Shiny, no knowledge of HTML, CSS, or JavaScript is required. The user interface is fully customizable, allowing the programmer to make use of any JavaScript visualization library. The Shiny framework automatically "reacts" to a variety of changes, including user input or even changes in the underlying data structures. This allows the R programmer to concentrate on the nature of the interactivity without worrying about event handlers or data transfer. Thus, Shiny bridges the statistical power of R and the interactive visualization of JavaScript.

Development tools

RStudio recently improved tools available for package development. We've introduced a build pane with package development commands and a view of build output and errors. There are also a number of tools for creating R package documentation including previewing, spell-checking, and roxygen-aware editing. We've also updated devtools (`https://github.com/hadley/devtools`), an R package that automates much of the package development process. R packages can actually be very simple to develop, and with the right tools it should be easier to use the package structure than not.

RStudio now provides an elegant working environment for C/C++ development, including syntax highlighting, quick navigation to compiler errors and warnings, and tight integration with the Rcpp package.

We also made many improvements to the core development experience including an overhauled find and replace feature, more intelligent automatic-indentation, and a new Vim editing mode. Looking to the future, our upcoming v0.98 release (preview available at `http://www.rstudio.com/ide/download/preview`) includes extensive debugging tools and a revamped workspace pane. The new release also includes the ability to create R presentations. The presentation feature enables users to weave together text, R code, graphics and even LaTeXequations into a presentation that can be played back either as a standalone HTML5 presentation in a web browser, or within RStudio itself.

Reference

Katz, Joshua. "Dialect Survey Maps." RStudio Shiny Hosting, June 2013. Web. 01 July 2013. `http://spark.rstudio.com/jkatz/SurveyMaps/`.

Tareef Kawaf (tareef@rstudio.com)
R Studio

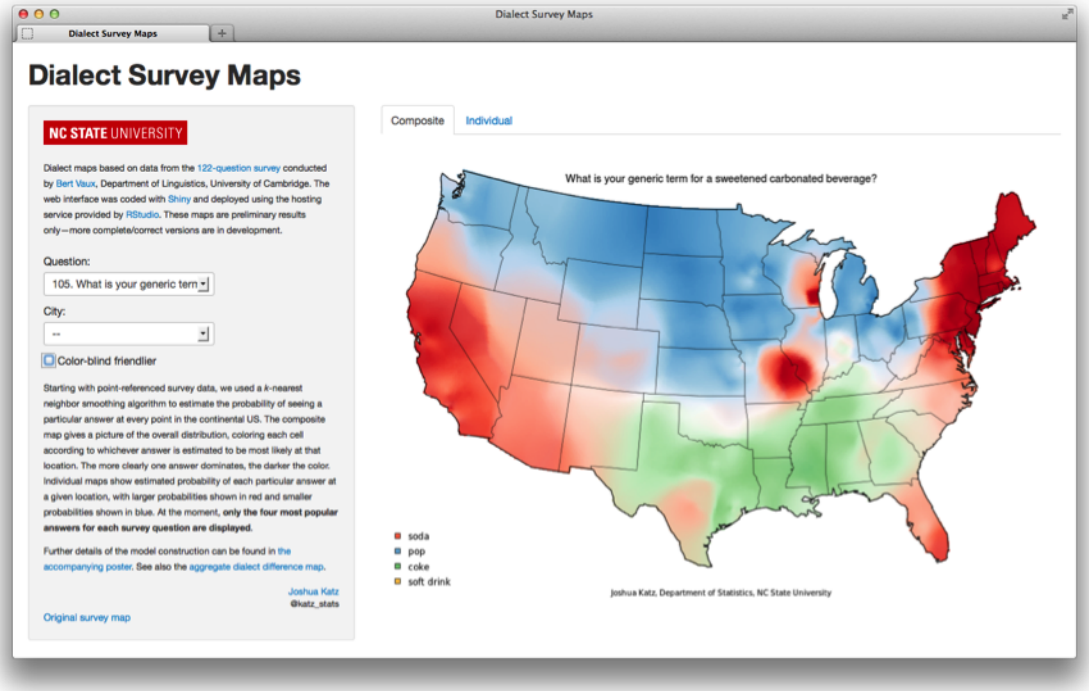

Figure 1: `http://spark.rstudio.com/jkatz/SurveyMaps/`

Amazon EC2, Big Data and High-Performance Computing

Overview

2013 has been an exciting year for the field of Statistics and Big Data, with the release of the new R version 3.0.0. We discuss a few topics in this area, providing toy examples and supporting code for configuring and using Amazon's EC2 Computing Cloud. There are other ways to get the job done, of course. But we found it helpful to build the infrastructure on Amazon from scratch, and hope others might find it useful, too.

Introduction

The term "recent advances" should be placed in context. Most of the fundamental computer science research beneath the technologies discussed here took place long ago. Still, innovation and software development of specific interest to statisticians and data scientists is one of the most important and impactful areas of R&D today. Let's say it together: "Yes, we are sexy!"

This note offers a high-level introduction to some of the recent changes of the R software environment (R Core Team, 2013b) as of versions $\geq$ 3.0.0 relating to high-performance computing. Specifically, updated indexing of vectors addresses a substantial size limitation on native R objects under versions $\leq$ 2.15.3. Native R objects are still limited to available memory (RAM), however, and many Big Data problems demand memory exceeding RAM on even the best-equipped modern hardware. To help address this

problem, we very briefly discuss package **bigmemory** (Kane and Emerson, 2013). Finally, we present the elegant framework for parallel computing using package **foreach** (Weston and Revolution Analytics, 2012).

Toy code examples are provided and were run on Amazon's Elastic Compute Cloud (EC2) running Ubuntu Linux. This isn't necessary, of course, so why do it? Because EC2 is relatively easy to use and scalable. Within a matter of minutes, anyone can request and create a cluster of instances that communicate with each other with low latency. A basic "how-to" is provided as supplementary material available online.

R Version 3.0.0 and Big Data

The ability to handle vectors of length greater than 2^{31} elements is arguably the most significant improvement provided by R versions $\geq$ 3.0.0. R versions $\leq$ 2.15.3 were unable to create such long vectors, as shown in the first code example:

```
> x <- raw(2^31 - 1)

> length(x)
[1] 2147483647
> object.size(x)       # Just over 2 GB
2147483688 bytes

> x <- raw(2^31)
Error in raw(2^31) : vector size specified is too large
```

In contrast, given sufficient RAM, R versions $\geq$ 3.0.0 have no trouble with this same example:

```
> x <- raw(2^31)

> length(x)
[1] 2147483648
```

The introduction of long vectors clearly has made R a more appealing choice for tackling Big Data problems, but there are still important qualifications. Although we technically have the ability to create large objects in R, we are practically limited by available RAM (about 23 GB on our EC2 instance):

```
> x <- raw(2^36)       # This would be about 64 GB
Error in raw(2^36) : vector size specified is too large
```

Unfortunately, there are still pitfalls even if an R object fits in available RAM. As discussed in Emerson and Kane (2012) and elsewhere, most non-trivial operations require the creation of multiple copies of objects, causing memory overhead (even if transient). When a single copy occupies a substantial proportion of RAM, the memory overhead rapidly becomes prohibitive. We provide a simple example, given in R Core Team (2013a), of copying overhead associated from the simple act of increasing the length of a long vector by one element. Here, although the object `x` occupies about 2 GB (2050.2 MB) of RAM, adding one element to the vector triggers the creation of a temporary copy; the peak memory usage during this operation has doubled to over 4 GB. An aside: `gc()` is R's "garbage collector" and can help study and manage some aspects of memory consumption.

```
> x <- raw(2^31 - 1)

> gc(reset=TRUE)[2,6]
2050.2
> x[2^31] <- as.raw(1)
> gc()[2,6]
4099.4
```

In contrast, modifying any one of the existing 2^{31} entries of this new vector `x` does not create extra copies of the vector:

```
> gc(reset=TRUE)[2,6]
2050.2
> x[2^31 - 1] <- as.raw(2)
> gc()[2,6]
2052.6
```

The memory overhead associated with almost any non-trivial operation is virtually unavoidable with native R objects. R Core recommends working with data sets that occupy at most 10-20% of available RAM in order to avoid such difficulties. Use of databases or other alternatives can help insulate the user from such problems, at least for the purpose of data management and basic manipulations. For an expanded discussion of these options, see Kane, Emerson, and Weston (2013) or an excellent page on CRAN: http://cran.r-project.org/web/views/HighPerformanceComputing.html.

In situations where the dataset is too large or the computations are too intensive, we may still need other options to be able to work efficiently in R. We will now discuss some of these options – **bigmemory**, **foreach**, and Amazon's EC2 – and demonstrate their roles in working with Big Data.

Big Data via bigmemory

The **bigmemory** family of packages enables the creation of matrices that exceed available RAM and, in fact, can be as large as the available hard-drive space. Here, we demonstrate the creation of a 100 GB matrix of single-byte "char" integers. All examples (except for one specifically noted) in this section run immediately without any lags. We use very basic toy examples here.

```
> library(bigmemory)
> N <- 10^9              # A billion rows
> K <- 100               # A hundred columns
> x <- big.matrix(N, K, type="char",
+                 backingfile="big.bin",
+                 descriptorfile="big.desc")
```

The created object `x` belongs to the `big.matrix` class. It has an associated memory-mapped "backing" file, `big.bin`, which resides on the hard drive and persists even after the R session is closed. The object can then be loaded back into R upon relaunch using the `attach.big.matrix` function. In the terminal, we can see the filebacking for the created object as well as the associated "descriptor" file, `big.desc`:

```
ubuntu@ip-10-170-20-92:/mnt/test$ ls -als
total 12
4 drwxr-xr-x 2 ubuntu ubuntu         4096 May 27 15:42 .
4 drwxr-xr-x 4 root   root           4096 May 27 15:32 ..
0 -rw-rw-r-- 1 ubuntu ubuntu 100000000000 May 27 15:42 big.bin
4 -rw-rw-r-- 1 ubuntu ubuntu          461 May 27 15:42 big.desc
```

We can now run some basic substitution and extraction operations on `x` as we would with a regular R object. The operating system manages available RAM and actively-used portions of the memory-mapped file with amazing speed and efficiency.

```
> dim(x)
[1] 1e+09 1e+02

> options(bigmemory.typecast.warning=FALSE)   # Avoids a warning message
```

```
> x[1:2, 1:2] <- 1:4       # A trivial assignment with regular matrix syntax
> x[, ncol(x)] <- 1        # Another assignment (takes 2 seconds, more work!)
> x[c(1:2, nrow(x)), c(1:2, ncol(x))]       # A 3x3 R matrix is returned
     [,1] [,2] [,3]
[1,]    1    3    1
[2,]    2    4    1
[3,]    0    0    1
```

In general, R functions that operate on matrices will not work on a `big.matrix`, but it is easy to extract subsets into a native R object in RAM as done in the example above. Packages **biganalytics**, **bigtabulate**, **bigalgebra**, and **synchronicity** offer more advanced functionality (references omitted for brevity, but available online), including k-means clustering, linear and generalized linear models, and more. A further advantage of the `big.matrix` data structure is its support for shared memory, which naturally lends itself to parallel computing (Kane, Emerson, and Weston, 2013). The following section discusses the **foreach** package which can be used in conjunction with **bigmemory** for parallel computing.

Parallel Programming via foreach

If multiple processor cores are at your disposal, then a large computational task might be completed more efficiently by making use of parallel computing. The **foreach** package extends the capabilities of standard looping constructs by delegating subtasks to multiple cores, if available, and collating results as work is returned. The framework provides the flexibility of using a variety of parallel transport mechanisms (multicore, snow, MPI, ...) without requiring code modification. Details are provided in Kane, Emerson, and Weston (2013).

We demonstrate the use of **foreach** on an EC2 cluster with a small example. We use the **doSNOW** library (Revolution Analytics, 2012) to manage the cluster from R. Our goal is to compute the column sums of a 4 by 250,000 matrix filled with integers ranging from 1 to 1,000,000.

```
>  x <- matrix(1:1e6, nrow=4)
```

Our first attempt uses one 4-core EC2 Quadruple Cluster Compute instance, taking 821 seconds:

```
> library(doSNOW)
> machines <- rep("localhost", each = 4)
> cl <- makeCluster(machines, type = "SOCK")
> registerDoSNOW(cl)
> system.time({
+   y <- foreach(j=1:ncol(x), .combine=c) %dopar% { return(sum(x[,j])) }
+ })

   user  system elapsed
249.395  18.929 820.651

> stopCluster(cl)
```

Note that this may be slower than running the same code using 4 cores on your own machine, because Amazon's compute nodes are far from being state-of-the-art. However, Amazon's strength lies in its scalability. With minimal effort, we can request a second Quadruple Cluster Compute instance, designate this second instance as a slave, and run the same code to take advantage of the 4 cores available in the master instance as well as the 4 cores available on the slave instance (identified in `/vol/nodelist` in our example):

```
ubuntu@ip-10-170-20-92:/mnt/test$ cat /vol/nodelist
ec2-174-129-178-209.compute-1.amazonaws.com
```

Back in R:

```
> machines <- readLines("/vol/nodelist")  # Get our slave node information
> machines <- rep(c("localhost", machines), each = 4)
> cl <- makeCluster(machines, type = "SOCK")
> registerDoSNOW(cl)
> system.time({
+   y <- foreach(j=1:ncol(x), .combine=c) %dopar% { return(sum(x[,j])) }
+ })

   user  system elapsed
513.124  21.849 668.685
```

Although we have doubled our computing resources, we have not halved the runtime (reduced only from 821 seconds to 669 seconds). The amount of speed-up is limited by the overhead of communication between machines over a network.

These two approaches shown above are far from optimal – parallel programming isn't trivial and doesn't always provide speed gains! R's native `apply` function (which uses a sequential loop on one core) easily beats both of the above solutions, requiring only 1.3 seconds.

```
> system.time({
+   y <- apply(x, 2, sum)
+ })

   user  system elapsed
  1.308   0.000   1.304
```

In the parallel `foreach` loop above, the entire computation is divided into 250,000 trivial subtasks which are then assigned to the available processor cores. The resulting communication overhead causes the computation to be extremely slow – far slower than the sequential solution.

Fortunately, we can further reduce the runtime by making use of the **itertools** package (Weston and Wickham, 2010), which offers a smarter way to delegate subtasks to individual cores to minimize the communication overhead. Combining **itertools** with **foreach** enables us to take full advantage of the available computing resources. Again working with all 8 cores, we can now beat the sequential `apply` solution, cutting the runtime down to about one second. Instead of creating 250,000 subtasks, **itertools** recognizes that there are 8 cores and so creates 8 subtasks for efficient execution and reduced communication overhead:

```
> system.time({
+   y <- foreach(j=isplitIndices(ncol(x), chunks=length(machines)),
+                .combine=c) %dopar% { return(apply(x[,j], 2, sum)) }
+ })

   user  system elapsed
  0.116   0.072   1.023
```

Amazon's EC2 is highly scalable in that we can request as many instances as needed, paying by the machine-hour. In this way, an iterative task that would require days to finish via a sequential loop can be rushed to completion by increasing the number of instances in the cluster.

Finally, we note that it is easy to create shared file systems on EC2 instances. Thus, if a task requires operations on a large data set (or produces results that need to be combined into a large matrix), the memory-mapped files of **bigmemory** can be stored on the shared disk space and accessed simultaneously by each of the worker processes without incurring the cost of copies. We can then make use of **foreach** and the shared-memory capabilities inherent in **bigmemory** to work with the data from any instance in the cluster. A more in-depth discussion of combining **foreach** with **bigmemory** is given in Kane, Emerson, and Weston (2013).

Jay Emerson and Xiaofei Wang
Yale University

Supplementary Material

We provide a friendly "how-to" on setting up Amazon EC2 instances and computing clusters. All examples shown in this paper were run on Amazon EC2 instances following this procedure. For more information on this and materials covered in this paper, please visit `http://www.stat.yale.edu/~jay/EC2`. We'll try to keep it updated. It currently installs both **shiny** and **FastRWeb**; these packages provide interactive web applications and CGI scripting with R (references omitted here). Topics for another day!

References

1. Emerson, J. W. and Kane, M. J. (2012). Don't drown in the data, *Significance*, **9**, 38–39.

2. Kane, M. J. and Emerson, J. W. (2013). **bigmemory**: *Manage Massive Matrices with Shared Memory and Memory-Mapped Files*, R package version 4.4.3,
Available from: http://CRAN.R-project.org/package=bigmemory

3. Kane, M. J., Emerson, J. W., and Weston S. B. (2013, to appear). Scalable Strategies for Computing with Massive Data, *Journal of Statistical Software*.

4. R Core Team (2013). Changes in R 3.0.0, *R News*,
Available from: http://cran.r-project.org/src/base/NEWS

5. R Core Team (2013). **R**: *A Language and Environment for Statistical Computing*,
Available from: http://www.R-project.org/

6. Revolution Analytics (2012). **doSNOW**: *Foreach parallel adaptor for the snow package*, R package version 1.0.6, Available from: http://CRAN.R-project.org/package=doSNOW

7. Weston, S. B. and Wickham, H. (2010). **itertools**: *Iterator Tools*, R package version 0.1-1, Available from: http://CRAN.R-project.org/package=itertools

8. Weston, S. B. and Revolution Analytics (2012). **foreach**: *Foreach Looping Construct for R*, R package version 1.4.0, Available from: http://CRAN.R-project.org/package=foreach

Section Officers

Statistical Computing Section Officers 2013

Montserrat Fuentes, Chair
fuentes@stat.ncsu.edu
(919) 515-1921

Karen Kafadar, Past Chair
kkafadar@indiana.edu
(812) 855-7828

Michael Minnotte, Chair-Elect
michael.minnotte@und.edu
(701) 777-4600

Jane L. Harvill, Secretary / Treasurer
Jane_Harvill@baylor.edu
(254) 710-1517

John Castelloe, COS Representative 2011-2013
John.Castelloe@sas.com
(919) 531-5728

Erik Iverson, COS Representative 2012-2014
erikriverson@gmail.com
(715) 252-2103

John W. Emerson, Program Chair
john.emerson@yale.edu
(203) 432-0638

Nicholas Lewin-Koh, Program Chair-Elect
lewin-koh.nicholas@gene.com
(650) 467-7955

Hadley Wickham, Publications Officer
h.wickham@gmail.com
(515) 450-8171

Rick Peterson, ASA Staff Liaison
rick@amstat.org
(703) 684-1221

Statistical Graphics Section Officers 2013

Webster West, Chair
websterwest@ncsu.edu
(803) 351-5087

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

Hadley Wickham, Chair-Elect
h.wickham@gmail.com
(515) 450-8171

Kenneth Shirley, Secretary / Treasurer
kshirley@research.att.com
(267) 243-6351

Michael Lawrence, COS Rep 2011-2013
michafla@gene.com
(515) 708-3239

Mario Morales, COS Rep 2013-2015
emetricz@hotmail.com
(917) 420-3433

Kary Myers, Program Chair
karymyers@gmail.com
(505) 606-1455

Michael Kane, Program Chair-Elect
michael.kane@yale.edu
(203) 737-4768

Rebecca Nugent, Publications Officer
rnugent@stat.cmu.edu
(412) 268-7830

Rick Peterson, ASA Staff Liaison (see left)

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:

John W. Emerson
Interim Newsletter Editor 2013
Sections on Statistical Computing and Graphics
`john.emerson@yale.edu`

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`