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

Winter 97 Vol.8 No.2/3

Statistical COMPUTING & GRAPHICS

A WORD FROM OUR CHAIRS

Statistical Computing

Daryl Pregibon is the 1997 Chair of the Statistical Computing Section. In his column he addresses the issue of making long term section initiatives a success.

The year has passed very quickly! Several months ago I wrote my inaugural address and already I am part of the transition team welcoming the new members of the Executive Committee. Thank goodness for term limits or the Section might end up with stodgy and uninspired leadership!

I thought that I would focus this address by reflecting on the way the Section is run in order to help put into perspective the extent to which new initiatives can be

CONTINUED ON PAGE 2

Statistical Graphics

Sally Morton is the 1997 Chair of the Statistical Graphics Section. In her column she outlines the section's plans for 1998.

As my year as Chair of the Statistical Graphics Section ends, I'd like to thank all the officers of the Section for their hard work. In particular, our 1997 ASA Joint Statistical Meetings Program Chair Dianne Cook produced an excellent program at the Anaheim meetings. I hope you also enjoyed the annual joint Graphics and Computing Mixer. The door prizes, ably gathered by our Treasurer Bob Newcomb with assistance from the Computing Section, were a high point as usual. Try

CONTINUED ON PAGE 3

SPECIAL FEATURE ARTICLE

Spatio-temporal Rainfall Processes: Stochastic Models and Data Analysis

By Richard Chandler, Valerie Isham and Paul Northrop

Introduction

Our aim, in the work reported here, has been to develop spatio-temporal models of rainfall fields for use by hydrologists when designing storm-water storage and drainage systems. Their need is for models that can be input, either directly or via the generation of synthetic (simulated) data, into mathematical models of soil run-off and drainage for a particular river or urban catchment. For many catchments, however, empirical rainfall data lacks either the spatial or temporal resolution required for this purpose. Therefore we choose to fit models where empirical data *are* available and adjust their parameters appropriately for the catchments of interest.

We emphasise that our models have *not* been developed with meteorological weather forecasting in mind and are unlikely to be suitable for this purpose. Rather, our intention is to reproduce the local structure of the wide variety of rainfall fields likely to be encountered at a particular catchment.

Our rainfall models have been constructed in continuous space and time, and have been fitted to fine-resolution radar data. Each radar picture is divided into an array of 2×2 km^2 pixels and represents a circular area with radius 76 km. Rainfall intensities are recorded for each pixel within the region. The radar beam takes one minute to make a complete revolution through 2π radians, and makes these revolutions at five-minute intervals. Thus, although each pixel value is obtained essentially instantaneously, the complete array is built up

CONTINUED ON PAGE 4

EDITORIAL

Double Issue

You received this issue of the newsletter after an unusually long hiatus. There are two main explanations (perhaps even justifications!) for the delay. First and foremost, for a variety of reasons, we were not as diligent and timely as we should have been. For this we apologize to you. Second, the recent flow of contributions has not been as steady and plentiful as we would have hoped.

Fortunately, in the end, we assembled successfully five articles of the highest quality. They constitute the core of this exciting double issue. However, the slow pace at which new submissions have been coming in and the dwindling number of regular contributors should make us pause to reflect for a moment.

Novel communication media, and especially the web, have modified the role and purpose of our newsletter. Access to electronic information has become so fast, affordable, and widespread among our readers that the usefulness of our news clippings and section notices has diminished considerably. By the time the newsletter comes out, most of you have already had an opportunity to learn about conference announcements and related news items from alternative sources.

So, the newsletter should have a different focus: the timely reporting of the latest trends and developments in the field of statistical computing and graphics. For this to happen, it is crucial that you keep sending in your contributions. After all, you are the key players in defining the paths along which our professional activities are evolving.

Of course, the newsletter is no substitute for peer reviewed journals. Instead, it should be regarded as the outlet of choice for reaching promptly a large, interested audience. Appropriate contributions would contain, for example, descriptions of the salient features of your most recent research projects, discussions of your experiences as developers or users of new software, and reviews of current advances in specific areas. The five main articles in this issue are excellent examples of the type of well-crafted, easy to follow, scientifically rigorous pieces that we would like to receive.

In the special feature article beginning on the cover page, Richard Chandler, Valerie Isham and Paul Northrop tackle the fascinating problem of modeling random processes that evolve both over time and space. The application they consider is to rainfall analysis and, more specifically, to the development of models that might be used in the design of storm-water storage and drainage systems. The interesting article by Nandini Raghavan and Prem K. Goel, beginning on page 10, also focuses on spatial modeling. Ultimately, their work on the morphology of the microstructure of heterogeneous materials will contribute to improve component design and manufacturing. The third article with a spatial content appears on page 31. In it, Dan Carr, Ralph Kahn, Kevin Sahr, and Tony Olsen describe a recently proposed standard for gridding information on the surface of the earth.

Beginning on page 17, Roger Koenker provides a historical comparison between ℓ_1 and ℓ_2 methods of combining observations, and describes recent advances in the development of effective computational techniques for minimizing the sum of absolute residuals. If you are considering setting up a new web site, or if you intend to spruce up your old one, make sure you turn to page 24 and read Michael Levi's advice on what the guiding principles of web site design should be.

For the two of us, this issue marks the end of an editorial collaboration that began two years ago and that has been very exciting and rewarding. Please join us in welcoming Anthony Unwin, the incoming graphics editor!

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

Mario Peruggia
Editor, Statistical Graphics Section
University of Virginia
mperuggia@virginia.edu

FROM OUR CHAIRS (Cont.)...

Statistical Computing

CONTINUED FROM PAGE 1

undertaken, and their likelihood of success. The harsh reality is that long term initiatives are problematic while short term ones have a high likelihood of success. In either case it is imperative that concerned members get involved in Section activities by either contributing their ideas or their time to maintain the vitality of the Section.

First of all, consider the positions on the Executive Committee that are viewed as most important, namely the elected positions of Section Chair and Program Chair. These positions have one year terms. Contrast

this to the elected positions of Secretary-Treasurer (2 year), Publications Officer (3 years), and Section Representatives (3 years). The implication is that the most visible positions on the Executive Committee have the shortest institutional memory! It is true that these two officials serve in an "elect-" position prior to their official capacity, but this theoretical "learning experience" doesn't work well in practice. The reason is that the Executive Committee meets only twice a year (at the JSM and Interface) and given the pressures of travel budgets and availability, it is not unusual to miss one of these meetings. The net result is that initiatives started by one of these Chairs rarely has the opportunity to come to fruition since the degree of involvement necessary to effect change is typically greater than the term of the Office (1 yr).

The Section recognizes this problem and to help ameliorate it, the Executive Committee has introduced a number of appointed positions to carry out long term section activities. Four such positions are the Newsletter Editor (Mark Hansen), the Electronic Communications Liaison (Tom Devlin), Continuing Education Officers (Ranjan Matra and John Miller), and the Awards Officer (Lionel Galway). These individuals have renewable three year terms and give tirelessly of their time.

So while it is sad to bid you farewell, it is very reassuring to know that the incoming officers, Karen Kafadar (Section Chair), Russ Wolfinger (Program Chair), and Merlise Clyde (Secretary-Treasurer) are exceptional, and together with continuing elected and appointed officers, they will deliver strong leadership to you, the Membership.

Daryl Pregibon
Statistics Research AT&T Labs
daryl@research.att.com

Statistical Graphics

CONTINUED FROM PAGE 1

your luck, and hear about all the latest section news, at this year's festivities in Dallas. Our 1998 Program Chair Ed Wegman has an interesting selection of talks planned for us, please be sure to check the schedule for the Graphics sessions.

In 1998 we welcome our incoming Newsletter Editor Antony Unwin, and thank Mario Peruggia for his dedication the past two years. We also congratulate our 1999 Chair-Elect Dianne Cook, 1999 Program Chair-Elect Deborah Swayne, and our new Council of Sections Representative David Scott. As always, Section members are encouraged to discuss any issues or concerns regarding the section or ASA matters with any section officer. Contact information is provided on the inside back page of this issue.

Section dues have been allocated to a number of exciting activities this past year. We have dedicated funding to our section's 1998 JSM Program, specifically to provide special audio-visual equipment for innovative graphics talks. We will be sponsoring the inaugural student graphics paper competition at the 1998 JSMs this August, paralleling that supported by the Computing Section. The winners will receive funding to come to Dallas and present their papers. We allocated funding both to the Undergraduate Data Analysis Contest, and to the K–12 Poster Competition. To find out about the former contest, visit http://wind.winona.msus.edu/~udac. The latter contest is one our section has co-sponsored with the ASA Center for Statistical Education since 1990. Through your section dues, we have been able to help promote and support the use of statistical graphics across a wide array of audiences. You can learn more about our funding allocations and other section news by reading the Minutes from our Executive Committee meetings posted on our website (http://orion.oac.uci.edu/~rnewcomb/statistics/graphics/graphics.html).

I am pleased to hand over the section to 1998 Chair Mike Meyer and look forward to the coming year under his leadership. I hope to see all of you at conferences and symposia throughout the year,

Sally Morton
RAND
Sally_Morton@rand.org

CONTINUED FROM PAGE 1

over one minute. However, we have assumed that the effect of this is negligible and treat each array as an instantaneous 'snapshot' of rainfall intensities. On the other hand, a single pixel value represents the average rainfall intensity over a region of 4 km^2, so that although the models are built in continuous space their properties have to be determined for spatially-aggregated data. This contrasts with the situation that applies for raingauge data, which are generally temporally-aggregated but correspond to point locations in space.

Model development has inevitably been an iterative procedure, incorporating data analysis to suggest appropriate model assumptions, model building and fitting, with assessment of model adequacy leading to refinement of the model structure, and so on. We have aimed for a parsimonious stochastic model with a modest number of parameters to represent the rainfall process. These parameters relate to underlying physical phenomena like rain cells (the smallest precipitation elements visible on radar, of the order of 50 km^2 in our data). In general, it is observed that rainfall fields exhibit effects on a hierarchy of spatial scales: rain cells cluster together to form (what we shall call) *storms*, while the storms themselves cluster together into larger-scale *rain events*. For simplicity, in the data analysis reported here, we have fitted our models with a single level of clustering (cells within storms) to data from the central portion (in both space and time) of single rain events, assuming stationarity (again, in both space and time).

Preliminary data analysis

The primary database underlying this work consists of three years' data from a weather radar station at Wardon Hill in the South West of England (in addition, contemporary data are available from a network of 49 raingauges in the catchment of the River Brue; however, these data have not been used directly in the work reported here). With an image being produced every 5 minutes when the station is working, one of the first issues to be tackled was how to deal with the sheer volume of data available. Data visualisation is an important part of the work, and software has been written to display animated sequences of radar images, with the option to print individual images or sequences of images which are of particular interest. Some specimen output is shown as Figure 1, which indicates four basic patterns into which UK rainfall may broadly be categorised.

In addition to this, programs have been written to calculate summary statistics for each radar image and produce time series plots of these statistics, on a daily/monthly basis. For example, Figure 2 shows some statistics for typical days in winter and summer months, illustrating the contrast between the generally widespread low-intensity rainfall of winter months and the much more spatial-concentrated high-intensity summer rain.

In this way, periods which are of interest for analysis and model fitting can be identified, and data quality problems can be spotted relatively easily. In choosing data sequences for the model fitting work reported here, the main considerations were: that the quality of the data should be reasonable; that the rainfall field should appear reasonably homogeneous in space and time throughout the duration of the sequence; and that the sequence should be long enough (at least 3 or 4 hours) to allow model parameters to be estimated reliably. Throughout, we focus upon the inner circle shown in Figure 1, where the data are generally more reliable than at greater distances from the radar.

In addition to these routine analyses, which are primarily a means of data reduction, it is necessary to quantify the structure of rainfall fields in space and time in order to inform any subsequent model development. One area of analysis has been the investigation of *Taylor's hypothesis* (Taylor, 1938), which relates spatial and temporal autocorrelation structure via the average velocity of a rainfall field. The velocity of a rainfall field can be efficiently estimated by tracking the centroid of the spatial autocorrelation surface at different time lags — analyses of Wardon Hill radar data in this way indicate that the hypothesis appears to hold for time lags of up to 30–40 minutes, in broad agreement with a previous study by Zawadski (1973).

The spatial variability of the total accumulated rainfall intensity (or *depth*) deposited by a storm over time is of particular relevance for hydrological applications. The storm is studied over a period of time through a circular 'data window', moving with the storm, and spatial statistics are computed over this window. An interesting feature, reported by Wheater *et al.* (1996), is that the spatial variance of the depth stops increasing after 3–4 hours in many of the storms analysed, although the mean continues to increase. One possible conclusion here is the presence of a mechanism whereby an area of high rainfall intensity within a storm inhibits further high intensity rainfall within that area later on. This conclusion is speculative, however, and the phenomenon bears further investigation.

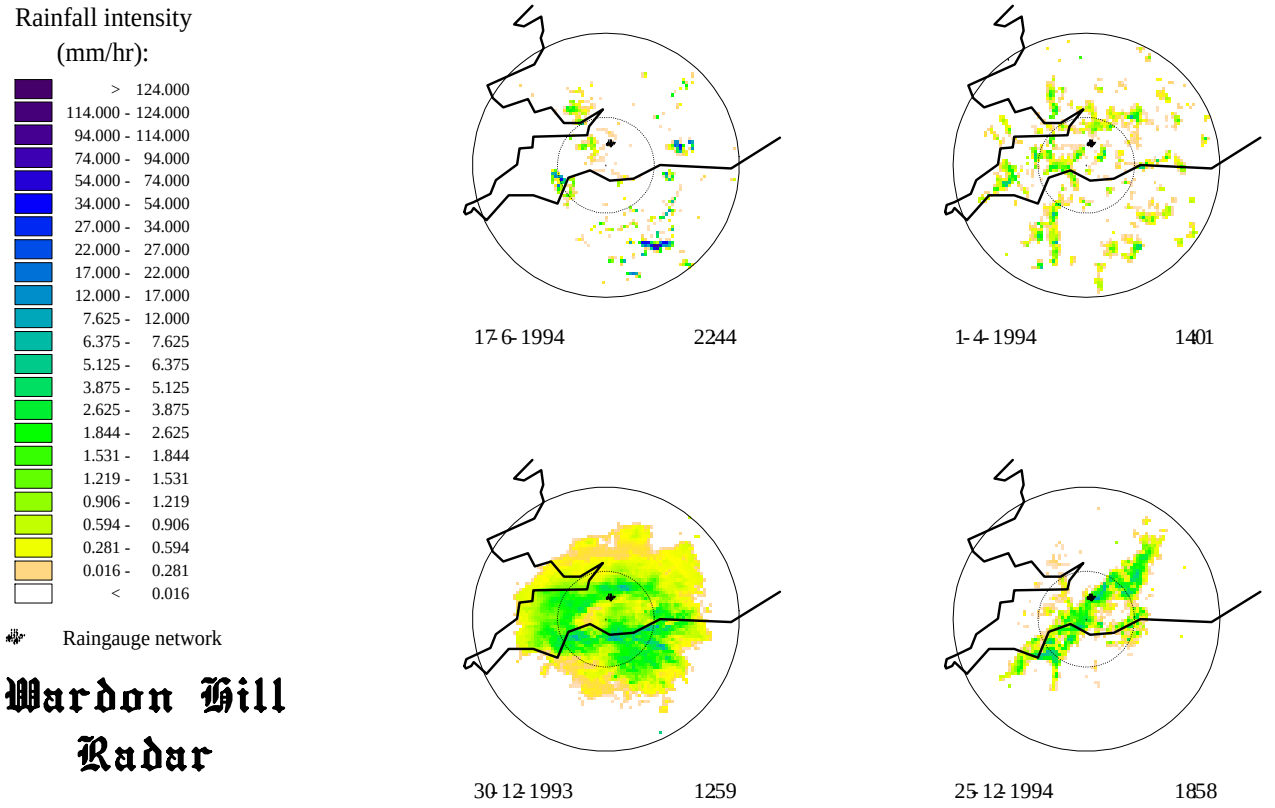


Figure 1. Different rainfall types observed with the Wardon Hill radar station. Clockwise from top left: scattered showers, widespread showers, rain band, stratiform rain.

Model specification

Our models for spatio-temporal rainfall fields are based on stochastic point processes, an approach which goes back to the pioneering work of Le Cam (1961). There was an upsurge of interest in such models in the 1980s, exemplified by Waymire *et al.* (1984), and our own work in this area began with Cox & Isham (1988). Models for the *temporal* rainfall process at a fixed spatial location, in which rectangular rain cells occur at times generated by a Poisson cluster process, have been used successfully Rodriguez-Iturbe *et al.* (1987 and 1988), and we seek to preserve the structure of these models in the marginal properties of our spatio-temporal models. For further details of the models described below see Northrop (1996).

The essential features of the models, in which storms consist of clusters of rain cells, are as follows:

- Storm locations are determined by a homogeneous Poisson process in (two dimensional) space and time. Each storm has a random velocity and an independent, exponentially distributed random duration.
- Within a storm, cell origins are clustered so that in time they follow the storm origin in a Bartlett-Lewis-type cluster (in a finite renewal process), while in space the clustering has a Neyman-Scott-type structure (independent, identically distributed displacements from the *moving* spatial storm origin) with Gaussian displacements, giving rise to the Gaussian displacements spatio-temporal model (GDSTM).
- Given the storm variables, characteristics of cells within a storm are independent and identically distributed. Spatially, each cell has either a circular or an elliptical cross-section. In the latter case, the ellipse is geometrically similar (*i.e.* has the same eccentricity and orientation) to the Gaussian covariance structure of the spatial displacement distribution. Each cell has a random spatial scale, a random temporal duration and a random intensity which is a constant over the spatial and temporal extent of the cell. Each cell in the storm moves with the velocity of the storm.

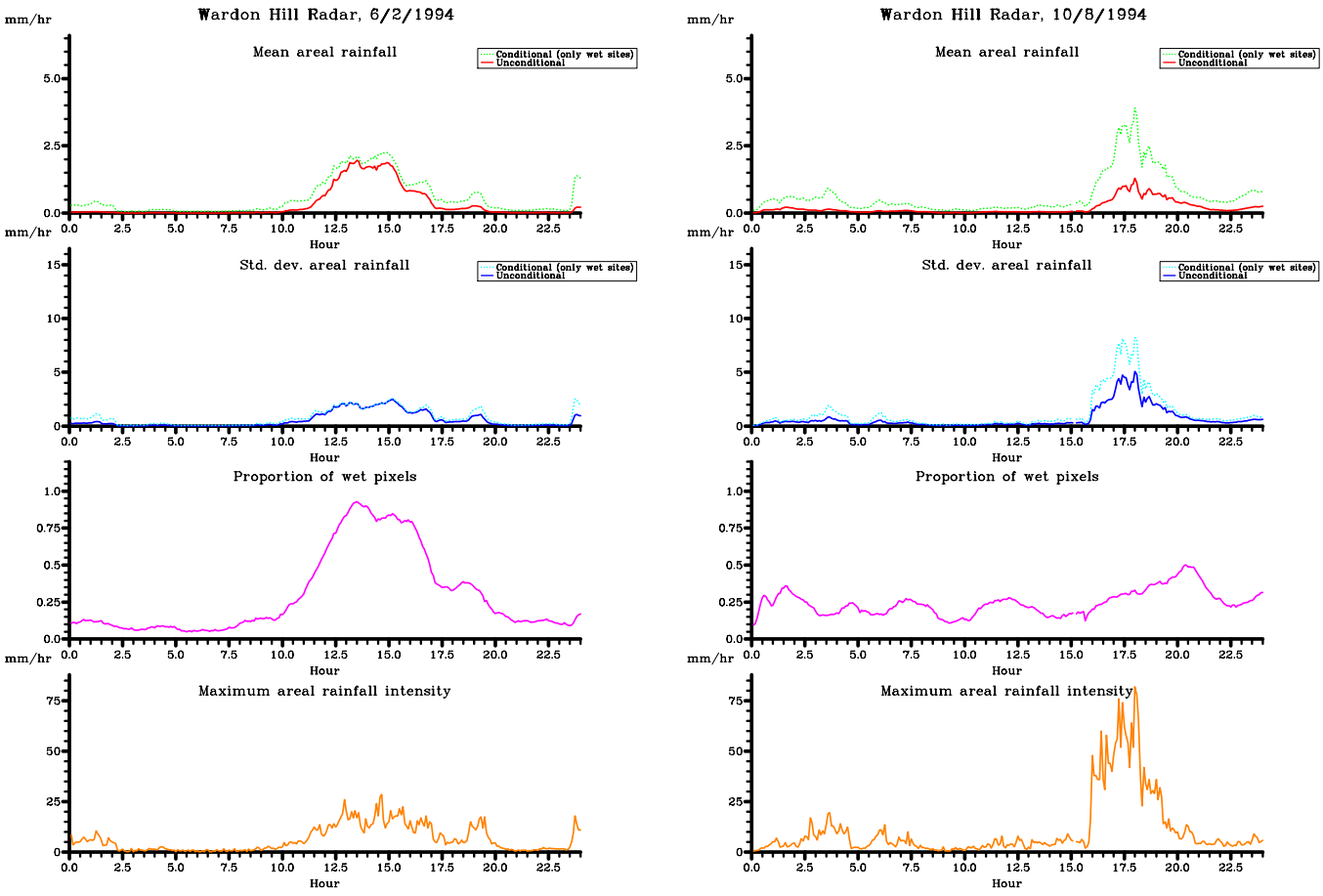


Figure 2. Time series plots of statistics for typical days in February and August 1994.

Finally, the rainfall intensity at a space-time point $(\boldsymbol{x}, t)$ is the total intensity $Y_{\boldsymbol{x}}(t)$ contributed by all cells (from all storms) that overlap $(\boldsymbol{x}, t)$ and $Y_{\boldsymbol{x}}^{(h)}(t)$ is the rainfall intensity averaged over a $h \times h$ km^2 pixel.

Explicit properties of this model are needed for use both in fitting and in assessing the adequacy of its fit. In particular, one can obtain algebraic expressions for properties of the marginal distribution of rainfall at a single site and the joint distribution of rainfall intensities at pairs of sites. For example, it is important to be able to find the fitted values of the cross-covariance, $\mathrm{cov}(Y_{\boldsymbol{x}}(t), Y_{\boldsymbol{x}+\boldsymbol{u}}(t+\tau))$, between rainfall intensities at sites separated by a spatial vector $\boldsymbol{u}$ and time τ, and joint probabilities such as $\mathrm{P}(Y_{\boldsymbol{x}}(t) = Y_{\boldsymbol{x}+\boldsymbol{u}}(t+\tau) = 0)$. Then for comparison with empirical data, similar properties for spatially aggregated data must be determined. Other properties, useful in assessing adequacy of model fit are not easy to obtain analytically and may be found by simulation of the fitted model. We discuss this further below.

Model fitting

We have explored the use of two alternative methods of model fitting. The first of these is based on a generalised method of moments, whereby an *objective function* comprising a weighted sum of squared differences between (suitably chosen) explicit and sample properties is minimised (Hansen, 1982). The second approach uses a spectral method (Chandler, 1997) in which approximate maximum likelihood is applied to the Fourier coefficients of the data. A third possibility that we have not investigated so far is to use Markov-chain Monte-Carlo fitting methods (Smith & Robinson, 1995), assuming a joint prior distribution for the unknown parameters and using sampling methods to obtain their corresponding posterior distribution given the data. The adequacy of the fit of the models is then assessed by comparison of fitted properties (both explicit and simulated) with empirical sample properties.

Some care is needed in the implementation of both of the fitting methods considered here; the problem in both cases is essentially one of minimising a highly nonlin-

ear function of several variables. Numerical methods are necessary owing to the lack of analytical solutions, and these can be unstable if not treated with due respect. All of our fitting work is carried out in FORTRAN, using NAG library routines to perform the minimisation. Typically, there are many local minima in the objective function, and it is necessary to experiment with a range of initial values to ensure that a global optimum has been found. For both methods each iteration of the minimisation algorithm requires a complete re-evaluation of the objective function, which is relatively cheap for the method of moments but can be extremely computationally demanding for the spectral method. It is well worth seeking analytical solutions for individual parameters, conditional upon values of the other parameters, if these exist — the programming effort expended in implementing such solutions is more than repaid by an increase in stability and by a reduction in processing time. Such an approach, relating to the spectral method, is described by Chandler (1996).

With these models, there is inevitably a tendency for some parameters to be difficult to identify. This results in there being distinct local optima in parameter space that give similar sets of fitted values. For example, there will be a trade-off between models having a few large clusters of cells and those with many small clusters, and between a small number of cells with high rainfall intensities and a large number of very light cells. However, these problems are most apparent when the available data are purely spatial (a single radar picture) or purely temporal (a time series from a single raingauge) and much less serious when, as here, fully spatio-temporal data are used for fitting, providing better information on individual rain cells.

Each of the two methods of fitting that we have investigated has some advantages and disadvantages. The moments method requires the subjective choice of those empirical properties to be included in the objective function, whereas the spectral method makes an apparently objective use of data (although this is somewhat illusory since the choice is being made to concentrate on fitting the second-order properties of the process). A possibly negative aspect of this objectivity in the context of applications is that, unlike the moments method, the spectral method cannot easily be forced to fit a chosen property especially well. An advantage of the spectral method is that it copes easily with assumptions of, for example, non-exponential cell durations whereas the algebraic derivation of model properties used in fitting via the moments method generally requires the model to satisfy temporal Markov properties. In addition, the spectral method is ideally suited to handle data filtering (spatial and/or temporal aggregation) and generally needs shorter runs of data. There is however a price to pay in terms of higher computational costs.

The results of these two fitting methods are compared in Wheater *et al.* (1996) but, in summary, the fitted parameter values are broadly similar from both methods, and the values and their seasonal variation are meteorologically sensible. As might be anticipated, the spectral method gives smoother variation from month to month, however it does not fit the dry periods particularly well, and this is a worrying feature for many hydrological purposes.

We present brief results from the fitting of the GDSTM to a sequence of radar data from an event on 6th February 1994, as recorded by the Wardon Hill radar station in south-west England. A generalised method of moments procedure has been used.

Figure 3 illustrates the goodness of fit of the model in terms of the temporal and spatial autocorrelation functions and the space-time variance, var$[Y_{\boldsymbol{x}}^{(h)}(t)]$, across scales of spatial aggregation h. $\rho^{(h)}(u_1, u_2, \tau)$ denotes corr$[Y_{\boldsymbol{x}}^{(h)}(t), Y_{\boldsymbol{x}+\boldsymbol{u}}^{(h)}(t+\tau)]$. There is good agreement between the observed and fitted values of these functions, the latter plot demonstrating that the model is able to reproduce the variability of the empirical data over a range of pixel sizes.

Figure 4 gives a visual illustration of the ability of the model to reproduce the internal structure of rain events. The sequence of radar images and the realisation simulated from the fitted model exhibit broadly similar features. Exact agreement between the two sequences is not expected just as exact agreement between two realisations from the same stochastic model would not be expected. The regularity imposed by the elliptical cell structure is apparent in the simulation. This feature of the model could be improved by inserting irregularity in the form of a high freqency 'jitter' (Rodriguez-Iturbe *et al.*, 1987).

Properties relating to the wet/dry pattern of rainfall were not used to fit the model and can be used, using simulations from the model, to assess its fit. Additionally, study of the pattern of rainfall intensities over *thresholds* may highlight similarities/differences between empirical and model simulated data (see Figure 5). For example, the relationship between the areas of rainfall *islands* over such thresholds and the level of the threshold can be investigated. Additionally, varying the resolution (pixels size) of the data used in such analyses will indicate the applicability of the model over different spatial scales.

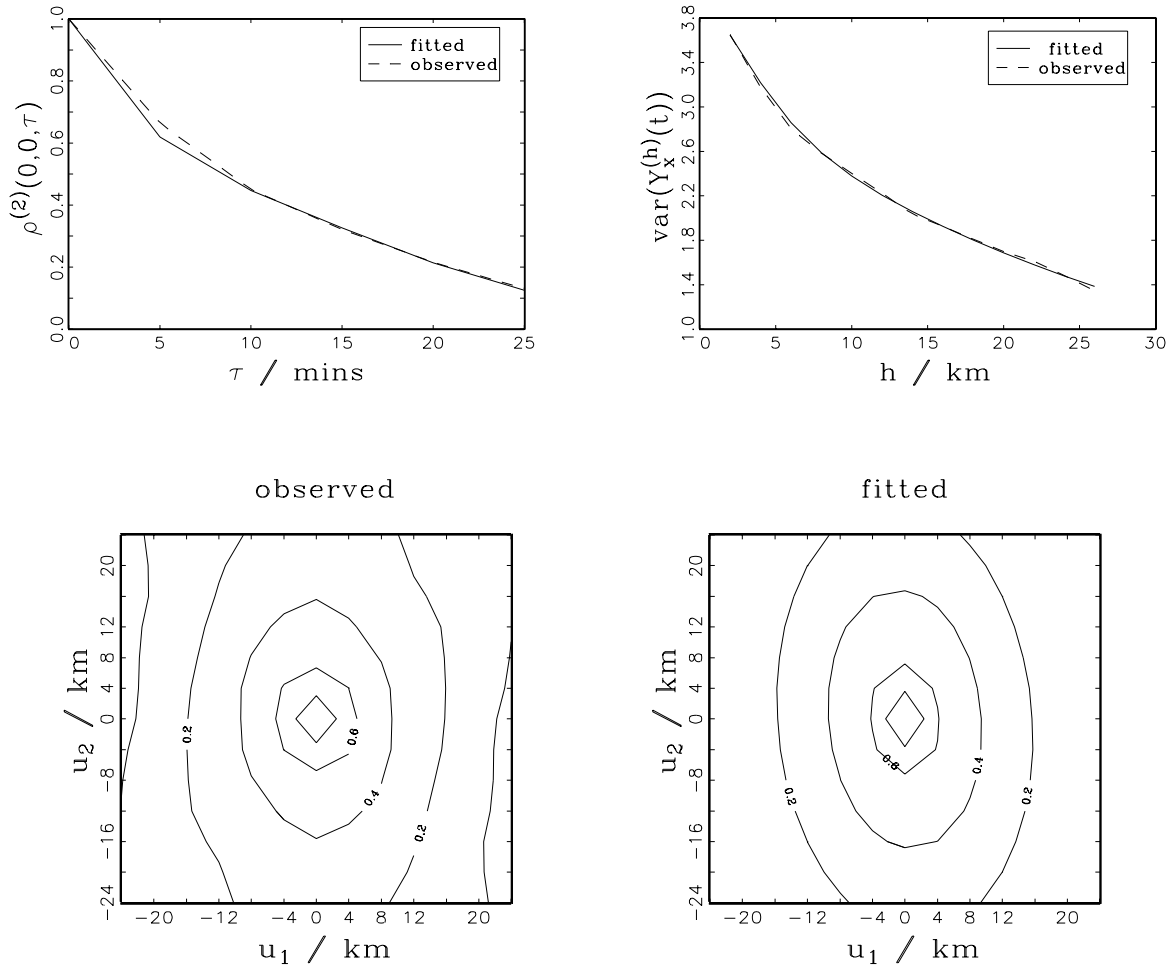


Figure 3. Assessment of the fit of the GDSTM to the storm of 6 February 1994, 1300–1400. Top left: temporal autocorrelation, top right: scaling of variance, bottom: spatial autocorrelation.

Current and future developments

The fit of our models has so far been assessed entirely in terms of their ability to reproduce statistical properties of the spatio-temporal rainfall field. However, our aim of using these models for hydrological design purposes makes it important to investigate the adequacy of their fit when used as input into soil-infiltration and run-off models. These are highly non-linear and may expose unexpected inadequacies of our fitted models by responding sensitively to apparently minor discrepancies. To investigate this, we plan to use both empirical data and simulated data from corresponding fitted models as input into the hydrological models, with the aim of comparing the statistical properties of the output (run-off).

An ultimate goal is to be able to simulate very long periods of synthetic rainfall over a catchment of hydrological interest and to answer such questions as 'how can a spatio-temporal rainfall field with given frequency for run-off (e.g. the 100 year flood) be characterised?' The current models have been developed and fitted to the interior of individual spatio-temporal rain events, and so the next stage is to embed this structure within a higher level model that will enable the continuous simulation (in time) of a long sequence of rain events over a particular spatial catchment. One way to do this is to use a non-homogeneous Poisson process to generate the times at which rain events reach the catchment and then sample the properties and parameters of each event from a space of potential rain events. The construction of this space requires the completion of a very substantial programme of exploratory data analysis and model fitting.

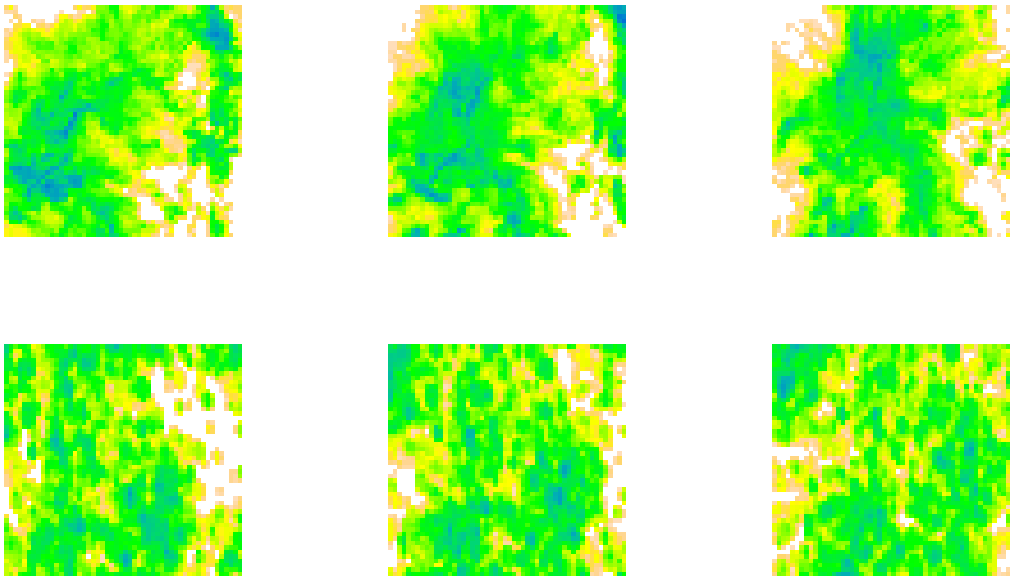


Figure 4. Visual assessment of the GDSTM fitted to the storm of 6 February 1994, 1300–1400. The temporal separation of the images is 15 minutes. Top: Observed data. Bottom: Simulation.

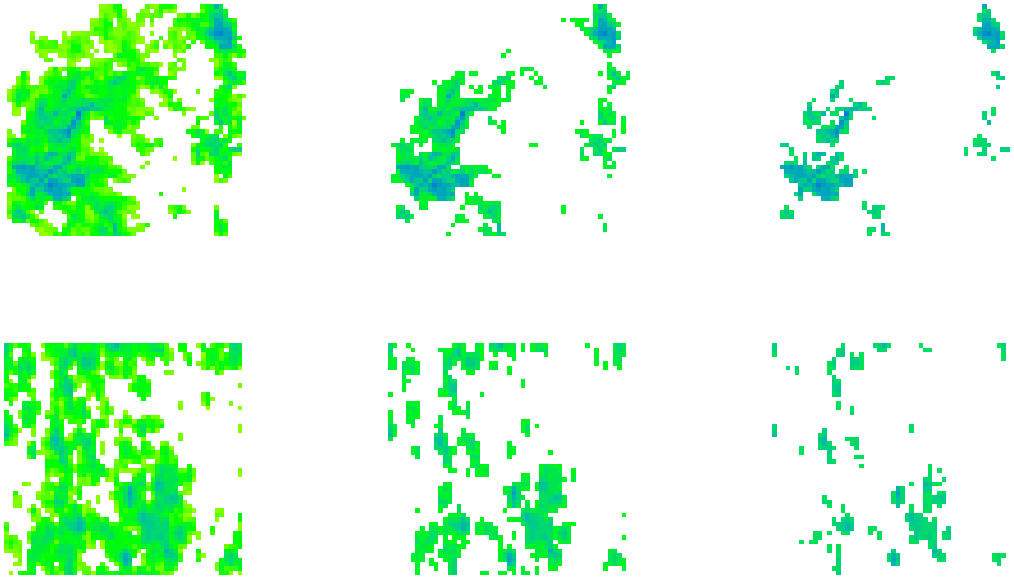


Figure 5. Thresholds of (left to right) 1, 2.5 and 4 mm/hr applied to the middle images from Figure 4. Top: radar data and bottom: model simulation.

Acknowledgements

This work was done as part of the Hydrological Radar Experiment (HYREX) funded by the UK Natural Environment Research Council (NERC) whose financial support is acknowledged with thanks. PJN also thanks the UK Engineering and Physical Sciences Research Council (EPSRC) for their support. We are especially grateful to all our collaborators on the HYREX project for their involvement, insight and helpful discussions on the work reported here: Sir David Cox, Anastasia Kakou, Christian Onof, Ignacio Rodriguez-Iturbe and Howard Wheater.

CONTINUED BOTTOM OF PAGE 10

Modeling and Characterizing Microstructures Using Spatial Point Processes

By Nandini Raghavan and Prem K. Goel

Introduction

Heterogeneous materials like reinforced composites of steel or aluminum consist of a primary material called the *matrix* to which *second-phase inclusions* are added in order to enhance mechanical properties like strength and wear resistance. Numerical and experimental studies in the field of Material Science (e.g. Brockenbrough et al., 1991; Christman et al., 1989; Moorthy and Ghosh, 1996; Spitzig et al., 1985; Rouns et al., 1992; Pyrz, 1994ab; and Ghosh, Nowak and Lee, 1997ab) suggest that the morphology of the microstructure i.e, the spatial distribution, size, shape and orientation of the *inclusions* is strongly related to the thermomechanical properties of the material. It is important to understand this relationship for effective design and manufacturing of components and for predicting component behavior and life. There is an increasing recognition in the Material Science community of the need to develop statistical models to achieve this.

The data consists of samples of micrograph windows, which are 2-D images of sections of the composite material available at desired magnifications and the corresponding material response fields. The response field for a given micrograph window is simulated using tessellation-based computer models (Moorthy and Ghosh, 1996). Studies have shown that these simulated responses closely approximate those in actual experiments. However no statistical methodology is presently used for estimating material response for the population, based on the sample of micrograph windows analyzed.

CONTINUED ON PAGE 11

SPECIAL FEATURE ARTICLE (Cont.)

References

CONTINUED FROM PAGE 9

Chandler, R. (1996), "A note on analytical solutions to the Whittle likelihood equation," Technical Report 173, Department of Statistical Science, University College London.

Chandler, R. (1997), "A spectral method for estimating parameters in rainfall models," *Bernoulli*. To appear.

Cox, D. & Isham, V. (1988), "A simple spatial-temporal model of rainfall," *Proc. R. Soc. Lond.* A415, 317–328.

Hansen, L. R. (1982), "Large sample properties of generalized method of moments estimators," *Econometrica* 50, 1029–54.

Le Cam, L. (1961), "A stochastic description of precipitation," *in* J. Neyman, ed., "Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability," 3, Berkeley, California, pp. 165–186.

Northrop, P. (1996), "Modelling and statistical analysis of spatial-temporal rainfall fields," PhD thesis, Department of Statistical Science, University College London.

Rodriguez-Iturbe, I., Cox, D. & Isham, V. (1987), "Some models for rainfall based on stochastic point processes," *Proc. R. Soc. Lond.* A410, 269–288.

Rodriguez-Iturbe, I., Cox, D. & Isham, V. (1988), "A point process model for rainfall: further developments," *Proc. R. Soc. Lond.* A417, 283–298.

Smith, R. & Robinson, P. (1995), "A Bayesian approach to the modelling of spatial-temporal precipitation data," *in* I. O'Muircheartaigh, ed., "Proceedings of the Sixth International Meeting on Statistical Climatology," Galway, June 1995.

Taylor, G. (1938), "Statistical theory of turbulence," *Proc. R. Soc. Lond.* A164, 476–490.

Waymire, E., Gupta, V. & Rodriguez-Iturbe, I. (1984), "A spectral theory of rainfall intensity at the meso-β scale," *Water Resources Research* 20, 1453–1465.

Wheater, H. S., Isham, V., Cox, D. R., Chandler, R. E., Kakou, A., Northrop, P. J., Oh, L., Onof, C. & Rodriguez-Iturbe, I. (1996), "Spatial-temporal rainfall fields: modelling and statistical aspects," Research Report 76, Department of Statistical Science, University College London.

Zawadski, I. (1973), "Statistical properties of precipitation patterns," *J. Appl. Meteor.* 12, 459–472.

Richard Chandler, Valerie Isham,
and Paul Northrop
Department of Statistical Science,
University College London

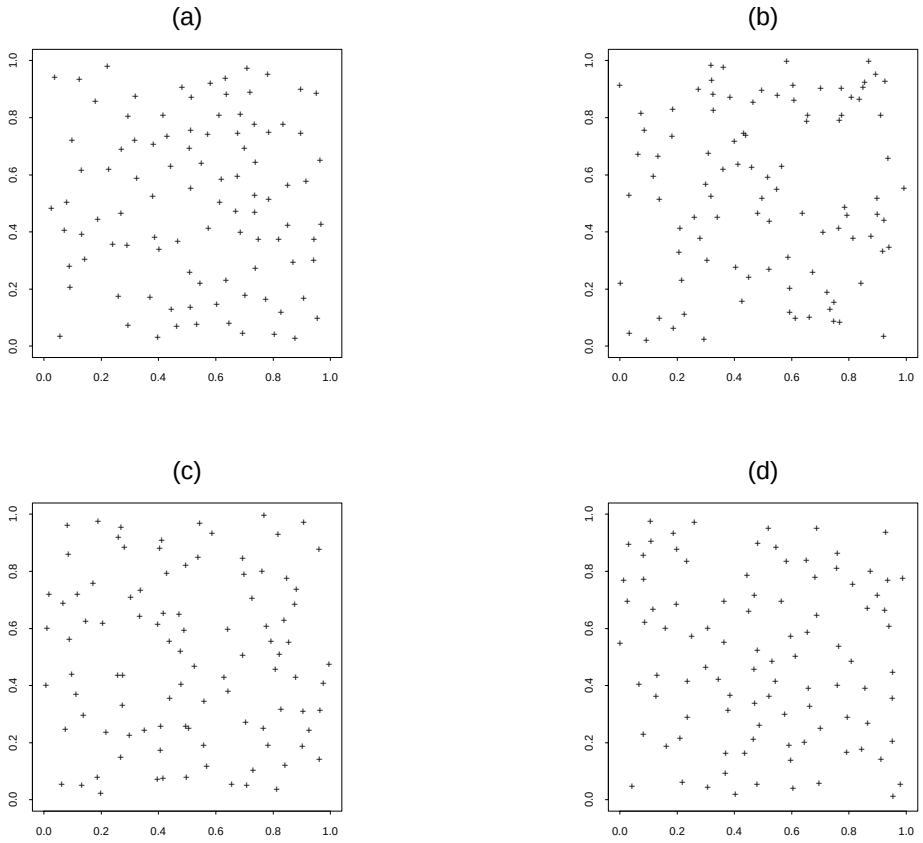


Figure 1. Data and simulated realizations from 3 spatial point processes.

In order to address this overarching goal, we will need spatial models which provide realistic depictions of actual micrographs, tools to characterize the underlying process which generates the micrograph data and subsequently, stochastic models which link the micrograph data to the material response field.

In what we describe below, we model the micrographs as realizations of spatial point processes. At low magnifications, where the *inclusions* are well-approximated as points, such models appear reasonable. This telescopic view also allows one to examine large-scale patterns such as clustering among the points, which may pinpoint irregularities in the production process. At larger magnifications, other aspects such as shape and size of the *inclusions* start coming into view and models that incorporate this additional information are also desirable.

Here, we present our attempt to address the question of characterizing or classifying a given micrograph window. To motivate the discussion, consider the four spatial point patterns in Figure 1. Can the reader identify which two belong together, if any? If the reader were further told that Figure 1 (a) represents actual data, describing the locations of the centroids of the *inclusions* for a section of an Aluminum matrix composite reinforced with Silicon particles, and that (b), (c) and (d) represent realizations from (possibly) three different spatial point processes, what would the answer be?

In Section 2, we describe some common models for spatial point processes. We also briefly describe some statistics and techniques currently used for summarizing such data and, thereby, for identifying the underlying process. One of the primary issues here is appropriate feature extraction from the binary image. In Section 3, we present our proposed methodology and use it on the point patterns in Figure 1. As we proceed, the answer to the question posed above will become increasingly apparent, as will some of the problems and issues that arise.

Models for Spatial Point Patterns

The most commonly encountered process is the homogeneous Poisson process where points are distributed uniformly (independently) on D, a bounded Borel set in $\Re^d$. Departures from this model attempt to incorporate repulsion and attraction among points.

Stochastic models, generally known as *pairwise interaction* models have been proposed for modeling inhibition among points. A pairwise interaction process is a finite point process $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ on D, a bounded Borel set in $\Re^d$ with density (see e.g., Baddeley and Moller, 1989)

$$f(X) = \alpha\beta^n \Pi_{i<j} g(\|\mathbf{x}_i - \mathbf{x}_j\|)$$

where α and β are positive constants, $g(\cdot)$ is a nonnegative real function and the cardinality of X, n, is given.

The Strauss interaction process (Strauss, 1975) arises when $g(\cdot)$ is of the form:

$$g(d) = \begin{cases} \gamma & \text{if } d \le r_s \\ 1 & \text{if } d \ge r_s \end{cases}$$

for $0 \le \gamma \le 1$. For $\gamma = 1$, this yields the Poisson process. Inhibition processes result when $\gamma < 1$ and, specifically, $\gamma = 0$ yields a $hardcore$ process, indicating that there are no points within hard-core distance r_s of any point. Strauss had proposed models for clustering with $\gamma > 1$. However Kelly and Ripley (1976) showed that the density of such processes was not integrable unless n is fixed.

Clustering can be modeled using the Neymann-Scott point process models. In particular, we used the Poisson-Poisson version of this model in our work. It consists of a parent Poisson process of intensity α defined on D and an independent daughter process of intensity β within a disk of radius r_c centered at each parent (see e.g., Ripley, 1981, pp. 165). The observed cluster process consists of the superposition of the daughter processes.

Superpositions of Point Processes

The superposition models described below allow for the generation of point patterns which exhibit regularity as well as clustering. Of particular interest in our application are superpositions of hard-core and cluster processes. This is a preliminary attempt to incorporate information about the size of an *inclusion* into the hard-core constraint and the clustering that typically results during the production process when heterogeneities are not well-dispersed in the matrix. See Raghavan, Goel and Ghosh (1997). We are currently investigating more realistic models based on marked point processes and random sets.

Let $n_s \sim$ Binomial(n, p), where $0 \le p \le 1$, denote the number of points generated from a Strauss hard-core process. Correspondingly $n_c = n - n_s$ is the number of points generated from a Poisson cluster process. Let X_s refer to the locations of the n_s points and similarly, X_c to the locations of n_c points and let f_s and f_c denote the respective pdfs of X_s and X_c.

Then the pdf of $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is:

$$\sum_{n_s=0}^{n} p^{n_s}(1-p)^{n_c} f_s(X_s) f_c(X_c) \qquad (1)$$

Other models have also been proposed for spatial point processes (see, e.g. Baddeley and Moller, 1989, for nearest-neighbor Markov point processes; Baddeley and van Leishout, 1995, for area-interaction point processes; and Granville and Smith, 1995, for more general Neyman-Scott cluster processes). Cressie (1993) also gives a general treatment of spatial point processes.

Summary Statistics

Statistics that have been used in the spatial point process literature have been based on near-neighbor distances, second-order descriptors and summaries of spatial tessellations.

Point-to-point nearest neighbor distance

Nearest neighbor distances are designed to capture small scale interactions between points. Let d_{ij} denote the Euclidean distance between two points $\mathbf{x}_i$ and $\mathbf{x}_j$ of the process and let $d_i = \min_j d_{ij}$. The empirical cdf $\hat{G}$ of these nearest-neighbor distances is given by:

$$\hat{G}(d) = n^{-1} \sum_i I(d_i < d)$$

where $I(x \in S)$ denotes the indicator function which is 1 if x belongs to S. Various statistics based on $\hat{G}$ have been proposed for testing randomness (see Upton and Fingleton, 1985). Often Monte Carlo confidence regions for the entire function are used to determine whether a given realization comes from a Poisson process.

Figure 2 (a) illustrates the plot of $\hat{G}(\cdot)$ for the point patterns in Figure 1.

The empty space function $F(\cdot)$ is correspondingly the cdf of the origin-to-point nearest neighbor distances and estimates the distribution of the distance of an arbitrary fixed point in D to the nearest point of the spatial point process. This is shown in Figure 2 (b).

The J-function, defined as

$$J(d) = \frac{(1 - G(d))}{(1 - F(d))}, \quad d > 0,\ F(d) \le 1$$

was recently proposed by van Leishout and Baddeley (1996) to quantify the strength and range of the interpoint interactions in a spatial point process. This is shown in Figure 2 (c).

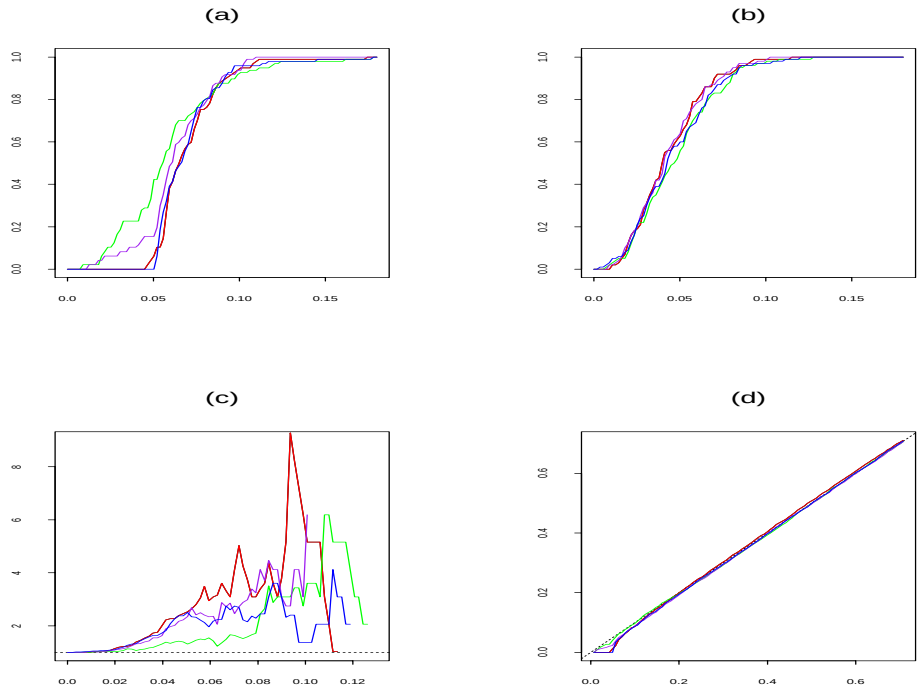


Figure 2. Plots of (a) $\hat{G}(\cdot)$, (b) $\hat{F}(\cdot)$, (c) $\hat{J}(\cdot)$ and (d) $\hat{K}(\cdot)$ for the point patterns in Figure 1 (a) red, (b) green, (c) purple, (d) blue.

Second Order Statistics

Second order properties of a point process are captured by the K-function (see Ripley, 1977) where

$$K(d) = \lambda^{-1} E[\text{number of points within a distance } d \text{ of an arbitrary point of the process}].$$

Given $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $K(d)$ is estimated by

$$\hat{K}(d) = n^{-2}|D| \sum \sum_{i \neq j} w_{i,j}^{-1} I(d_{i,j} < d).$$

where n, D and d_{ij} are as described earlier, $|D|$ is the area of D, w_{ij} is the proportion of the circle with center at i and passing through j which lies in D. This estimate corrects for edge effects. A transformation of $K(d)$ is given by $L(d) = \sqrt{K}(d)/\pi$. For a homogeneous Poisson process, $K(d) = \pi d^2$ and this gives a 45^o line for the plot of $L(d)$ vs. d. Therefore $L(d)$ is often used for detecting departures from a homogeneous Poisson process. Figure 2 (d) shows $\hat{L}(d)$ for the point patterns in Figure 1.

Spatial Tessellations

A spatial tessellation (see Okabe et al., 1992; or Upton and Fingleton, 1985, pp. 96-97) of a given realization of a point process on domain D in R^2 is a partitioning of the area of D into polygons (also called tiles or Voronoi cells), enclosing each point of the realized process in such a way that every point in the polygon is closer to that particular point of the process than to any other. For $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \Re^2$, the Voronoi cell associated with $\mathbf{x_i}$ is:

$$C(\mathbf{x_i}|\mathbf{X}) = \left[\mathbf{s} \in \Re^{\mathbf{2}} : \|\mathbf{x_i} - \mathbf{s}\| \leq \|\mathbf{x_j} - \mathbf{s}\| \ \ \forall \mathbf{j} \neq \mathbf{i}\right]$$

One could think of these polygons as being the regions of influence of their particular points. The boundaries of each tile are segments of the perpendicular bisectors of the lines joining a point to its neighbors.

Summary statistics have been based on the number of sides, the length of the perimeter and the areas of the tiles. By studying the sampling distributions of the Voronoi cell areas for the various processes, we elicited statistics that would capture differences among the processes. These included the CV, skewness and kurtosis of the cell area distribution for a given realization.

Very little is known about the theoretical properties of empirical distributions based on realizations from these models except the Poisson process, which has been studied extensively. Inference for point processes has typically concentrated on parameter estimation within specific models and significance tests for randomness. The tests are based on measures such as near-neighbor

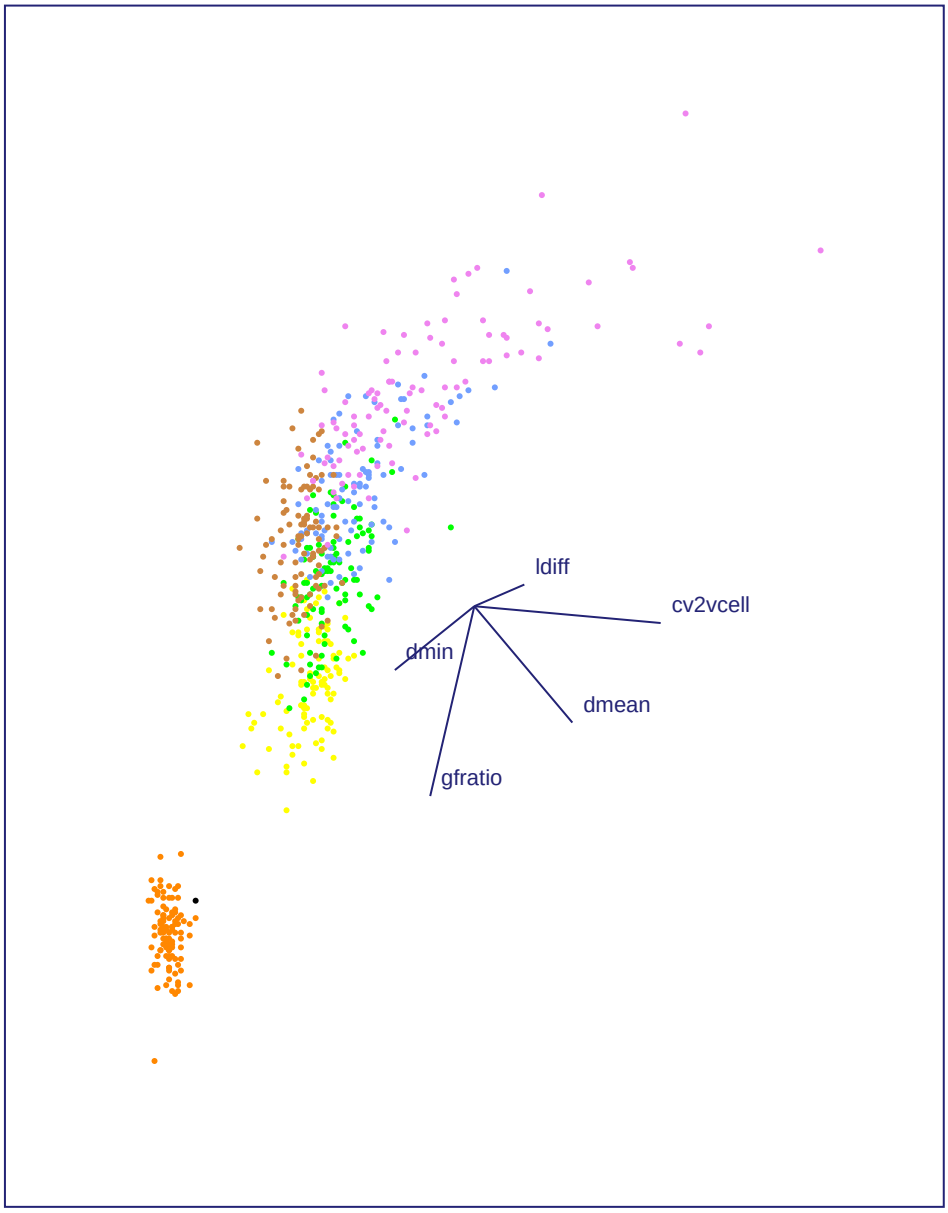


Figure 3. A 2-dimensional projection of the following 5 features (using XGobi): $\min d_i$ (dmin), $\bar{d}$ (dmean), the area between $\hat{L}(\cdot) - L_p(\cdot)$ (ldiff), the CV of the Voronoi cell areas (cv2vcell) and the area under the ratio $\hat{G}(\cdot)/\hat{F}(\cdot)$ (gfratio). This is based on 100 realizations from the following six processes: Poisson (peru), Strauss (orange), Strauss-Cluster 75:25 mixture (yellow), Strauss-Cluster 50:50 mixture (green) Strauss-Cluster 25:75 mixture (blue), Poisson cluster (orchid), Actual Data (black). Parameters: Strauss $r = 0.05$, Poisson cluster $\alpha = 10, r = 0.25$.

distances, second order statistics such as the K-function (Ripley, 1977) and the areas of cells of spatial tessellations. While they give important information about the spatial distribution of locations, it is well-understood that no one function can be regarded as characterizing a point pattern (Silverman and Baddeley, 1984). Multiple tests based on a variety of statistics against null-Poisson models do not necessarily lead to a conclusive decision. This causes difficulties when comparing two or more non-nested, in particular non-Poisson, models. This point is driven home by the differing inferences the reader would have likely drawn, regarding the underlying model for the point pattern in Figure 1 (a) based on the four different empirical functions in Figure 2.

The Classifier

We propose and construct a supervised pattern recognition scheme for classifying spatial point patterns into one of several specified classes of processes. Our idea in

attempting to build such a classifier is to exploit the joint empirical distribution of several statistics. The procedure thus provides a unified way of comparing two or more models for spatial point patterns. We demonstrate the procedure on the point patterns in Figure 1.

One of the primary issues here is feature extraction from the image. In building the classifier for the supervised pattern recognition scheme, we utilized various moment-based, quantile-based and order statistics of the nearest neighbor distances and of the spatial tessellations. We also proposed and utilized summary statistics based on the $K(\cdot)$ and $J(\cdot)$ functions. Data visualization tools such as XGobi, (Buja et al., 1996) and Monte Carlo methods were used to study the marginal and joint distributions of features under various models. Sampling properties of these various statistics were also examined. The reader is referred to (Raghavan, Goel and Ghosh, 1997) for details of these. Figure 3 shows a two-dimensional projection in the five-dimensional feature space defined by the following five features which were among those used in the classifier: $\min d_i$, $\bar{d}_i$, the area between $\hat{L}(\cdot) - L_p(\cdot)$ in the initial portion of the curves, the CV of the Voronoi cell areas and the area under the ratio $\hat{G}(\cdot)/\hat{F}(\cdot)$. The projection was obtained using XGobi. Virtually all of the summary measures we used extract information from the entire realization, not just from typical points or cells, and thus try to capture the dependence structure in the point pattern.

The classification procedure assumes that the set of processes which could potentially have generated a given pattern is known. This implies having some prior knowledge about reasonable parameter sets or ranges for these underlying processes. Here, we estimated the parameter r_s in the Strauss hard-core distance by $\min d_i$ (the minimum nearest neighbor distance). Parameter settings for the Poisson-Poisson cluster process were harder to come by and were obtained by trial and error.

For this set of parameter specifications, we considered the six classes of processes; the Poisson process, the Strauss hard-core process, the Poisson cluster process, the superposition process with $p = 0.75$, with $p = 0.50$ and with $p = 0.25$, as defined in (1). We built the classifier using 100 realizations from each of the six classes.

The *training sample* consisted of the various summary statistics computed for each of these realizations. The classifier was constructed using the `tree` function in S-plus (see e.g. Venables and Ripley, 1994, Chapter 13). The function `predict.tree` was used to classify each of the four spatial point patterns in Figure 1 into one of the six classes, based on the maximum posterior probability.

Results and Discussion

Table 1 gives the posterior probabilities of the six classes for each of the point patterns in Figure 1. The actual data (Figure 1 (a)) is classified categorically into the Strauss hard-core model. The point patterns in Figure 1 (b), (c) and (d) were generated from a Strauss-Cluster mixture with p = 0.5, 0.75 and 1 respectively. The 50:50-mixture in Figure 1 (b) is misclassified as a Poisson but the other two are correctly classified.

Class	25:75	50:50	75:25	C	P	S
Fig. 1 (a)	0	0	0	0	0	1
Fig. 1 (b)	0	0	0.333	0	0.667	0
Fig. 1 (c)	0	0.029	0.97	0	0	0
Fig. 1 (d)	0	0	0	0	0	1

Table 1. Posterior Probabilities for the four point patterns in Figure 1.

A plausible explanation for the misclassification is that these two processes are virtually indistinguishable from each other. If so, labeling a point pattern as coming from one or another process is a fairly arbitrary exercise. That this may be the reason is evidenced by the plot in Figure 3 which indicate substantial overlap among five of the six classes in the 5-dimensional feature space. In particular the 25:75 mixture, the 50:50 mixture and Poisson process appear to be hard to separate.

The complexity of the classifier, indicated to some extent by the number of terminal nodes, also suggests this. As the underlying processes become inherently less separable, the complexity of the classifier necessarily increases as more features and finer partitions are necessary to separate the classes. The number of terminal nodes here was 36, indicating a fairly complex classifier. In other experiments, we have seen as few as 3.

Computer experiments described in (Raghavan, Goel and Ghosh, 1997) illustrate the feasibility of using a supervised pattern recognition scheme to successfully classify spatial point patterns. The question of what is the most appropriate classifier to use in this context is one that warrants enquiry and one that we are currently investigating.

Acknowledgements

The authors would like to thank Somnath Ghosh, Department of Aerospace Engineering, Applied Mechanics and Aviation, The Ohio State University for providing the data.

References

Baddeley, A. and Moller, J. (1989), "Nearest Neighbor Markov Point Processes and Random Sets," *Intl. Statist. Review*, **57**, 89-121.

Baddeley, A. and van Lieshout, M. N. M (1995), *Ann. Inst. Stat. Math.*, **47**, 4, 601–619.

Baddeley, A. and Silverman, B. W. (1984), *Biometrics*, **40**, 1089–1093.

Brockenbrough, J. R., and Wienecke, H. A. (1991), "Deformation of metal-matrix composites with continuous fibers: Geometrical effects of fiber distribution and shape," *Acta Metall. et Mater.*, **39**, 735-752.

Buja, A., Cook, D. and Swayne, D. F. (1996), "Interactive High-Dimensional Data Visualization," *J. Comp. and Graph. Stat.*, **5**, 78–99.

Christman, T., Needleman, A. and Suresh, S. (1989), "An experimental and numerical study of deformation in metal-ceramic composites," *Acta Metall. et Mater.*, **37**, 3029–3050.

Cressie, N. (1993), *Statistics for Spatial Data*, Wiley, New York.

Ghosh, S., Nowak, Z. and Lee, K. (1997a), "Quantitative characterization and modeling of composite microstructures by Voronoi cells," *Acta Metall. et Mater.*, **45** 6, pp 2215-2234.

Ghosh, S., Nowak, Z. and Lee, K. (1997b), "Tessellation based computational methods in characterization and analysis of heterogeneous microstructures," *Comp. Sci. Tech.*, (in press).

Granville, V. and Smith, R. L. (1995), "Clustering and Neyman-Scott process parameter simulation via Gibbs sampling," *Technical Report, Statistical Laboratory, University of Cambridge* Cambridge, England.

Kelly, F. P. and Ripley, B. D. (1976), "On Strauss' model for clustering," *Biometrika*, 63, 357–360.

Moorthy, S. and Ghosh, S. (1996), "A model for analysis of arbitrary composite and porous microstructures with Voronoi cell finite elements," *Int. Jour. Numer. Meth. Engrg.*, **39**, 2363-2398.

Okabe, A., Boots, B. and Sugihara, K. (1992), *Spatial Tessellations*, Wiley, New York.

Pyrz, R. (1994a), "Quantitative description of the microstructure of composites. Part I: Morphology of unidirectional composite systems," *Comp. Sci. Tech.* **50**, 197-208.

Pyrz, R. (1994b), "Correlation of microstructure variability and local stress field in two-phase materials," *Mater. Sci. Engrg.* **A177**, 253-259.

Ripley, B. D. (1977), "Modelling Spatial Patterns," *J. Royal Statist. Soc. Ser. B.*, **39** 172–212.

Ripley, B. D. (1981), *Spatial Statistics*, Wiley, New York.

Rouns, T. N. , Fridy, J. M., Lippert, K. B. and Richmond, O. (1992), "Quantitative characterization and modeling of second phase populations through the use of tessellations," *Simulation and Theory of Evolving Microstructures*, ed. M.P. Intl. Conf. Aluminum Alloys, **2**, 333-340, Trondheim, Norway.

Spitzig, W. A., Kelly, J. F. and Richmond, O. (1985), "Quantitative characterization of second phase populations," *Metallography*, **18**, 235-261.

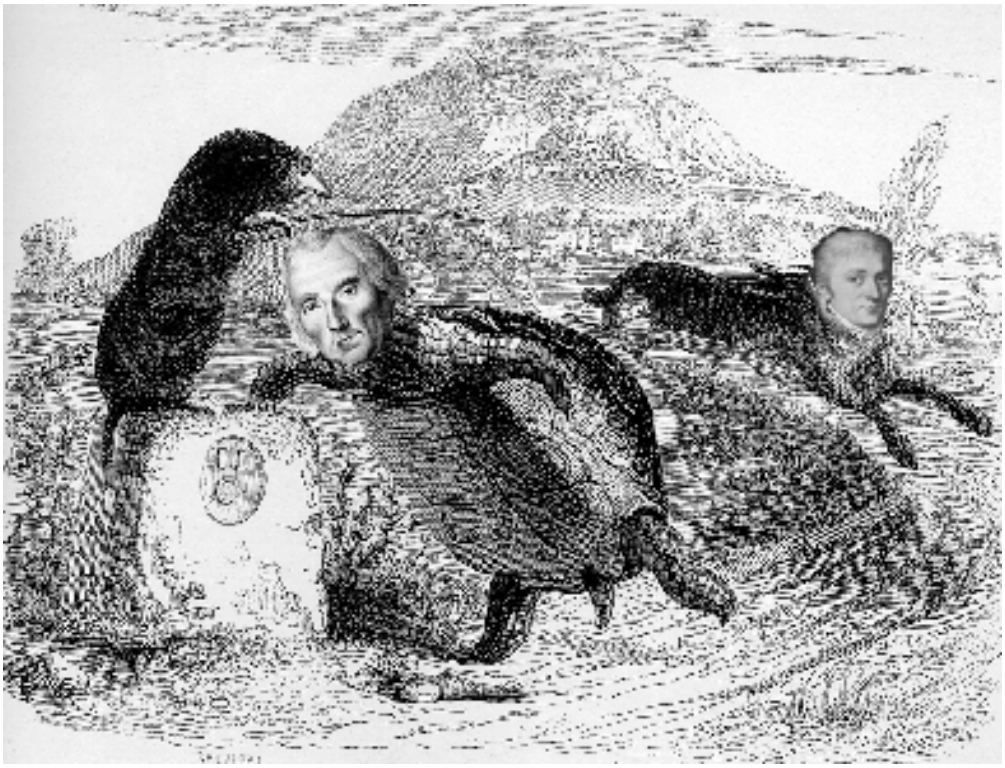
Strauss, D. J. (1975), "A model for clustering," *Biometrika*, **62**, 467–475.

Upton, G. and Fingleton, B. (1985), *Spatial Data Analysis by Example, Vol. I*, Wiley, New York.

van Lieshout, M. N. M, and Baddeley, A. (1996), *Stat. Neerlandica*, **50**, 3, 344–361.

Venables, W. N. and Ripley, B. D. (1994), *Modern Applied Statistics with S-plus*, Springer-Verlag, New York.

Nandini Raghavan
raghavan@stat.ohio-state.edu
and
Prem K. Goel
goel@stat.ohio-state.edu
Department of Statistics
The Ohio State University



COMPUTING AND STATISTICAL MODELING

ℓ_1 Tortoise Gains on ℓ_2 Hare

By Roger Koenker

> *Tartu, Estonia: June 10, 1998.*
> Archival sources announced here today the discovery of an unusual wood-etching/"photoprint" from the late 18th century depicting the conclusion of Stage 11 of Round G of the *Tour de Monde Statistique.* Previous *reportage* from this race by M. La Fontaine had suggested the possibility that the perennially over-confident Hare "Gauss" had been overtaken by the tortoise representing M. Laplace. The print, reproduced above, confirms quite conclusively, that the tortoise was leading at the stone marker ending stage G-11. However, attempts to obtain the official records of this stage of the race, presumably registered by the unidentified mole in sunglasses with the dowsing stick, have proven, thus far, unsuccessful.

An Historical Introduction

Since Gauss it has been generally accepted that ℓ_2 methods of combining observations by minimizing sums of squared residuals have significant computational advantages over earlier ℓ_1 methods based on minimizing sums of absolute residuals, advocated by Boscovich, Laplace and others. In 1887, six years after publishing his path breaking work in economics *Mathematical Psychics*, F.Y. Edgeworth began a series of papers "On a new method of reducing observations relating to several quantities." In fact the method was not *entirely* new. In the 1760's the Croatian Jesuit Roger Boscovich proposed estimating the ellipticity of the earth by solving a problem of the form:

$$\min \sum_{i=1}^{n} |y_i - \alpha - x_i'\beta|$$

subject to:

$$\sum_{i=1}^{n} (y_i - \alpha - x_i'\beta) = 0,$$

where y_i was the length of one degree of latitude measured at latitude θ_i and $x_i = \sin^2 \theta$. Somewhat later, Laplace showed that this problem could be solved by computing a weighted median, and provided an astonishingly complete theory of the limiting distribution of the resulting estimator $\hat{\beta}$, when $\alpha = 0$.

Edgeworth's new method, which he called the "plural median" was intended to revive the Boscovich/Laplace approach as a direct competitor to the least squares approach championed by Galton and others. Edgeworth (1888) proposed dropping the zero-mean constraint on residuals, arguing that it conflicted with the median intent of the absolute error approach. And appealing to the univariate results of Laplace, he conjectured that the plural median should be more accurate than the least squares estimator when the observations were more "discordant" than those from the Gausssian probability law. Finally, he proposed a rather arcane geometric algorithm for computing the plural median and remarked rather cryptically:

> ...the probable error is increased by about 20 percent when we substitute the Median for the Mean. On the other hand, the labour of extracting the former is rather less: especially, I should think in the case of many unknown variables. At the same time, that labour is more "skilled". There may be needed the attention of a mathematician; and, in the case of many unknowns, some power of hypergeometrical conception.

The "20 percent" is a bit optimistic. At the normal model the median would have confidence intervals which are about 25 percent wider than those based on the mean. But many of the details of Edgeworth's conjectures concerning the improvements achievable by the plural median over comparable least squares methods of inference in discordant situations have been filled in over the last 20 years. See, for example, Bassett and Koenker (1978), Koenker and Bassett (1982), Powell (1986), Gutenbrunner and Jurečková (1992), Gutenbrunner, Jurečková , Koenker and Portnoy (1993), and the three conference volumes of Dodge (1987,1992,1997).

Even in the elementary one-sample setting we can see the merit in Edgeworth's claim: the median is less laborious to compute by hand, but perhaps requires more skill than finding the mean. In large samples, if we adopt the naive idea that finding the median requires a complete sorting of the n sample observations it is easy to reach the conclusion that the median needs $\mathcal{O}(n^2)$ operations while the mean can obviously be computed in $\mathcal{O}(n)$ operations. More careful sorting quite easily reduces $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$, but considerable "skill" was required to design the $\mathcal{O}(n)$ algorithm of Floyd and Rivest (1975). Can something comparable be done to make ℓ_1 regression competitive with least squares in terms of computational complexity?

Edgeworth's conjectures about the computational complexity of ℓ_1 versus ℓ_2 regression methods were also partially vindicated by the development of the simplex method of linear programming by Dantzig and others in the late 1940's. Refinements of simplex designed for the ℓ_1 regression problem including the widely implemented Barrodale and Roberts (1973) algorithm proved quite competitive with least squares computational speed for problems up to a few thousand observations. In Figure 1 we illustrate this based on experience in Splus on a Sparc 20. Note that up to about 3000 observations for $p = 4$ parameters, or up to about n=1000, for $p = 8$, or $n = 300$ for $p = 16$, the Splus function `l1fit` embodying the algorithm of Barrodale and Roberts is actually faster than the QR decomposition algorithm for least squares embodied the the Splus function `lm()`. However in larger problems the simplex approach founders badly, exhibiting quadratic growth in cpu-time with n. By the time that we reach $n = 100,000$, with $p = 16$ for example, `l1fit` requires nearly an hour of Sparc 20 time while the equivalent least squares computation takes about 10 seconds.

Interior versus Exterior Methods

A recent paper, Portnoy and Koenker (1997), explores two approaches which, taken together, provide some reason for optimism. I will briefly describe both approaches, relying on the full paper to fill in the details. Consider the median regression problem,

$$\min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} |y_i - x_i' b| \tag{2}$$

which may be formulated as the linear program,

$$\min\{\ e'u + e'v \ \mid\ y = Xb + u - v, (u,v) \in \mathbb{R}^{2n}_+\}. \tag{3}$$

Note that we have simply decomposed the regression residual vector into its positive and negative parts, calling them u and v, and written the original problem as one of minimizing a linear function of the $2n$-vector (u, v) subject to n linear equality constraints and $2n$ linear inequality constraints. This "primal" linear program formulation of the ℓ_1-regression problem has an associated "dual" formulation in which we maximize with respect to a vector, $d \in \mathbb{R}^n$, which may be viewed as the vector of Lagrange multipliers associated with the equality constraints of the primal problem. This dual formulation is,

$$\max\{y'd \ \mid\ X'd = 0, \quad d \in [-1,1]^n\}, \tag{4}$$

or equivalently, setting $a = d + \frac{1}{2} e_n$,

$$\max\{y'a \ \mid\ X'a = \frac{1}{2} X' e_n, \ \ a \in [0,1]^n\}, \tag{5}$$

where e_n denotes an n-vector of ones. The simplex approach to solving this problem may be briefly described

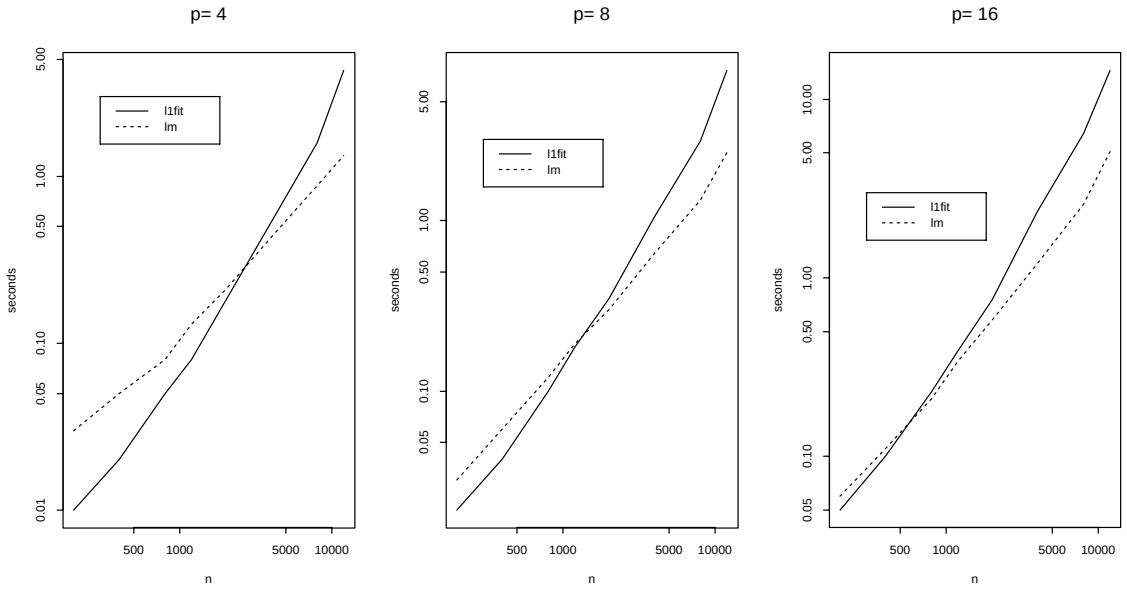


Figure 1. Timing comparison of ℓ_1 and ℓ_2 algorithms: Times are in seconds for the median of five replications for iid Gaussian data. The parametric dimension of the models is $p+1$ with p indicated above each plot, p columns are generated randomly and an intercept parameter is appended to the resulting design. Timings were made at 8 design points in n: 200, 400, 800, 1200, 2000, 4000, 8000, 12000. The solid line represents the results for the simplex-based Barrodale and Roberts algorithm, `l1fit(x,y)` in Splus, and the dotted line represents least squares timings based on $\texttt{lm}(\texttt{y} \sim \texttt{x})$.

as follows. A p-element subset of $\mathcal{N} = \{1, 2, ..., n\}$ will be denoted by h, and $X(h)$, $y(h)$ will denote the submatrix and subvector of X, y with the corresponding rows and elements identified by h. Recognizing that solutions to (2) may be characterized as planes which pass through precisely $p = \dim(b)$ observations, or as convex combinations of such "basic" solutions, we can begin with any such solution, which we may write in the primal formulation as,

$$b(h) = X(h)^{-1} y(h). \tag{6}$$

We may regard any such "basic" primal solution as an extreme point of the polyhedral, convex constraint set. In the dual formulation since the index set h identifies the active constraints of the primal problem, i.e. those observations for which both u_i and v_i are zero, $a(h)$ lies in the interior of the p dimension unit cube, and the complement of h corresponds to coordinates of a which lie on the boundary: if $u_i > 0$ then $a_i = 1$, while if $v_i > 0$ then $a_i = 0$. A natural algorithmic strategy is then to move to the adjacent vertex of the constraint set in the direction of steepest descent. This transition involves two stages: the first chooses a descent direction by considering the removal of each of the current basic observations and computing the gradient in the resulting direction, then having selected the direction of steepest descent and thus an observation to be removed from the currently active "basic" set, we must find the maximal step length in the chosen direction by searching over the remaining $n - p$ available observations for a new element to introduce into the "basic" set. Each of these transitions involves an elementary "simplex pivot" matrix operation to update the current basis. The iteration continues in this manner until no direction is found at which point the current $b(h)$ can be declared optimal.

To illustrate the shortcomings of the simplex method, or indeed of any strategy for solving linear programs which relies on an iterative path along the *exterior* of the constraint set, consider the problem depicted in Figure 2. We have a polygon whose vertices lie on the unit circle and our objective is to find a point in the polygon that maximizes the sum of its coordinates, that is, the point furthest north-east in the figure.

Since any point in the polygon can be represented as a convex weighting of the extreme points, the problem may be formulated as

$$\max\{e'u | X'd = u, \;\; e'd = 1, \;\; d \in \mathbb{R}^n_+\}, \tag{7}$$

where e denotes a (conformable) vector of ones, X is an $n \times 2$ matrix with rows representing the n vertices of the polygon and d is the vector of convex weights to be determined. Eliminating u we may rewrite (4.1) somewhat more simply as

$$\max\{s'd | e'd = 1, \;\; d \in \mathbb{R}^n_+\}, \tag{8}$$

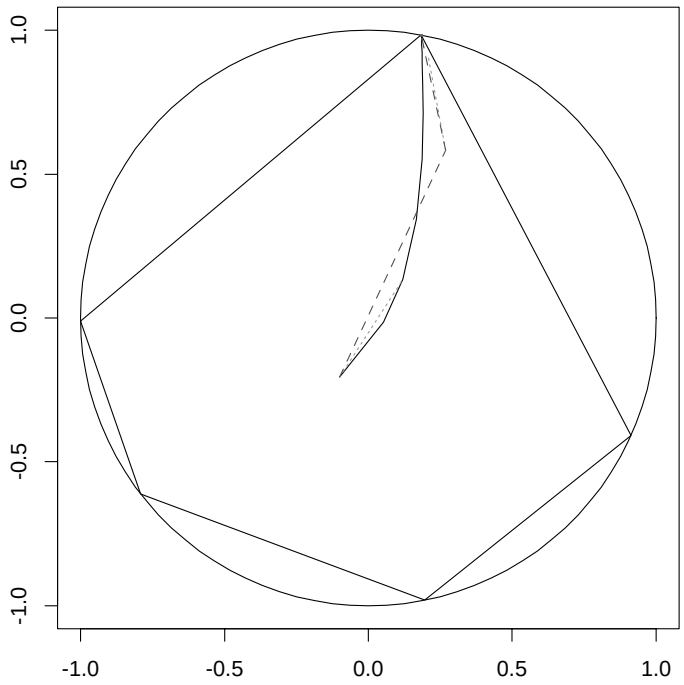


Figure 2. A Simple Example of Interior Point Methods for Linear Programming: The figure illustrates a random pentagon of which we would like to find the most northeast vertex. The central path beginning with an equal weighting of the 5 extreme points of the polygon is shown as the solid curved line. The dotted line emanating from the center is the first affine scaling step. The dashed line is the modified Newton direction computed according to the proposal of Mehrotra (1992). Subsequent iterations are unfortunately obscured by the scale of the figure.

where $s = Xe$. This is an extremely simple linear program which serves as a convenient geometric object for studying various approaches to solving such problems. Simplex is particularly simple in this context, because the constraint set *is* literally a simplex. If we begin at a random vertex, and move around the polygon until optimality is achieved, we pass through $\mathcal{O}(n)$ vertices in the process. This does not appear to be too onerous in Figure 2 where $n = 5$, but in ℓ_1 regression where the number of vertices is $\binom{n}{p}$ it can prove to be quite burdensome. Of course, a random initial vertex is rather naive, and one could do much better with an intelligent "Phase 1" approach that finds a *good* initial vertex. In effect, we can think of the "interior point" approach we will now describe as a class of methods to accomplish this, rendering unnecessary further travel around the outside of the polygon.

Although prior work in the Soviet literature offered theoretical support for the idea that linear programs could be solved in polynomial time, thus avoiding certain pathological behavior of simplex, the paper of Karmarker (1984) constituted a watershed in the numerical analysis of linear programming. It offered not only a cogent argument for the polynomiality of interior point methods of solving LP's, but also provided for the first time direct evidence that interior point methods were demonstrably faster than simplex in specific, large, practical problems.

It is an interesting irony, illustrating the spasmodic progress of science, that the most fruitful practical formulation of the interior point revolution of Karmarker (1984) can be traced back to a series of Oslo working papers by the economist Ragnar Frisch in the early 1950's. The basic idea of Frisch (1956) was to replace the linear inequality constraints of the LP, by what he called a log barrier, or potential, function. Thus, in our example, we may reformulate the problem as,

$$\max\{s'd + \mu \sum_{i=1}^{n} \log d_i | e'd = 1\} \qquad (9)$$

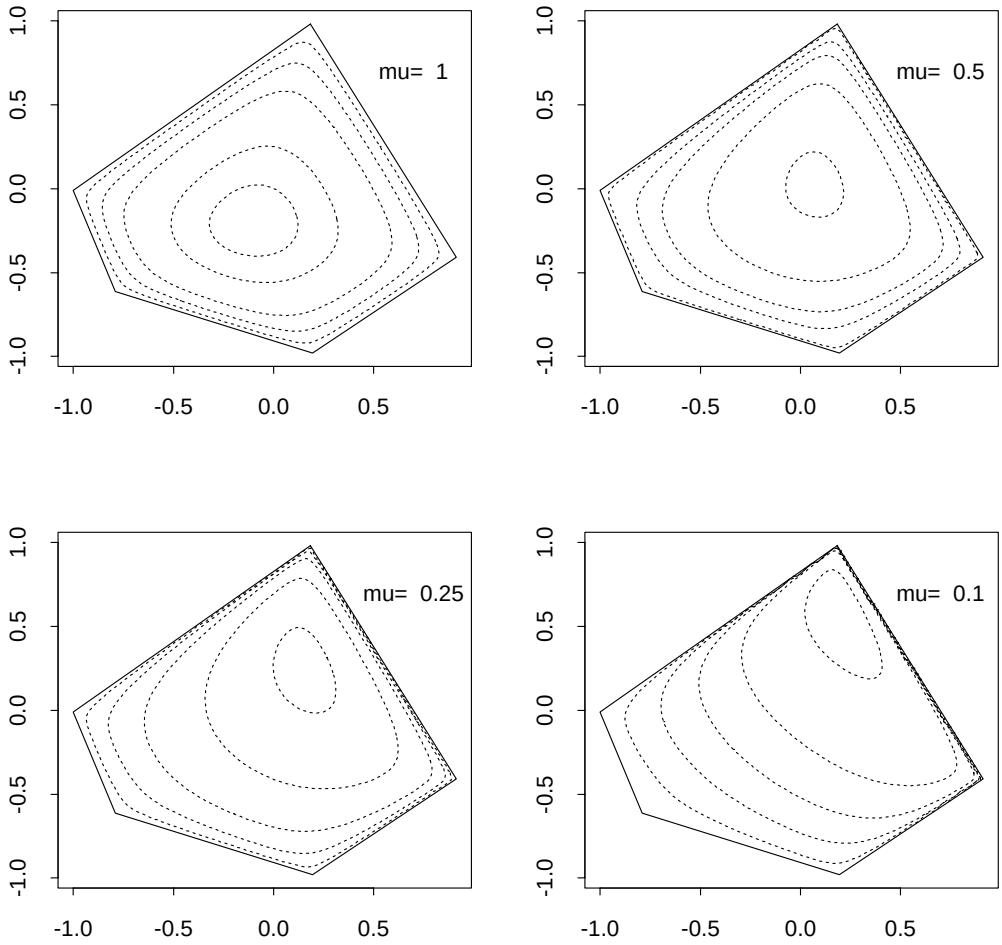


Figure 3. Contours of the Log Barrier Objective Function for the Simple Polygonal Linear Program: The figure illustrates four different contour plots of the log barrier objective function (2.8) corresponding to four different choices of μ. In the first panel, $\mu = 1$ and the contours are centered in the polygon. As μ is reduced the penalized objective function is less influenced by the penalty term and more strongly influenced by the linear component of the original LP formulation of the problem. Thus, for $\mu = .1$ we find that the unconstrained maximum of the log barrier function occurs quite close to the optimal vertex of the original LP. The locus of solutions to the log barrier problems for various μ's is called the central path, and is illustrated in Figure 2 by the solid curved line.

where now the barrier term $\mu \sum \log d_i$ serves as a penalty which keeps us away from the boundary of the positive orthant. By judicious choice of a sequence $\mu \to 0$ we hope to converge to a solution of the original problem.

The salient virtue of the log barrier formulation is that, unlike the original formulation, it yields a differentiable objective function which is consequently attackable by Newton's method and inherits the quadratic convergence of Newton's method.

In Figure 3 we try to illustrate the log barrier approach by plotting 4 versions of the contours corresponding to the penalized objective function for four distinct values of the penalty parameter μ. In the first panel, with $\mu = 1$ we are strongly repelled from the boundary of the constraint set and the unconstrained maximum of the barrier function occurs near the center of the polygon. In the next panel, with μ reduced to $\frac{1}{2}$ the barrier penalty exerts a somewhat weaker effect and the contours indicate that the unconstrained maximum occurs somewhat closer to the upper vertex of the polygon. This effect is further accentuated in the $\mu = \frac{1}{4}$ figure, and in the last figure with $\mu = \frac{1}{10}$ we find that the maximum occurs quite close to the vertex. The path connecting the maximum of the family of fixed-μ problems is generally called the central path. As emphasized by Gonzaga (1992) and others, this central path is a crucial construct

for the interior point approach. Competing algorithms may be usefully evaluated on the basis of how well they are able to follow this path. Clearly, there is some tradeoff between staying close to the path and moving along the path, thus trying to reduce μ, iteration by iteration. Improving upon existing techniques for balancing these objectives is the subject of a vast outpouring of current research. Excellent introductions to the subject are provided in the survey paper of Margaret Wright (1992) and the recent monograph of Stephen Wright (1996).

In Figure 2 we have also illustrated the central path for our simple polygonal example. The solid curved line indicates the central path, the dotted line indicates the much superior first step taken by an affine scaling version of an interior point algorithm, while the dashed line indicates the first step of a primal dual version of the interior point method. The primal dual form of the algorithm requires somewhat more work per iteration, but takes fewer iterations and is somewhat more robust with respect to degeneracy and non-uniqueness of the solution. Formal computational complexity results indicate that for large problems primal dual implementations of interior point methods for solving ℓ_1 problems require $\mathcal{O}(np^3 \log^2 n)$ operations which is considerably better than the quadratic in n behavior of simplex, but still inferior to the $\mathcal{O}(np^2)$ behavior of least squares.

Preprocessing

But further gains are possible from careful preprocessing of ℓ_1 type problems. Preprocessing rests on an extremely simple idea: if, by preliminary estimation, or some other form of statistical necromancy, we could determine the signs of a significant group of observations, we could then combine observations with positive residuals into a single "globbed" observation, and similarly glob together the negative observations, so that the original problem,

$$\min \sum_{i=1}^{n} |y_i - x_i' b| \tag{10}$$

would be equivalent to,

$$\min \sum_{i \in N \setminus J_L \cup J_H} |y_i - x_i' b| + |y_L - x_L' b| + |y_H - x_H' b| \tag{11}$$

where $N = \{1, 2, ..., n\}$, $x_K = \sum_{i \in J_K} x_i$ for $K \in \{L, H\}$ and y_L and y_H can be chosen arbitrarily small and large respectively, to ensure that the corresponding residuals on the globbed observations remain negative and positive. In this process we have reduced the problem of n original observations to $n - \sharp\{J_L, J_H\} + 2$ observations so if the cardinality of the J-sets is large we have gained substantially. Under plausible sampling assumptions we can, based on a preliminary subsample of m observations, make a prediction region for $\{x_i\beta : i = 1, 2, ..., n\}$ of width $\mathcal{O}(p/\sqrt{m})$, so assigning observations above this region to J_H and observations below this region to J_L, we would have $M = \mathcal{O}_p(np/\sqrt{m})$ observations falling inside the region. This is illustrated in Figure 4.

Minimizing the computational effort required to compute the preliminary fit based on m observations plus the effort required for the solution of the globbed problem (3.2) with M observations, we obtain $m^* = \mathcal{O}((np)^{2/3})$, which under our conjectured performance of the underlying interior point algorithm yields a complexity for the full problem of

$$C = \mathcal{O}_p(n^{2/3} p^3 \log^2 n) + \mathcal{O}(np^2), \tag{12}$$

where the first term comes from the solution of the two median regression problems of size $\mathcal{O}(n^{2/3})$ and the second term arises from the computation of the confidence band.

Further details are provided in Portnoy and Koenker (1997) and I will comment only briefly here on the important fact that any implementation of this preprocessing approach must verify that the solution to the globbed problem actually agrees with the predicted signs based on the confidence region. The simultaneous confidence region can be chosen to assure this with arbitrarily high probability, and the eventuality that we may need to repeat the cycle to remedy some inaccurately predicted signs introduces another multiplicative factor which does not affect the orders in probability in the complexity computation.

All of the foregoing discussion may be extended immediately from the median case to a general quantile regression setting as introduced in Koenker and Bassett (1978). Quantile regression, by offering a means of estimating a complete family of conditional quantile functions in the familiar setting of linear regression, significantly extends the domain of applicability of ℓ_1 methods. In effect, it permits the researcher to examine the effect of specified covariates in particular ranges of the conditional distribution of the response variable, and thus to distinguish differential effects in the tails, or on dispersion, from the classical regression effect characterized by a simple location shift.

The crucial consequence of the formal complexity theory and the extensive concomitant empirical testing of our implementation of the algorithm is that the computational effort required for quantile regression can be made comparable with the effort required for least

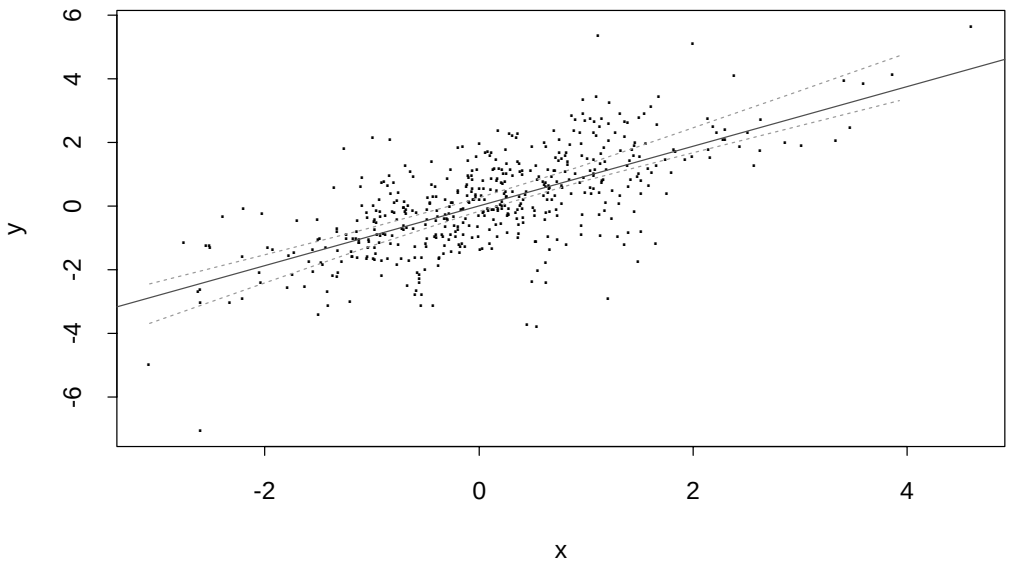


Figure 4. A Bivariate Example of Quantile Regression Preprocessing: The figure illustrates a bivariate scatter plot of 500 observations with y conditionally student t on 10 degrees of freedom. The curved dotted lines describe a confidence band for the response variable based on the median regression fit for a sub-sample of 126 observations. After globbing there are only 107 observations, including the two globbed observations. All the points outside the band are collapsed into this pair of pseudo-observations. The fit to the globbed sample is indicated by the solid line; since it falls inside the band we are assured that the globs are correct and that this solution is identical to a fit of the entire original sample.

squares over the full range of currently plausible problem dimensions. In the final empirical example of Portnoy and Koenker (1997), we compare timings for a typical large econometric application of quantile regression with $n = 113,547$ and $p = 6$. With the new algorithm, quantile regression estimates take about 10 seconds on a Sparc-Ultra, comparable to the least squares time of 8 seconds. Interior point methods applied to the full problem before preprocessing requires about a minute for these problems. Simplex solution of the same quantile regression problems requires approximately an hour on the same machine.

Future Prospects

There are many open questions posed by the rapid development of computational methods for ℓ_1 methods and quantile regression more generally. But we are, I believe, on the verge of fully vindicating Edgeworth's old claim that the "plural median" is less laborious, as well as more robust, than its least squares competitor. Laplace's Tortoise has a real chance against Hare "Gauss", and it is to be hoped that this will provide a continued impetus for the further development of these methods in statistics.

This research has been partially supported by NSF Grant SBR 93-20555.

References

Barrodale, I. and Roberts, F.D.K. (1974), "Solution of an overdetermined system of equations in the ℓ_1 norm," *Communications ACM*, **17**, 319-320.

Bassett, G. and Koenker, R. (1978), "Asymptotic Theory of Least Absolute Error Regression," *Journal of the American Statistical Association*, 73, 618-622.

Dodge, Y. (.ed) (1987), *Statistical Data Analysis Based on the L_1 Norm and Related Methods*, Amsterdam, North Holland.

Dodge, Y. (.ed) (1992), *L_1 Statistical Data Analysis and Related Methods*, Amsterdam, North Holland.

Dodge, Y. (.ed) (1997), *L_1 Statistical Procedures and Related Topics*, IMS Monograph Series, **31**, Hayward, CA.

Edgeworth, F.Y. (1888), "On a new method of reducing observations relating to several quantities," *Philosophical Magazine*, **25**, 184-191.

Floyd, R.W. and Rivest, R.L. (1975), "Expected Time Bounds for Selection," *Communications of the ACM*, **18**, 165-173.

Frisch, R. (1956), "La Résolution des problèmes de programme linéaire par la méthode du potential logarithmique," *Cahiers du Séminaire d'Econometrie*, **4**, 7-20.

Gonzaga, C.C. (1992), "Path-following methods for linear programming," *SIAM Review*, **34**, 167-224.

Gutenbrunner, C. and Jurečková , J. (1992), "Regression quantile and regression rank score process in the linear model and derived statistics," *Annals of Statistics* **20**, 305-330.

Gutenbrunner, C., Jurečková , J., Koenker, R. and Portnoy, S. (1993), "Tests of linear hypotheses based on regression rank scores," *J. of Nonparametric Statistics*, 2, 307-33.

Karmarkar, N. (1984), "A new polynomial time algorithm for linear programming," *Combinatorica*, **4**, 373-395.

Koenker, R. and Bassett, G. (1978), "Regression quantiles," *Econometrica*, 46, 33-50.

Koenker, R. and Bassett, G. (1982), "Tests of linear hypotheses and l_1 estimation," *Econometrica*, 50, 1577-1584.

Koenker, R. and Portnoy, S. (1997), "The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-error vs. Absolute-error Estimators," *Statistical Science*, 12, 279-299.

Mehrotra, S. (1992), "On the implementation of a primal-dual interior point method," *SIAM Journal of Optimization*, **2**, 575-601.

Powell, J.L. (1986), "Censored regression quantiles," *J. Econometrics*, 32, 143-155.

Wright, M.H. (1992), "Interior methods for constrained optimization," *Acta Numerica*, 341-407.

Wright, S.J. (1996), *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia.

Roger Koenker
Department of Economics
University of Illinois Urbana-Champaign
roger@ysidro.econ.uiuc.edu

WRITING FOR THE WWW

A Shaker Approach to Web Site Design

By Michael D. Levi

"That is best which works best"

"Beauty rests on utility"

"Simplicity is the embodiment of purity and unity"

-Shaker Maxims

The Shakers

Who are the Shakers, and what do they have to do with the World Wide Web?

In brief, the Shakers were a religious denomination who split off from the English Quakers and immigrated to the New World shortly before the American revolution. The Shakers, or United Society of Believers in Christ's Second Appearing, reached a peak of approximately 6000 members in 18 communities by the mid-nineteenth century, and began to decline after the Civil War. Today only a few remain.

The Shakers believed in communal property, pacifism, equality of the sexes, celibacy, open confession of sins, and consecrated labor. They lived in largely self-sufficient communities separate from the "world's people." Unlike some other religious denominations, Shakers were not averse to technology. On the contrary, Shaker communities were early adopters of machinery and techniques that would facilitate their work.

The Shakers are probably best remembered for their furniture, their architecture, and other crafts. All are characterized by an extraordinary level of skill and beauty.

It is the contention of this paper that the three key attributes of Shaker artifacts:

- Simplicity
- Elegance
- Quality

along with the overriding Shaker emphasis on utility, can serve as the guiding principles for designing and developing World Wide Web sites.

A Shaker chair is useful. It is well suited for sitting. Its clean, smooth lines have no excess adornment; the beauty of the wood is adornment enough. So should

Web sites be structured for maximum utility, allowing the value of their content to attract and hold user interest.

Measuring Web Success

It is fairly easy to determine whether a chair is 'successful' or not. From a vendor's point of view, if the chair is purchased and not returned it is a success. If the sale of a chair leads to repeat business, it is even more successful.

From a purchaser's point of view a chair is successful if it is comfortable, does not break when sat upon, has the desired physical dimensions, and fits the aesthetics of the room in which it is located. Depending on the circumstances, other criteria may also be important. Lawn chairs must withstand precipitation and temperature change. Nursery furniture should not poison infants.

Shaker chairs were (and continue to be) successful primarily because they meet the above criteria. Above all else, Shaker chairs are useful. They are designed to seat humans; all other functionality is of secondary consideration.

Determining whether a Web site is successful is less straightforward. The most common metric used is synonymous with popularity: a successful site has a lot of "hits."

This is insufficient. To start with, "hit" is a very poorly defined term. Depending on the site and the tool set used, a reported hit count may or may not include graphic elements such as pictures, icons, and buttons, as well as internal technical details of the Web software such as 'image maps.' So at the very best a hit count can only legitimately be used to measure changes in popularity over time, but not as a comparative cross-site measure.

Another flaw in using hit counts as a basic metric is its leveling effect. A hit count does not distinguish between a content-rich page, a list of internal or external links, or an 'under construction' page. A hit is a hit, regardless of substance.

Worst of all, a hit count often rewards exactly the wrong thing. A site that has been redesigned to streamline user access to the most frequently retrieved material may well show a reduction in hits precisely because its utility has been enhanced.

There are certainly quantitative metrics that can be used: number of sessions, number of users, and number of repeat users are all valuable statistics to gather and analyze. But qualitative measures are equally important. The best measure of success is how well a site facilitates users and sponsors in accomplishing their goals. In short, the value of a site can be assessed by measuring its utility and quality.

In order to determine whether a site's goals have been met, it is useful to begin by introducing a rough taxonomy of site types and Web users.

Site Types

With the explosion of the Web over the past three years, Web sites have sprung up for almost every imaginable purpose. Despite this diversity, when analyzed from the standpoint of interactivity there are only three basic types of Web site:

- Ego, or Vanity sites;
- Resource, or Information sites;
- Transaction or Two-way sites.

A personal home page is the embodiment of an ego site. Its purpose is to say "I'm here. I'm cool. Look at me!" Ego sites may serve as a communications channel for a very small audience (directions to a wedding, for example), or as an advertisement (putting one's resume online) but their primary reason for being is personal gratification for the creator. For an ego site, ongoing interactivity is secondary to the original act of creation.

The success of an ego site can be assessed quite simply: is the creator satisfied or not? No further evaluation is required.

Resource sites attempt to provide information or some other data resource to their readers. This is primarily one-way communication, with the site developer providing information he/she wants to share in some way and site visitors consuming this resource. Most government sites, and most statistical sites, follow this model. Many corporate sites are primarily resource oriented, providing company profiles, product descriptions, and other marketing documents. The resources need not be serious, however. Most entertainment-oriented sites, such as online magazines, lists of jokes, concert schedules, etc. can be categorized as resource sites.

The success of a resource site should be assessed by asking whether most users found the information they were looking for (or discovered that the desired information was – legitimately – not present on the site), how quickly they found it, how many wrong turns they made along the way, and how satisfied they were with the quality of the interaction.

Transaction sites, by contrast, involve two-way communication. Both the site owner and the user have something to contribute to the interaction. Commercial sites, such as online bookstores or stock brokerages, are typically transactional in nature (the site owner provides a list of products; the user assembles a list of desired goods, then adds a shipping address and a credit card number; finally the site owner physically sends a product). Non-commercial transaction sites would include online surveys.

The success of a transaction site can be measured with the same tools as a resource site, adding a measure of how many successful transactions were completed.

The remainder of this paper will focus primarily on resource sites. All site types, however, can benefit by emulating Shaker craftsmen and women and concentrating on the essentials of the task at hand.

Web Users

Like site types, users of the Web can be broken into three categories:

- Non-discretionary;
- Discretionary, directed;
- Discretionary, casual.

Non-discretionary users include those who, for whatever reason, are required to access a particular Web site. This might be an organizational intranet which holds the only copy of a current telephone directory. Non-discretionary users do not have the freedom not to use a particular Web site.

Discretionary, directed users are those who are not required to use a particular Web site, but who are strongly motivated to do so. This may be because they believe other channels for obtaining the necessary information are too cumbersome, expensive, or slow. A discretionary, directed user will try very hard to make sense out of a Web site, and to obtain what he/she is looking for, but will leave if the experience proves too frustrating or the desired information does not appear to be present.

Discretionary, casual users are the stereotype "surfers". Such users may stumble upon a given site, and may be willing to spend a little time investigating, but are not particularly motivated to stay and are likely to leave as soon as they get frustrated or bored.

Different audiences require different approaches to design and implementation. In general, non-discretionary and discretionary, directed users are likely to focus immediately on how well a site supports their particular task, while casual users may be more 'seducible' by superficial attractions. Ultimately, however, even a surfer requires meaningful and well-organized content to remain at a site or return to it.

Naturally there are many other user and usage characteristics that are relevant to the Web site designer: familiarity with the topic, familiarity with the technology, frequency and duration of sessions, etc.

Site Perspectives

The single most important predictor of site utility is the perspective from which the site has been designed. Just as there are two participants in any interaction between a user and a Web site (the user and the site designer), so there are two perspectives around which the site can be designed.

The first is provider-centered. This is a site where the designer cataloged everything his/her organization had to offer, then built a site around the material thus assembled. Provider centered sites often are structured to follow a corporate organization chart. Provider-centered home pages frequently feature a list of links that begins with "Welcome to Our Site", perhaps a picture of the company president or agency head, "About the Organization", perhaps an organization chart, annual report, etc. Most of this information may seem extremely important to employees of the organization, but is seldom of interest to outsiders who are looking for concrete information.

A site that organizes information about product recalls by the date of the press release in which the recall was announced is provider-centered. It is easy to imagine that internal to the company such documents are stored and referenced chronologically. A user of the site, however, is likely to be looking for a particular product, and will probably not know when the recall was announced.

By the same token, any site that has an 'under construction' page is showing signs of provider centricity. The real meaning of 'under construction' is the site developer saying to the user "I know there is something missing here. It's just not ready yet. Don't bug me about it." Meanwhile the user's expectations have been raised, he/she has been tricked into an extra click, and has paid a penalty in unnecessary wait time and possible confusion. 'Under construction' pages are invariably frustrating for the user; they serve only to bolster the self-esteem of the developer.

Provider-centered sites, regardless of the intent of the creator, end up being ego sites more often than not.

User-centered sites, by contrast, are usually structured to follow specific tasks. Here the designer has attempted to determine what users are likely to wish to accomplish when visiting the site, and has structured the site around those anticipated demands. Whereas corporate information like organization charts, mission statements, or annual reports may still be present on the site—after all, some users are likely to be interested—it will be low on the list, not the first thing a user will see. Instead, items of interest to the user will be displayed most prominently. The site as a whole will be structured so that information of interest to the largest number of users is the easiest to reach, while less popular pages might take a few more clicks.

Building a user-centered site requires a very good understanding of the user population. This understanding can come from past interactions with a customer base, from studies of projected user behavior and demands, or from analysis of what users are actually doing on an existing Web site.

Developer intuition concerning user needs and desires is typically not very good. People steeped in either a technology or an organizational culture are frequently not able to leave their knowledge behind and project the experience of either a less knowledgeable individual or an individual with a markedly different perspective. Some sort of objective user study is usually required.

Shakers were both farmers and designers of barns; they understood the task and the technologies very well. Since Web developers cannot always be experts in everything required to design appropriate sites (the task domain, the technology, and the techniques of user-centered design) they can draw on outside resources such as guidelines documents and work with usability professionals.

Many designers are thrown by the potential diversity of users. "How can I possibly design a site for a particular user population when literally anyone in the world could stumble across my site?" It is certainly true that the Web user population is huge. But some initial development decisions must be made, without which no trade-offs can ever be approached systematically. This can be as simple as assuming understanding of the English language, or as demanding as requiring facility with college level statistics and a sophisticated understanding of survey methodology. Different portions of a site can be optimized for distinct user populations (several sites currently suggest different paths depending on whether the user is a high school student, a teacher, a researcher, or a legislative aide), but no site can be built for "everyone."

A user-centered site will usually require several iterations to get right, and can be expected to evolve along with the user population.

Components of Web Site Design

There are two main components of Web site design: page design and dialog design.

Page design comprises all the elements that can been seen in one piece through a browser: text and pictures along with headers, footers, icons, banners, buttons, and links arranged in different typefaces and fonts, columns, tables, etc. The objective of good page design is to position elements in such a way that important information is easily recognized, less important information can be located with only a little effort, and unimportant information has been eliminated. Information, in this context, may include text, graphics, menu items and navigation buttons, or site identifiers.

Web page design corresponds to screen or window design in traditional interactive systems, or page layout in paper-based graphic design. A great deal of research has been done in this area, much of which can be transferred to the Web with only slight modification.

Dialog design has to do with the back-and-forth between a user and a Web site. Typically the user enters a site (through a link, a bookmark, or by typing a URL) and is presented with a page. The user spends some time looking at the page, then clicks on a link or a browser button. The system responds by presenting a new page. The user looks at that, then clicks somewhere, and so forth. There is an interaction between the user and the Web site. With any luck, the Web site is responding to the user's actions in a meaningful and predictable way, and the user, in turn, is reacting to the Web site. This can be seen as a conversation between user and Web site, or between user and site designer.

The objective of good dialog design is to facilitate the user finding the page he/she is looking for (i.e. the page or pages on which the desired information resides) in the most effective way possible. The quality and efficacy of a particular dialog design can only be assessed in the context of its use: specific users performing specific tasks.

Nobody's job description reads "use the computer". Instead, workers are made responsible for specific tasks, and (one hopes) given the necessary tools to accomplish those tasks. The computer is such a tool; the Web site becomes such a tool. Thus effective dialog design is directed towards facilitating task completion in an accurate, comprehensive, and rapid manner.

The Software Development Life Cycle

Successful Web sites rarely happen by accident. A careful development process will increase the chances of a useful site.

In the past an illusion of ease ("just HTML'ize these documents and put them up on the server") coupled with unrealistic deadlines resulted in a plethora of chaotic sites, sites with little or no advance planning, little or no stylistic coordination, and no process in place for controlled updates, corrections, or expansions to the site once it had been released.

The traditional software development life cycle (SDLC) can serve as a model for a proper development process:

1. Analysis
2. Testing
3. Design
4. Testing
5. Implementation
6. Testing
7. Iterate until correct

Analysis is the phase during which the analyst determines what needs to be accomplished. In a Web context this means identifying the target user population(s); determining what these users want or need from the proposed Web site; and determining what organizational goals are to be met through the Web site. The functionality of a system is specified during analysis.

Design is the life cycle phase during which the designer determines how the required functionality will be provided. In a Web context this includes identifying the content to be placed online, determining the site structure, and developing a uniform style guide.

Implementation is where the site is actually built. Existing documents are converted to HTML. New documents and graphics are created, as are forms, CGI scripts, and applets. Links are inserted. The Web server is brought up and pointed to a home page.

Testing is an ongoing activity, particularly important towards the end of each life cycle phase. The earlier an error or misunderstanding can be identified, the cheaper it will be to correct. It must be recognized, however, that later phases tend to uncover insufficiencies in earlier phases. Thus design may well highlight missing functionality, and implementation will surely point out gaps in the design. So developers should be willing to revisit each life cycle stage more than once. Development can be viewed less as a waterfall, where each phase leads clearly to the next, and more as a spiral, with each phase revisited as often as necessary.

All of the above phases have a traditional functionality component and a usability component. The two are not fully independent, but neither are they the same thing. Insofar as careful methodologies have been applied to Web development in the past, most of it has been directed towards functionality: is the HTML correct, do the links point to the correct pages, is the appropriate content present, are the scripts and applets fully debugged. This is critical to a fully functioning, successful site. Equally important are other questions: Does this make sense to our expected user population? Can new users figure out where to go? Can experienced users take advantage of shortcuts? Does the site assist the users in accomplishing their underlying goals? Usability analysis, usability design, and usability testing run parallel to their functional counterparts.

The SDLC is not unique; some general equivalent exists in most professional fields. Publishing, for example: common advice to authors would include "identify the audience before beginning to write," "clarify in your own mind what you wanted to say," "begin with an outline, then flesh out the outline, then have others read and critique the work."

Shaker Principles Tailored to Web Systems

The preceding sections have focused primarily on utility and quality – how these can be determined and how they can be produced. It is time to return to the other two Shaker principles, simplicity and elegance.

Webster's Dictionary gives one definition of simplicity as "absence of affectation or pretense," and elegance as "grace and restraint of style." In the Web context, these characteristics define a design philosophy which concentrates on supporting, enhancing, and emphasizing the underlying content by careful arrangement and use of graphical elements, but never allows the design itself to become prominent. If the user becomes aware of the design beyond an almost subliminal sense of aesthetic satisfaction, the effort has failed.

In an earlier paper Levi and Conrad (1996) took a set of general-purpose usability principles originally proposed by Jakob Nielsen and modified them to apply specifically to Web sites. Of the nine principles thus derived, five are relevant to this paper because they deal with simplicity and elegance:

Design aesthetic and minimalist systems

Create visually pleasing displays. Eliminate information which is irrelevant or distracting.

Simple does not mean ugly or boring. Shaker crafts are quite beautiful. Web sites can be engaging and visually pleasing without succumbing to excess. The key point is that no amount of graphic or other technical virtuosity can make up for insufficient, inaccurate, or poorly organized content. On the contrary, baroque or otherwise unnecessary stylistic embellishment (see below for examples) frequently hides or draws attention away from the substance of a site.

Be Consistent

Indicate similar concepts through identical terminology and graphics. Adhere to uniform conventions for layout, formatting, typefaces, labeling, etc.

The importance of consistency cannot be overemphasized. Not only will use of uniform vocabulary and graphic layout add to site aesthetics, but it will enhance utility as well. In a 1996 study Mahajan and Shneiderman report up to 30 percent degradation in speed of task completion when synonyms such as question/inquiry, search/browse, or counselor/advisor were used on different screens of an application. According to this study, such inconsistencies in wording decrease user's performance regardless of the user's level of expertise.

Speak the users' language

Use words, phrases, and concepts familiar to the user. Present information in a natural and logical order. Avoid jargon wherever possible. Use acronyms cautiously, usually spelling out the complete phrase upon first use (note that in a hypertext system the designer cannot assume that the reader has already traversed any given page, so such clarification might be required every time an acronym is used).

A site designed explicitly for users highly familiar with a given task domain will probably employ a specialized vocabulary; a site designed for a general audience must use a more accessible vocabulary.

Build flexible and efficient systems

Accommodate a range of user sophistication and diverse user goals. Lay out screens so that frequently accessed information is easily found. Provide instructions where useful.

The goal is to build a Web site that is clear enough for a novice or intermittent user to navigate with relative ease, yet is powerful enough or contains sufficient shortcuts so that an experienced user is not slowed down unnecessarily.

Simple does not mean simplistic. Web sites can and should contain a wealth and richness of material. A well-structured page, for example, can support scores of links without being confusing or overwhelming.

Don't lie to the user

Eliminate erroneous or misleading links. Do not refer to missing information.

Some Things to Avoid

As a general rule, anything that detracts from ease of use, including legibility, comprehension, ability to focus, etc., should be eliminated. New techniques, such as extensions to HTML or programming languages such as Java or ActiveX, can usually be used either to facilitate users' tasks or to complicate them. Web developers should be aware that technology for technology's sake (cool for cool's sake) is unlikely to help the users.

Backgrounds

Since at this time most computer screens are rather grainy, the pixels in background graphics tend to interfere with foreground text, making it difficult to read. Virtually all patterned or textured backgrounds fall into this category. Since there is already a performance penalty in reading off a screen rather than reading from reasonable-quality paper, such further degradation just adds insult to injury.

Equally problematic are poor choices of background and foreground color. Pastel on pastel, or bright contrasting colors, may provide a striking and memorable display, but will not make reading any easier.

Scrolling text, blinking text, animated graphics

These can all be lumped together as "gratuitous moving stuff". Movement anywhere in one's field of vision is very seductive and distracting; this draws the user's attention away from content and degrades performance and possibly comprehension.

Some concepts can be very effectively illustrated through animation; one of the great strengths of a computer screen as opposed to a printed sheet of paper lies in the ability to dynamically change what is displayed on a CRT. Designer should use such capabilities wherever they enhance or explicate the subject matter. But waving flags, grinning monsters, or slithering marquees are unlikely to add much value.

Frames

Frames are an attempt to solve a real problem: how to present multiple views of an information space or task domain. Unfortunately they are a failed attempt. Frames break the common metaphors of the Web (one page at a time), and don't work well with many

browsers (the back button behaves erratically, bookmarks and printing may not work as expected). What's more, they don't even do a very good job at the problem they were intended to solve. Coordinating multiple frames so that an action in one frame updates multiple other frames is difficult if not impossible; certainly few if any sites have implemented such functionality correctly.

PDF files

Adobe's Portable Document Format (PDF) is perhaps the best way to support perfect facsimile printing, but what works well on paper does not necessarily work well on a computer monitor. PDF files typically do not render well on screen (the typeface can be fuzzy depending on how the file is produced). Layout options such as multiple columns depend on having an entire page visible at once; this is typically not the case on a screen, so the user ends up paging down one column, then returning to the top and paging down again, and then perhaps yet a third time. Such scrolling is time consuming, error-prone, distracting, and a waste of precious screen real estate. Finally, having a PDF reader as a browser add-in makes the user's work environment that much more complicated by adding another set of buttons and commands that may work subtly different from the buttons and commands which are part of the standard browser.

If the designer's intent is to provide material to be printed, PDF is a very good choice of formats. If the intent is for users to view documents on screen, provide HTML. If users can be expected to do both, provide both formats.

Conclusion

The World Wide Web is an interactive information dissemination mechanism unprecedented in scale and accessibility. Never before has it been so easy to provide computer-supported data, news, knowledge, analysis, opinions, and advertising to so many people around the globe. And never before has it been possible for a computer to confuse, frustrate, and annoy so many people in such a short time.

Given the vastness of the information space represented by the Web, only the best sites can be expected to thrive. Slick technology tricks are fascinating the first time, but become boring after the second or third iteration. Fancy design curlicues become stale equally quickly.

If a site is to endure, it is the value of its content that will make the difference. For the value of content to be apparent, the site must be designed to sustain and enhance that content. Shaker handicraft, and the philosophy that gave rise to it, provides the model to do just that.

In a paper on Shaker architecture, Robert P. Emlen wrote:

> "Nineteenth-century visitors wrote of the unity of design in Shaker villages. Developed according to the community's standards and requirements, the buildings in a Shaker village are more consistent in appearance than those of the neighboring farms. Their clustering on the land, the way they relate to one another in function and scale, the consistency of aesthetic choices employed by Shaker craftsmen, all attest to that communal society of spiritual brethren and sisters devoted to creating an ideal life on earth."

If a few words in the above quote were changed, it could describe a well-crafted Web site. Anyone should be very proud to have created such a Web site

References

Levi, M. D. and Conrad, F. G. (1996), "A Heuristic Evaluation of a World Wide Web Prototype," *Interactions Magazine*, July/August.

Mahajan, R. and Shneiderman, B. (1996) "Visual and Textual Consistency Checking Tools for Graphical User Interfaces," Human-Computer Interaction Laboratory, University of Maryland.

Emlen, R. P. (1995), *The Distinctiveness of Shaker Architecture*; Originally published as the foreword in: Nicoletta, J. (1995), *The Architecture of the Shakers*, Countryman Press.

Michael D. Levi
U.S. Bureau of Labor Statistics
levi_m@bls.gov

ISEA Discrete Global Grids

By Dan Carr, Ralph Kahn, Kevin Sahr,
and Tony Olsen

1. Introduction

This article describes a recently proposed standard, ISEA discrete global grids, for gridding information on the surface of the earth. The acronym ISEA stands for icosahedral Snyder equal area. The grid cells not only have equal areas, they are hexagons when projected onto an icosahedron! Being an advocate of hexagon binning, and corresponding graphics, my (Dan) enthusiasm is such that I want to call attention to this new approach.

Jon Kimerling, a geosciences professor, won the Oregon State University (OSU) Milton Harris Award for his research on the topic. Since I was peripherally involved in the development of the grids, I asked Kevin, Ralph and Tony to help me with portions of this article. They all interacted with Jon in the research and we all shared the desire to promote ISEA grids. Kevin, an admirer of Buckminster Fuller, developed many of the ISEA algorithms and graphics. Along with OSU collaborators Mathew Gregory and Larry Hughes, he developed a web site, `http://bufo.geo.orst.edu/tc/firma/gg/`, that contains foldable figures, descriptions and three reference lists. At the Milton Harris Seminar, Ralph provided an excellent overview on the relevance to summarization and presentation of data from the Earth Observing System (EOS). EOS is a series of NASA satellites designed to detect and monitor global climate change, starting in summer 1998. Material from Ralph's talk is available at the above web site and we include portions here. Tony was the instigator of the whole development with his push to develop a globally consistent environmental sampling methodology. Tony and his EMAP co-workers helped EPA Regions, states and nations develop environmental sampling plans using an EMAP grid that was developed in the early stages of the research.

The structure of this article is as follows. Section 2 traces some of the history behind the ISEA grids. Sections 3 describes the ISEA grid. Section 4 introduces a potential application, storage of summaries derived from Earth Observing System sensors. Section 5 discusses graphics for hexagon grids, Splus algorithms for low resolution ISEA grid smoothing on the globe, and new, resolution 9 (about 200,000 cells) binned map of global elevation data. Section 6 closes with challenges for future research.

2. The Recent Historical Development of ISEA Grids

The impetus for the global grid system came from what many would call an unusual perspective - survey sampling. In 1989, Denis White and Scott Overton, a geographer and a statistician from Oregon State University, held a workshop in Corvallis, Oregon, to discuss the geographic requirements for a general survey design. The survey design would be the foundation for all surveys conducted as part of the Environmental Monitoring and Assessment Program (EMAP) (Messer et al., 1991; Stevens, 1994). Scott Overton, leader of the survey design effort, recommended basing the design on a systematic grid with a random start (Overton et al., 1990). We all know how to accomplish that for planar surfaces but when the design must cover the United States, we are faced with a non-planar surface - the earth. Jon Kimerling, along with the other geographers at the workshop, devised a discrete grid system that satisfied the needs of EMAP at that time (White et al., 1992). The system used a truncated icosahedron model of the earth with a triangular point grid applied to the large hexagon plates. It worked for the contiguous 48 states. However, the initial discrete grid system did not solve all the underlying issues and the embedded triangular grid structure had elements that were arbitrary. As an example, the EMAP team also applied the grid to China, Russia, and Indonesia. The team knew problems would exist for China and Russia, as a single large hexagon plate would not cover either country. Although small in area, Indonesia is stretched out and also is not covered! The initial discrete grid system had problems at the boundaries of the plates.

In 1993, Tony Olsen, faced with these inadequacies, initiated a research effort with Jon Kimerling, Kevin Sahr, and Denis White in the OSU Geosciences department to investigate an alternative discrete global grid system. Tony required the system to be truly global and result in an equal area tessellation. He also had a preference for compact areas, minimal shape distortion, a triangular point grid, and a hierarchical grid structure allowing multiple grid densities. These characteristics would enable global implementation of survey designs for continuous spatial populations (Stevens, 1997).

My (Dan) formal involvement did not start until Oregon State researchers held a workshop on discrete global grids at Santa Barbara in 1994. Others in attendance were Denis White, Jon Kimerling, Michael Goodchild, Waldo Tobler, Tony Olsen, Geoff Dutton, Frank Davis, and David Mark. Many in the group had already developed their own approaches for global grids. Waldo Tobler was already using his methodology to show populations on the global. Geoff Dutton had developed a gridding system that modeled the earth as an octahedron with an appropriate map projection. Kimerling and White presented their icosahedral alternative to the EMAP (truncated) icosahedron model. There are of course additional approaches that work more directly on the globe. (For a recent discussion of distributing points on a sphere, see Saff and Kuijlaars 1997). All methods must deal with the fact that there is no perfect regular partition for the surface of a sphere. One member noted that there is always at least one singularity, as he humorously pointed to the bald spot on his head. Michael Goodchild suggested that the meeting produce a list of desirable properties for gridding systems. The list appears below. Tony knew my objective when I proposed cells being "compact" (having a small dimensionless second central moment - see Conway and Sloane 1988). It was my attempt to promote hexagon cells.

At the Santa Barbara workshop Michael Goodchild proposed a prioritized attribute list for a discrete global grid system. The elements of the list are: the domain is the globe (sphere, spheroid), areas exhaustively cover the domain, areas are equal in size, areas are compact, areas are equal in shape, areas have same number of edges, edges of areas are of equal length, edges of areas are straight on some projection, areas form a hierarchy preserving some properties for $m < n$ areas, each area is associated with only one point, points are maximally central within areas, points are equidistant, points form a hierarchy preserving some properties for $m < n$ points, addresses of points and areas are regular and reflect other properties.

With methods and evaluation criteria at hand, the group planned two sessions at the GIS/LIS94 meeting. To have something to contribute, I, on the spur of the moment, concocted a method based on projecting 3-D lattice points "near" a sphere surface onto the surface. My subsequent attempts with different lattices, packings, and notions of near, did not lead to hexagon patterns over the whole sphere. The redeeming features of my talk at GIS/LIS94 turned out to be the color anaglyph stereo viewgraphs and brevity. The other presentations carried the two sessions.

After GIS/LIS94, work proceeded on the icosahedron model (see Kimerling, Sahr, Song, White, and Iltis, 1995). I called the research to the attention of Ralph Kahn (NASA-JPL) who was looking for better ways to summarize the global data expected from EOS. Tony sought to involve Noel Cressie for dealing with spatial estimation issues.

Jon Kimerling subsequently won the Milton Harris Award. In May 1997, he held the Milton Harris Award Symposium on Global Grids: New Approaches to Global Data Analysis. In addition to presentations by team members Kevin Sahr and Denis White, he invited Ralph Kahn (NASA -JPL), Noel Cressie (Iowa State University), Ross Kiester (USDA-Forest Science Laboratory), Tony Olsen (USEPA-Corvallis) and myself to make presentations. The following sections cover selected topics from the Symposium and Kevin's web site.

3. Icosahedral Snyder Equal Area (ISEA) Grids

The S in ISEA refers to John P. Snyder. He came out of retirement specifically to address projection problems with the original EMAP grid (see Snyder, 1992). He developed the equal area projection that underlies the gridding system. His work at the U.S. Geological Survey on map projections is known by all who spend any time with map projections. John Snyder died this year. By all reports, he was a modest man who would not seek to have procedures named after him. Nonetheless, in honor of his contributions to the field of map projections, those developing the gridding system have desired to use his name.

ISEA grids are simple in concept. Begin with a Snyder Equal Area projection to a regular icosahedron (see the stereo pairs in Figure 1) inscribed in a sphere. In each of the 20 equilateral triangle faces of the icosahedron inscribe a hexagon by dividing each triangle edge into thirds (see the large gray hexagon in Figure 2). Then project the hexagon back onto the sphere using the Inverse Snyder Icosahedral equal area projection. This yields a coarse-resolution equal area grid called the resolution 1 grid. It consists of 20 hexagons on the surface of the sphere and 12 pentagons centered on the 12 vertices of the icosahedron.

To form higher resolution grids, tessellate each equilateral triangle in the planar view with more hexagons and use the inverse projection back to the sphere. The details of the regular tessellation are as follows: Always center a hexagon about the center point of the

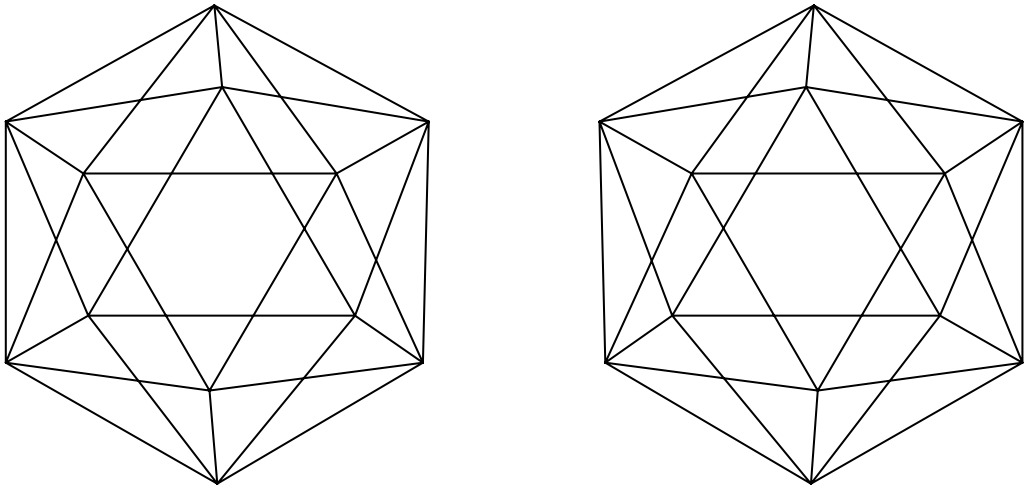


Figure 1. Stereo pairs of a regular icosahedron.

equilateral triangle. For odd resolution grids, orient the hexagon so its base is parallel to the base of the triangle. For even resolution grids orient the hexagon so a vertex points at the base of the triangle. (Figure 2 shows the central hexagons for resolutions 1 and 2 in gray and black, respectively.) Select the edge length of a resolution $r+1$ hexagon so it is $1/\sqrt{3}$ times the edge length of a resolution r hexagon. Thus, the area of a hexagon reduces by a factor of 3 with each increase in resolution. As the resolution increases by 1, the tessellation procedure produces a hexagon centered on each hexagon vertex and center point of the lower resolution tessellation.

As illustrated in Figure 2, the procedure partitions a lower resolution hexagon cell into one central cell and six fractional (1/3) cells. This is not as simple as partitioning a large square into exactly four smaller squares. While the merits of strictly nesting cells within cells depend on the context, one clear merit is aggregation simplicity. The ISEA fractional cells create aggregation and disaggregation problems that are currently under investigation.

The orientation of the icosahedron relative to the globe is an important consideration. The selected orientation for the ISEA grid creates symmetry about the equator. This is desirable for numerical modeling purposes. There are always 12 pentagon cells about the vertices of the icosahedron. The selected orientation places 11 of the pentagon cells over water areas, so that most land mass views will be completely composed of hexagons.

Table 1 on the next page (taken from Kevin's web site) gives the number of cells and characteristic hexagon edge lengths for ISEA grids of increasing resolution.

The advantages of the ISEA grids are (1) they have irregularities (12 pentagon cells) that are minor nuisances rather than being pathological singularities, (2) they are suitable for modeling on all parts of the globe including the poles, (3) they preserve symmetry about the equator, (4) they provide an infinite nesting of equal-area sub-grids, and (5) they provide a basis for uniform global density of sampling for data at all spatial resolutions. The grid facilitates comparisons between high and low latitude data and high and low spatial-resolution data. The grid also improves the isotropy of finite-difference quantities compared to those calculated for rectangular grid schemes. For example Fisch, Hasslacher and Pomeau (1986) note that two-dimensional Navier-Stokes implementations are optimal with hexagons. Finally, no ambiguity exists about nearest neighbors as all nearest neighbor cells share an edge with a reference cell and their distances to the center of a reference cell are nearly equal.

4. EOS and the Potential Application of ISEA Grids

There are many potential applications of ISEA grids. We are particularly mindful of NASA's Earth Observing System and the wealth of global earth science data that it will collect. The EOS AM-1 Platform is scheduled for launch in June 1998. The summarization of this data provides a rapidly approaching opportunity to use ISEA grids.

Resolution	Number of Cells	Length Scale (km)
1	32	4,684.2571
2	92	2,694.2932
3	272	1,553.6212
4	812	896.6139
5	2,432	517.5892
6	7,292	298.8166
7	21,872	172.5192
8	65,612	99.6035
9	196,832	57.5060
10	590,492	33.2011
11	1,771,472	19.1687
12	5,314,412	11.0670
13	15,943,232	6.3896
14	47,829,692	3.6890
15	143,489,072	2.1299
16	430,467,212	1.2297
17	1,291,401,632	0.7100
18	3,874,204,892	0.4099

Table 1. The number of cells and the characteristic hexagon edge lengths for ISEA grids of increasing resolution.

More specifically, ISEA grids are relevant to Level 3 Products in the EOS Data Product Classification. Level 1 Products involve raw radiances with geometric and radiometric calibration. Level 2 Products are geophysical parameters at the highest resolution available. These data sets preserve the non-uniform spatial and temporal sampling of the satellite instruments. Level 3 products are globally and temporally uniform data sets. Level 3 products are needed where large-scale, uniform coverage is required (e.g., global-scale budgets, and problems that depend on data sets from multiple sources). Various tradeoffs will drive the selection of spatial and temporal scales chosen for Level 3 standard products so a multiple-resolution equal-area global grid system is immediately relevant.

The massive amount of data and the resolution issues drive the need for professional algorithms. For example, one instrument on the platform (MISR) will help characterize, on a global basis, atmospheric aerosol type and optical depth, surface bi-directional reflectance properties, and cloud properties. The amount of data to be collected from this one sensor is enormous. With a spatial resolution of 16 values per km^2 and 36 channels, a global description will involve 2.9×1011 basic measurements. The MISR collection rates will be 40 Gbytes/day of raw data, 300 Gbytes/day total data, and

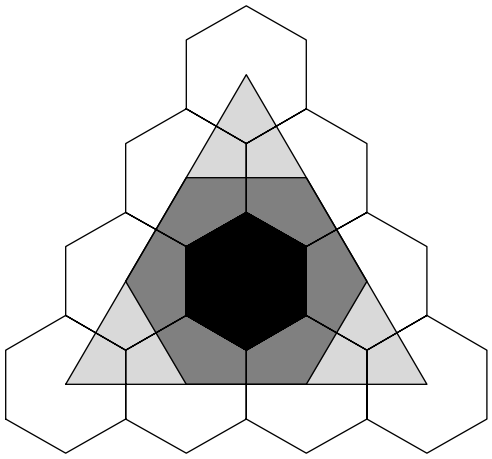


Figure 2. Subdividing the faces of a regular icosahedron: Gray and black regions represent the central hexagons for resolutions 1 and 2, respectively.

15-100 Tbytes/yr for at least 5 years. The computing tools developed for the graphics in this article will not handle such data.

Of course there is the old alternative to handle Level 3 gridding, equal angle grids. The equal angle grid relies on the global latitude-longitude system and uses a cylindrical map projection. It typically has a spatial resolution of 1° (112 km) and sub-grids based on equal-angle divisions at 0.5° (56 km) and 0.25° (28 km). The advantages of the equal angle grid are that the latitude-longitude system is convenient, familiar, and entrenched. Also very important is the fact that the results are easy to represent in 2-D arrays. However, the equal angle approach leads to several issues such as rapidly changing spatial resolution at high latitudes, non-uniform resolution for fine scales, ambiguity of nearest neighbor operations and problems in representing data at multiple scales. The current solutions to the multiple scale problem are discipline-specific variations, for example, specialized grids for polar and for local high-resolution applications. The ISEA approach, among other things, would provide compatible grids across disciplines.

Those seeking additional information on alternative grids and EOS sensors can access Ralph's descriptions at `http://bufo.goe.orst.edu/tc/firma/gg/kahntoc.html`. Of particular interest is an example that shows the huge discrepancies that can result from changing from one grid to another and back. Those seeking more information on Level 2 products or discussion of problems in validating satellite derived parameters can start with Kahn et al. (1991).

5. Graphics for Hexagon Cells, Global Binning and Foldable Figures

Many graphics are available for hexagon cells. Some of these graphics involve spatial smoothing. The figure on page 35 (adapted from Yang and Carr, 1995) shows a breeding bird diversity map based on smoothing to the previous EMAP grid. The brute force smoothing of ten year prevalence data for 615 bird species to the 13000 cell grid involved close to 8 million local logistic regressions! Soon after an article on mortality map smoothing (Carr and Pickle, 1993), Andrew Carr and I created a point and click Splus function (UNIX only) for selecting U.S. cancer mortality rates and smoothing the rates to hexagon grids. The smoother in that context was `loess`. This collection of functions is available as an Splus `data.dump` file, `nchs.dmp`, by anonymous ftp to `galaxy.gmu.edu`. It is located in `pub/dcarr/newsletter/nchs`. While hexagon cell maps are relatively uncommon, the general notion of choropleth maps is, of course, not new.

There are several sources for innovative hexagon graphics. Carr et al. (1987), and Carr (1991) present various density representations and a practical bivariate generalization of box plots. Kevin and Ron Keister (personal communication) have developed an approach for showing the change from cell to cell by coloring triangles within the hexagons. Papers of Carr (1989), Carr (1991), and Carr, Olsen and White (1992) address symbol congestion control with the first showing a stereo regression diagnostic and the last two papers focusing attention on maps. The idea is to partition the map (or plot) using hexagon cells and provide symbols to represent the summary for each cell with data. For example, the angle of a ray glyph can represent a continuous variable, such as a trend estimated from a time series. The rays can point down (below horizontal) for small values or negative trends and up for large values or positive trends. The rays can plot on top of confidence arcs that represent associated confidence bounds. Two rays with common origin, one pointing to the left and one to the right can easily represent two continuous variables on a map.

Splus derivatives of my 2-D lattice functions now facilitate hexagon binning, gray level erosion, smoothing, hexagon plotting and ray plotting. Familiarity and convenience suggested following the conventions in this software when developing binning, smoothing and display procedures for global grids. The result is a set of closely related Splus functions for low resolution grids.

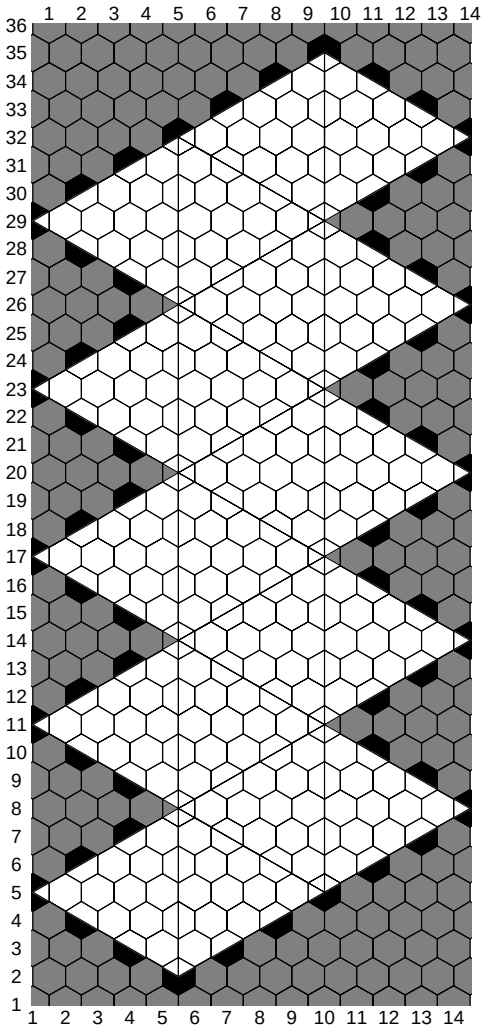


Figure 3. A flattened icosahedron foldable figure for resolution three.

The task addressed here is binning of global ETOPO $5'$ minute elevation data into an ISEA resolution 9 grid. Conceptually the computation of a cell id for a latitude longitude pair involves four steps. First, the Snyder Equal Area projection produces coordinates in one of the 20 triangles of the icosahedron. Second, an affine transformation (one for each of the twenty triangles) maps the coordinates into a flattened icosahedron foldable figure as shown for resolution three in Figure 3. Next, a hexagon index routine (like `xy2cell` in Splus) produces a planar cell id. The last step uses a look-up table (a vector) to convert planar cell id into a globe cell id. (Globe cell ids are integers ranging from 1 to the number of globe cells.) Given globe cell ids, computation of summary statistics for data falling in the cells is straightforward. The work was in generating the pla-

nar cell id to globe cell id conversion vector. The re-indexing omits unused hexagons that cover the Figure 3 rectangle. The re-indexing also accounts for split foldable figure planar cells (for example those containing the five left triangles tips) that are really parts of the same cell on the globe.

Rather than reading the large ETOPO file into Splus, I modified one of Kevin's programs. The program imports the resolution 9 re-indexing vector generated in Splus and the bins on the fly. After reading the binned results into Splus, I used three additional bookkeeping vectors to compute hexagon boundaries and colors for all cells (196832 globe cells and the 832 cloned cells) in the foldable figure. The figure on page 36 shows the average elevation for each cell. Jon Kimerling suggested the basic elevation and depth coloring scheme. A further refinement requiring additional data would be to distinguish land hexagons that are slightly below sea level from ocean floor hexagons. I had problems producing the whole postscript file for the figure on page 36 so I wrote out pieces and connected them using Unix tools. A procedure that reads a value for a location and writes a hexagon directly to a file would be better for graphics output.

My Splus routines for odd resolution ISEA grids are available via anonymous ftp to `galaxy.gmu.edu`. Change directory to `pub/dcarr/newsletter/isea`. There is a README document describing the various functions. For example, one function produces a globe cell near neighbor pointer matrix (for low resolutions). Another function uses this matrix for smoothing values on the globe. (More sophisticated smoothers could restrict domains to land masses, oceans land-ocean boundaries, or address flow constraints.) A script file shows the process of producing a foldable icosahedron. The script starts by randomly generating a vector of 2432 values that implicitly correspond to globe cells in a resolution 5 grid. After smoothing the values and converting them into colors for hexagons, the script plots the hexagons along with tabs for gluing. Creasing along the lines shown in Figure 3 helps in the construction. I have made several figures for the holiday season. Postscript files for different examples and sizes are in the above directory. Kevin's web site contains more examples including one of my favorites. The favorite is an amazing gift from the antiquity of basketry, a six great circle weave.

For stereo presentations on a globe, a simple approach partitions each hexagon into six triangles. The plotting step then renders triangles whose vertices result from the inverse Snyder equal area projection.

6. Additional Challenges and Closing Remarks

This article describes a 1-D indexing system that is viable for modest odd resolution grids. The basic indexing is for hexagon cells that cover a rectangle bounding the planar icosahedron view. A re-indexing vector, whose length is the number of covering cells, removes the unused and redundant indices. After binning with the new indices, pre-computed x and y vectors provide plotting positions for the planar icosahedron view. The binned results correspond to cells on the globe so a short subscript vector extracts values corresponding to split cells in the planar view. The result of concatenating the two vectors corresponds to the planar x and y coordinates. No doubt a similar approach will work for even resolution grids but the bookkeeping to handling unused and split cells will require some work.

When the grid involves billions of cells, the indexing based on the rectangle bounding the foldable icosahedron planar view may be too wasteful. A first challenge is to develop a more efficient indexing system. Quite possibly this will just cover the twenty triangles with hexagons and handle the cells that cross the edges of touching triangles. A second challenge is to move from a demonstration system to professional quality algorithms for high resolution grids.

There are many tasks to be addressed for a collection of algorithms to be complete. Tony is interested in indexing optimized for subsets of the globe such as the continental U.S. Perhaps the most crucial task is to provide fast, conceptually acceptable algorithms for changing resolutions. As indicated earlier, lack of strictly nested cells at different resolutions poses a problem. The equal area projection approach easily adapts to strictly nested triangles, but that would give up some of the merits of hexagon cells.

A second challenge area is to consider the use of spatial models in producing cell summaries. For example, Ralph has noted that the current procedure for producing pixel values for satellite images involves a simple near neighbor averaging process. Noel Cressie addressed some of the spatial modeling possibilities in his talk at the Harris Seminar.

Assuming the computation issues are solved, we will then face the biggest challenge of all, institutional inertia. Proposing a standard is one thing. Getting scientists in different nations and different disciplines to use it is another.

Acknowledgments

Research related to this article was supported by EPA under cooperative agreement No. CR820820-01-0. The article has not be subjected to the review of the EPA and thus does not necessarily reflect the view of the agency and no official endorsement should be inferred.

References

Carr, D. B. (1989), "Discussion of Regression Diagnostics with Dynamic Graphics," *Technometrics*, 31(3), 293-296.

Carr, D. B. (1991), "Looking at Large Data Sets Using Binned Data Plots," *Computing and Graphics in Statistics*, eds. A. Buja and P. Tukey, Springer-Verlag, New York, New York, 7-39.

Carr, D. B., Littlefield, R. J., Nicholson, W. L. and Littlefield, J. S. (1987), "Scatterplot Matrix Techniques For Large N," *Journal of the American Statistical Association*, 82(398), 424-436.

Carr, D. B., Olsen, A. R., and White, D. (1992), "Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographical Data," *Cartography and Geographic Information Systems*, 19(4), 228-236, 271.

Carr, D. B. and Pickle, L. W. (1993), "Plot Production Issues and Details: Smoothed Cancer Rates and Hexagon Mosaic Maps," *Statistical Computing & Graphics Newsletter*, 4(2), 16-20.

Conway, J. H. and Sloane, N. J. A. (1988), "Coverings, Lattices, and Quantizers," *Sphere Packings, Lattices and Groups*, New York. Springer-Verlag, 56-62.

Fisch, U., Hasslacher, B. and Pomeau, Y. (1986), "Lattice-Gas Automata for the Navier Stokes Equation," *Physical Review Letters*, 56(14), 1505-1508.

Kahn, R., Haskins, R. D., Knighton, J. E., Pursch, A. and Granger-Gallegos, S. (1991), "Validating a Large Geophysical Data Set: Experiences with Satellite-Derived Cloud Parameters," *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, Interface Foundation of North America, Fairfax Station, VA, 133-140.

Messer, J. J., Linthurst, R. A., and Overton, W. S. (1991), "An EPA Program for Monitoring Ecological Status and Trends," *Environmental Monitoring and Assessment*, 17, 67-78.

Overton, W. S., Stevens, D. L. and White, D. (1990), Design Report for EMAP, Environmental Monitoring and Assessment Program, EPA 600/3-91/053, U.S. Environmental Protection Agency, Office of Research and Development, Washington, D.C.

Saff, E. B. and Kuijlaars, A. B. J. (1997), "Distributing Many Points on a Sphere," *Mathematical Intelligencer*, 19(1), 5-11.

Snyder, J. P. (1992), "An Equal-Area Map Projection for Polyhedral Globes." *Cartographica*, 29(1), 10-21.

Stevens, D. L., Jr. (1994), "Implementation of a National Environmental Monitoring Program," *Journal of Environmental Management*, 42, 1-29.

Stevens, D. L., Jr. (1997), "Variable Density Grid-Based Sampling Designs for Continuous Spatial Populations," *Environmetrics*, 8, 167-95.

White, D., Kimerling, A. J., and Overton, W. S. (1992), "Cartographic and Geometric Components of a Global Sampling Design for Environmental Monitoring," *Cartography and Geographic Information Systems*, 19(1), 5-22.

Yang, K. S., Carr, D. B. and O'Connor, R. J. (1995), "Smoothing of Breeding Bird Survey Data to Produce National Biodiversity Estimates," *Computing Science and Statistics, Proceeding of the 27th Symposium on the Interface*, Vol. 27, M. M. Meyer and J.L. Rosenberger (eds.), Interface Foundation of North America, Fairfax Station, VA, 405-409.

Dan Carr
Institute for Computational Sciences and Informatics
George Mason University
dcarr@voxel.galaxy.gmu.edu

Ralph Kahn
Jet Propulsion Laboratory
ralph.kahn@jpl.nasa.go

Kevin Sahr
University of Oregon
sahrk@cs.uoregon.edu

Tony Olsen
EPA National Health and Environmental Effects Research Laboratory
tolsen@mail.cor.epa.gov

1997 Student Paper Competition Winners!

Congratulations to the winners of the 1997 Student Paper Competition sponsored by the Computing Section. Pictured at the 1997 JSM are (from left to right) Wenjiang J. Fu, Gareth James, Alan Gous and Ramani S. Pilla.

Best Contributed Paper Competition

This year at the 1998 Joint Statistical Meetings in Dallas the Statistical Computing Section is sponsoring, for the first time, a competition for best contributed paper. The aim is to recognize outstanding work in statistical computing research among the attendees at the Joint Statistical Meetings.

All papers presented in the contributed sessions sponsored by the Statistical Computing Section are eligible. The selection will be made from papers submitted over the summer to the chairs of the sessions. The award will be presented to the author(s) of the best paper at the joint business meeting/mixer of the Statistical Computing and Statistical Graphics Sections.

ASA Election Results

Below are the names of the candidates who were elected in the recent balloting by ASA members. These results were reported in a memorandum from Ray A. Waller, Executive Director of the ASA, dated June 1, 1998.

President-elect, 1999 (President 2000)
Michael O'Fallon
Mayo Clinic

Vice President (1999-2001)
Fritz Scheuren
Ernst and Young

A number of officers were also elected to the Statistical Computing and Graphics Sections.

Statistical Computing

Chair-Elect
Russell Wolfinger
SAS Institute, Inc.

Program Chair-Elect
J. S. Marron
University of North Carolina

Representative to the Council of Sections (1999-2001)
Leland Wilkinson
SPSS, Inc.

Statistical Graphics

Chair-Elect
Edward J. Wegman
George Mason University

Program Chair-Elect
Daniel Carr
George Mason University

Secretary/Treasurer
Robert L. Newcomb
University of California

Publications Officer
Michael C. Minnotte
Utah State University

Council of Sections Representative (1999-2001)
Bradley Jones
SAS Institute

Newly elected Section Officers should attend our Business meeting/mixer at the Dallas JSM.

We especially want to thank those of you who were candidates, as well as those of you who took the time to vote. Your continuing interest in and support for our Sections' activities are greatly appreciated.

1998 Student Paper Competition

Statistical Computing Section

As in previous years, the Statistical Computing Section sponsored a paper competition for graduate students. The requirements were that the student be the first author of a paper in the area of statistical computing, which might be original methodological research, a novel application, or a software-related project. The four winners are invited to present their papers at a special contributed session at the Joint Statistical Meetings, and the Section pays their expenses to attend. The competition is open to applicants who are students in the fall of the year prior to the competition.

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

- ***Alessandra Brazzale***
  Department of Mathematics
  Swiss Federal Institute of Technology
  "Approximate Conditional Inference in Logistic and Loglinear Models"

- ***Matt Calder***
  Department of Statistics
  Colorado State University
  "Scompile: A Compiler for SPLUS"

- ***Steven Scott***
  Department of Statistics
  Harvard University
  "Bayesian Analysis of a Two State Markov Modulated Poisson Process"

- ***Yan Yu***
  Statistics Center
  Cornell University
  "Fitting Trees to Curve Data: An Application to Time of Day Patterns of International Calls"
  with Diane Lambert

The students will be recognized at the Statistical Computing/Statistical Graphics business meeting at the 1998 Joint Meetings in Dallas, and will make presentations based on their papers in a special contributed session, currently scheduled for Thursday, August 13 at 10:30 am. The papers will also be published in an upcoming issue of the Journal of Computational and Graphical Statistics.

SECTION OFFICERS

Statistical Graphics Section - 1997

Sally C. Morton, Chair
310-393-0411 ext 7360
The Rand Corporation
Sally_Morton@rand.org

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

William DuMouchel, Past–Chair
908-582-7180
AT&T Laboratories
dumouchel@research.att.com

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

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

Mario Peruggia, Newsletter Editor (96-97)
804-924-8298 University of Virginia
mperuggia@virginia.edu

Robert L. Newcomb, Secretary/Treasurer (97-98)
714-824-5366
University of California, Irvine
rnewcomb@uci.edu

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

Lorraine Denby, Rep.(96-98) to Council of Sections
908-582-3292
Bell Laboratories
ld@bell-labs.com

Colin R. Goodall, Rep.(95-97) to Council of Sections
Health Process Management
colin@hdsys.com

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

Statistical Computing Section - 1997

Daryl Pregibon, Chair
908-582-3193
AT&T Laboratories
daryl@research.att.com

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

Sallie Keller-McNulty, Past–Chair
913-532-6883
Kansas State University
sallie@cecil.stat.ksu.edu

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

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

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

Evelyn M. Crowley, Secretary/Treasurer (97-98)
317-494-6030
Purdue University
crowley@purdue.edu

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

MaryAnn H. Hill, Rep.(95-97) to Council of Sections
312-329-2400 SPSS
hill@spss.com

Janis P. Hardwick, Rep.(96-98) Council of Sections
313-769-3211
University of Michigan
jphard@umich.edu

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

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

SECTION OFFICERS

Statistical Graphics Section - 1998

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

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

Sally C. Morton, Past–Chair
310-393-0411 ext 7360
The Rand Corporation
Sally_Morton@rand.org

Edward J. Wegman, Program Chair
703-993-1680
George Mason University
ewegman@gmu.edu

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

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

Robert L. Newcomb, Secretary/Treasurer (97-98)
714-824-5366
University of California, Irvine
rnewcomb@uci.edu

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

Lorraine Denby, Rep.(96-98) to Council of Sections
908-582-3292
Bell Laboratories
ld@bell-labs.com

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

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

Statistical Computing Section - 1998

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

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

Daryl Pregibon, Past–Chair
908-582-3193
AT&T Laboratories
daryl@research.att.com

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

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

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

Merlise Clyde, Secretary/Treasurer (97-98)
919-681-8440
Duke University
clyde@isds.duke.edu

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

Janis P. Hardwick, Rep.(96-98) Council of Sections
313-769-3211
University of Michigan
jphard@umich.edu

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

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

INSIDE

Statistical COMPUTING & GRAPHICS

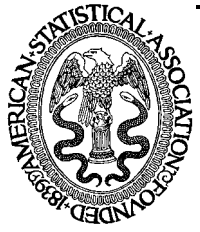
The *Statistical Computing & Statistical Graphics Newsletter* is a publication of the Statistical Computing and Statistical Graphics Sections of the ASA. All communications regarding this publication should be addressed to:

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

Mario Peruggia
Editor, Statistical Graphics Section
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Health Sciences Center Box 600
Charlottesville, VA 22908
(804) 924-8298 • FAX: 924-8437
mperuggia@virginia.edu
http://stat.ohio-state.edu/~peruggia

All communications regarding membership in the ASA and the Statistical Computing or Statistical Graphics Sections, including change of address, should be sent to:

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3402 USA
(703) 684-1221 • FAX (703) 684-2036
asainfo@amstat.org



Nonprofit Organization
U. S. POSTAGE
PAID
Permit No. 50
Summit, NJ 07901

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3402
USA

This publication is available in alternative media on request.