# Keynote Address: Risk Analysis in Geoscience and Remote Sensing

*Thursday, 9:00 am – 10:00 am, Gold and Grey Room:*

David R. Brillinger (Statistics Department, University of California, Berkeley)

**Abstract**

Risk analysis, that is the problem of estimating the probabilities of rare and damaging events, unifies the geosciences. One can mention the risks from: floods, earthquakes, forest fires, space debris. The probabilities may be fed into the computation of insurance premiums. The Poisson process often plays a prominent role, while in the talk marked point processes will have a basic function. Various ways to collect and extrapolate data will be described and examples from various fields will be presented.

# KDD Featured Session: Data Mining (Invited Session)

Organizer and Chair: Arnold F. Goodman

*Thursday, 10:30 am – 10:55 am, Gold Room:*

### Challenges for Data Miners, Statisticians, and Clients Who Depend Upon and Fund Them

Arnold F. Goodman (University of California, Irvine, Center for Statistical Computing), agoodman@uci.edu

**Abstract**

To discover real knowledge: the results must work most (not only some) of the time, account for uncertainty outside (not only inside) the data in the client's world, and add value to such a world. If data mining is to become knowledge discovery, it must refine questions it mines from the data into answers it develops within the data, to then discover knowledge it will find outside the data.

Knowledge discovery rests on three balanced legs of existing client knowledge, computer science and statistics; and it will not stand on either one leg or two legs, or even on three imbalanced legs. No amount of elaborate data manipulation or pragmatic computer science will be a meaningful substitute for statistics in handling uncertainty inside and outside the data or adding client value.

Data mining and statistical methodologies will grow closer to each other during the next 10 years. Perceptive data miners and statisticians will prosper, while the others might consider job hunting.

"The successful... will be those who sense oncoming changes and respond to them. The unsuccessful will be those who remember the past too fondly and too well."

*Thursday, 10:55 am – 11:20 am, Gold Room:*

### Data Mining: Are We There Yet?

Herb Edelstein (Two Crows Corporation), herb@twocrows.com

**Abstract**

Data mining started its move out of the statistics and machine learning ghettos and into the mainstream almost 10 years ago. With great fanfare and a large influx of venture capital, data mining was going to change the very nature of business. Yet data mining products have had relatively modest success in the marketplace because of mistakes on the part of the data mining vendors, resistance from the statistics community, and The reasons include limitations and misplaced a lack of readiness by many prospective end users. This session will look at where vendors have succeeded and failed with their products, what expectations statisticians and end users should have, and suggestions for achieving the potential of this exciting and valuable technology.

*Thursday, 11:20 am – 11:45 am, Gold Room:*

### Magical Thinking in Data Mining

Charles Elkan (Department of Computer Science and Engineering, University of California San Diego), elkan@cs.ucsd.edu

**Abstract**

CoIL challenge 2000 was a supervised learning contest that attracted 43 entries. The authors of 29 entries later wrote explanations of their work. This paper discusses these reports and reaches three main conclusions. First, naive Bayesian classifiers remain competitive in practice: they were used by both the winning entry and the next best entry. Second, identifying feature interactions correctly is important for maximizing predictive accuracy: this was the difference between the winning classifier and all others. Third and most important, too many researchers do not appreciate properly the issue of statistical significance and the danger of overfitting. Given a dataset such as the one for the CoIL contest, it is pointless to apply a very complicated learning algorithm, or to perform a very time-consuming model search. In either case, one is likely to overfit the training data and to fool oneself is estimating predictive accuracy and in discovering useful correlations.

*Thursday, 11:45 am – 12:10 pm, Gold Room:*

### Data Mining for the Development of Public Policy

Greg Ridgeway (Rand Corporation), gregr@rand.org

**Abstract**

While data mining is not primarily associated with public policy, there are many public policy applications where using data mining methods on massive secondary datasets can contribute substantially. I will describe how data mining methods have been instrumental in designing the Medicare rehabilitation payment system, learning the predictability of high school dropouts, and estimating the effect of drug treatment programs.

## Practical Challenges in the Collection, Processing, and Analysis of Geoscience Data (Invited Session)

Organizer and Chair: Amy Braverman

*Thursday, 10:30 am – 11:05 am, Grey Room:*

### A New High-resolution Blended Real-time Global Sea Surface Temperature Analysis

H. Jean Thiebaux (Dalhousie University, Department of Oceanography), thiebaux@phys.ocean.dal.ca

**Abstract**

A new blended high-resolution real-time global sea surface temperature analysis (RTG_SST), developed specifically for use in operational numerical weather forecasting models, was implemented in NCEP's operational job stream on 30 January 2001. Each daily analysis uses the most recent 24-hours receipts of in situ and satellite-derived surface temperature data and provides a global SST field on a $0.5^\circ \times 0.5^\circ$ (latitude, longitude) grid. The RTG_SST provides the sea surface temperature fields for the regional Meso Eta Model, replacing the previously used NESDIS 50 km satellite-only SST analysis.

Here we describe the development and implementation of the RTG_SST; compare its properties with those of the Reynolds-Smith (RS) analysis (1994) and the NESDIS sea surface temperature analysis.

*Thursday, 11:05 am – 11:40 am, Grey Room:*

### "Volume Closure": A Next Step in Validating Our Interpretation of Atmospheric Observations by Satellites?

Ralph Kahn (Jet Propulsion Laboratory), Ralph.Kahn@jpl.nasa.gov

**Abstract**

We in the satellite remote sensing community have come up with quite a number of "retrieval algorithms". These are aimed at interpreting the radiances measured by Earth-viewing satellites in term of atmospheric properties, such as the amount and type of particles suspended in the air. Since aerosol retrieval is an under-determined problem, assumptions are routinely made about properties of the underlying surface, scene variability, and aerosol vertical distribution. Prior knowledge is applied to constrain the range of aerosol properties considered by the retrieval.

Quantitatively testing the validity of these algorithms, under a range of natural conditions, is proving to be a major challenge. The most ambitious approaches involve surface stations, ships, and aircraft carrying

an array of direct sampling and remote sensing devices intended to characterize the surface and atmosphere as the satellite flies overhead. These field campaigns typically last about a month, during which fundamental questions about experiment design, spatial and temporal sampling, and the degree to which particular meteorological conditions are representative of other times and the radiances measured by Earth-viewing satellites in term of atmospheric properties, such as the amount and type of particles suspended in the air. Since aerosol retrieval is an under-determined problem, assumptions are routinely made about properties of the underlying surface, scene variability, and aerosol vertical distribution. Prior knowledge is applied to constrain the range of aerosol properties considered by the retrieval.

Quantitatively testing the validity of these algorithms, under a range of natural conditions, is proving to be a major challenge. The most ambitious approaches involve surface stations, ships, and aircraft carrying an array of direct sampling and remote sensing devices intended to characterize the surface and atmosphere as the satellite flies overhead. These field campaigns typically last about a month, during which fundamental questions about experiment design, spatial and temporal sampling, and the degree to which particular meteorological conditions are representative of other times and places, are debated as flight decisions are being made. Practical considerations must be folded in, such as restricted waters and air space, and altitude and flight duration limits. Once the data are collected, the debate continues, focusing on how best to analyze, and how to interpret, the results.

“Column Closure” represents a conceptual advance in satellite aerosol retrievals that developed in the mid 1990s. The idea is to collect and interpret the field measurements with the goal of obtaining characteristic surface and atmospheric properties over a region sampled by the satellite. From these characteristics, the net radiation at the top and bottom of the atmosphere can be calculated. The net radiative flux at the top and bottom of the atmosphere is measured as well, and the comparison between the measured and calculated fluxes is used as an indication of the degree to which the environment has been well-characterized.

For the new generation of satellite instruments, such as the Multi-angle Imaging SpectroRadiometer (MISR) that flies aboard NASA’s Terra satellite, the calibration is so good that spatial variability, even over “uniform” ocean sites several kilometers in size, must be taken into account. In planning for the ACE-Asia and CLAMS field campaigns, which took place in April and July 2001, we extended the Column Closure idea to “Volume Closure”, aimed at characterizing the variability of the surface and atmospheric radiative properties along with the representative values. Experiment designs included high-temporal-sampling with surface and airborne radiometers, and “stacked L” flight trajectories. This talk will describe what we were able to measure, will describe how we are planning to analyze the results, and will ask for thoughts about how we might do better.

*Thursday, 11:40 am – 12:15 pm, Grey Room:*

### Creating a Data Mining Environment for Geosciences Data

Sara Graves (University of Alabama, Huntsville), sgraves@itsc.uah.edu

**Abstract**

In addition to feature extraction and anomaly detection, scientific data mining encompasses the discovery and acquisition of knowledge. A Data Mining Environment is required to provide these capabilities for scientific analysis as well as for hypothesis generation. The many challenges that must be addressed with an approach that integrates information science and geoscience expertise will be highlighted in this presentation.

## Best of the Journal of Computational and Graphical Statistics (Invited Session)

Organizer and Chair: David Scott

*Thursday, 10:30 am – 11:05 am, Blue Room:*

### A Brief History of the Mosaic Display

Michael Friendly (Psychology Department, York University), friendly@yorku.ca

**Abstract**

This paper provides an illustrated history of the visual and conceptual ideas leading to the development of mosaic displays. We trace the origins of the use of rectangles and area to depict data quantities and their relations, of early forms of mosaic displays including sub-divided bar-like charts and various cartograms, to the modern forms used in log-linear analysis and in space-filling tree maps. (Keywords: data visualization, space-filling displays, cartogram, thematic cartography, log-linear models, mosaic matrix, tree map.)

*Thursday, 11:05 am – 11:40 am, Blue Room:*

### Detecting Clusters and Nonlinearity in 3D Dynamic Graphs

John Fox (Department of Sociology, McMaster University), jfox@mcmaster.ca,
Robert Stine, Georges Monette, and Neeru Vohra

**Abstract**

Three-dimensional dynamic scatterplots can reveal certain features of data that cannot be apprehended in marginal two-dimensional displays. Using graduate students as subjects, we sought to establish whether the detection of clusters and nonlinearity in 3D plots varies by easily characterized properties of the data and the design of the display. We found that the probability of detection of clusters increased smoothly with cluster separation, and that, at a fixed level of separation, 'diagonally' displaced clusters were easier to detect than 'horizontally' displaced clusters. Cluster detection appeared to be affected to a smaller extent by the design of the display. Three further experiments addressed the detection of nonlinearity in 3D dynamic scatterplots. Most subjects were able to respond in a reasonable manner to properties of the data, so that the probability of detection of nonlinearity increased with its level, particularly when the signal was strong. As in the experiment on cluster detection, subjects' performance was also affected, though to a lesser extent, by characteristics of the displays; for example, spinning the display horizontally in the regression plane was particularly effective. We discuss the implications of these results for the design of statistical software incorporating dynamic 3D scatterplots.

*Thursday, 11:40 am – 12:15 pm, Blue Room:*

### Collaborative Visualization Environments

Edward J. Wegman (Center for Computational Statistics, George Mason University),
ewegman@galaxy.gmu.edu

**Abstract**

The PlatoCAVE and the MiniCAVE are immersive stereoscopic projection-based virtual reality environments oriented toward group interactions. As such they are particularly suited to collaborative efforts in data analysis and visual data mining. In this paper we discuss design considerations for the construction of these environments including 1 wall versus 4 wall implementations, augmented reality, stereoscopic placement, head tracking, the use of LCD devices, polarized light stereo, voice control, and image synchronization.

## Bioinformatics (Contributed Session)

Chair: Sue Bell

*Thursday, 10:30 am – 10:50 am, Green Room:*

### Correlation of Gene Expression Levels with Clinical Covariates in Microarray Experiments

William Shannon (Washington University in St. Louis School of Medicine), shannon@ilya.wustl.edu,
Mark A. Watson, Arie Perry, and Keith Rich (Washington University in St. Louis School of Medicine)

**Abstract**

Microarrays, or gene chips, are used in medical research to simultaneously measure the quantitative expression of 1,000's of genes in tissue samples. These experiments can be used to contrast tissue samples (e.g., benign versus malignant tumors) to learn how combinations of genes play a role in clinically important measures (e.g., disease progression, diagnosis, outcome). Analyzing microarray data generally starts with a clustering algorithm (e.g., hierarchical cluster analysis, self-organizing maps) to separate genes into subgroups with similar expressions, and then compares the clinical measure of interest across the identified clusters. If the clinical measure, such as tumor grade, is disproportionately present across one or a few clusters, the genes within those clusters may be functionally related to the measure. This "two-stage" analysis is likely to be sub-optimal since the analysis of the gene expressions and clinical measures (i.e., covariates) are performed sequentially. Ideally, a statistician would prefer a single analysis (e.g., correlation or regression) using gene expression and covariate data simultaneously. The Mantel correlation statistic, and its extension to the partial correlation and regression frameworks, allows the simultaneous analysis of expression data and covariates in microarray studies. In this paper, we describe this family of statistics and apply them to a previously published study of seven human oligodendrogliomas (brain tumors) where the expression levels of 1,013 genes and five covariates were analyzed using the "two-stage" approach described above. In the previous analysis, qualitative relationships were found between gene expressions and two of the clinical covariates. In this analysis, the Mantel statistics quantify these relationships, and using permutation tests, provide P values of statistical significance. We also show how the Mantel statistics can be used to rank subsets of genes that are sufficient for tumor grade determination. We propose the use of the Mantel statistics as a valuable contribution to the future of microarray study data analysis.

*Thursday, 10:50 am – 11:10 am, Green Room:*

### Assessing Gene Expression Measurements

Lidia Rejto (Statistics Program, University of Delaware), rejto@udel.edu

**Abstract**

DNA microarrays are capable of providing genome-wide patterns of gene expressions across many different conditions over time. The analysis of these patterns requires detecting whether observed differences in expressions are significant or not. Normalization is the term used to describe the process of removing noise and variability of the data. Current methods are unsatisfactory due to lack of systematic framework that can accommodate such noise, variability, low replication of microarray data, and missing value estimation.

We developed a parametric likelihood framework for microarray data analysis. A Bayesian component is introduced, it is supposed that for each gene there is a time dependent probability of different expression levels under different conditions. In this model the probability of different expression levels for the individual genes can be estimated. We included missing data in the model, using appropriate techniques. We used different estimation techniques in order to validate the parametric likelihood model and the parameters of the model in the actual data set. Furthermore we used bootstrapping to estimate the variances of the estimated parameters. Based on the parametric likelihood model we developed clusters of gene expression levels using principal component analysis, $k$-means clustering, and variance clustering techniques.

*Thursday, 11:10 am – 11:30 am, Green Room:*

### Some Recent Methods in Clustering Microarray Gene Expression Data and Applications

Kim-Anh Do (Department of Biostatistics, University of Texas M.D. Anderson Cancer Center), kim@odin.mdacc.tmc.edu, and
G.J. McLachlan (Department of Mathematics, University of Queensland)

**Abstract**

We discuss the statistical development of two methods of clustering gene microarray data: "Gene shaving" as developed by Hastie et al. (2000) and the mixture-model based clustering program, EMMIX-GENE of McLachlan et al. (2000). We applied these methods to the analysis of some well-known data sets: the colon data of Alon et al. (2000), and the leukemia data of Golub et al. (2000). We also analyzed the NCI 60 data using Gene-shaving alone. A close correspondence is found between the gene clusters found by both methods in the case of the Alon data. We also comment on the two distinct tissue clusterings which are found in this data set, which can be explained by the external classification of tumor/non-tumor tissues and by the protocol change explained in Getz et al (2000). In the Golub data, clusters are found which produce a division of tissues corresponding closely to the external classification for both methods. The pros and cons of both methods will be compared and contrasted in details.

Specifically, we will also demonstrate the software GENECLUST developed at MDACC (Biostatistics) based on the idea of gene-shaving.

*Thursday, 11:30 am – 11:50 am, Green Room:*

### Conditional Maps of Continuous Variables

Blair Christian (Department of Statistics, Rice University), blairc@stat.rice.edu, and
David W. Scott (Department of Statistics, Rice University)

**Abstract**

Chloropleth maps offer a simple approach to the display of cross-correlation of two variables. Variable levels are cross-tabulated in a $2 \times 2$ or $\times 3$ matrix, and a color or cross-hatched plotted in regions of the map. We believe a continuous shading or coloration coupled with continuous conditioning offer advantages to understanding the spatial distribution of not only individual variables, but also their cross-correlation. We describe the algorithm, which is based upon multivariate kernel regression, and its implementation in ArcView. Examples from NCI databases will be displayed.

*Thursday, 11:50 am – 12:10 pm, Green Room:*

**Sequential Bayesian Gage R and R Studies: A Graphical Approach**

David A.T. Andrews (Department of Mathematics, University of Dallas), andrews@cs22.math.udallas.edu

**Abstract**

The outcomes of a measurement study can be modeled with a random-effects model with two random components: one for parts, the other for measurement device. For a measurement procedure to be acceptable, the device-to-device variance and error variance should be small in comparison to the part-to-part variance. Using a Bayesian approach yields an easy sequential analysis to decide on the acceptability of the measurement procedure, but the decision rule for increments (add another part, measuring device, or measurement) is not obvious.

We examine the so-called correlations– the variance of one component over the sum of all three– to give some help A posterior sample (generated using MCMC methods) of these quantities can be plotted on a 2-dimensional triangle. This graphical presentation leads to an easily explained increment rule. We describe this rule and explore its behavior through simulation studies.

*Thursday, 12:10 pm – 12:30 pm, Green Room:*

**Algorithm for Bayesian Inference about an Information Index of Fit for Categorical Data Analysis**

Thomas A. Mazzuchi (The George Washington University),
Ehsan Soofi (School of Business Administration, University of Wisconsin-Milwaukee), soofi@csd.uwm.edu, and
Refik Soyer (The George Washington University)
Joseph J. Retzer (Maritz Marketing Research Inc.)

**Abstract**

Akaike information criteria and its descendants serve the purpose of model comparison only, and do not provide diagnostic about the fit of the model to data. We will present the algorithm for a new approach to inference about the fit of a model in categorical data analysis. This approach combines ideas that are well known in information theoretic statistics (maximum entropy characterization of the model) and Bayesian statistics (Dirichlet prior) and is referred to as Maximum Entropy Dirichlet (MED). The procedure assumes that the data generating distribution is unknown, uses moments to derive a tentative model, and incorporates uncertainty about the model. The MED generates prior and posterior distributions of an information index for assessing the model fit. As byproducts, MED also produces priors and posteriors that map uncertainty about the model parameters and the moments. Applications will be presented.

# Random Graphs for Statistical Pattern Recognition (Invited Session)

Organizer and Chair: Carey Priebe

*Thursday, 1:30 pm – 2:05 pm, Gold Room:*

**Proximity Graphs for Nearest Neighbor Decision Rules: Recent Progress**

Godfried Toussaint (School of Computer Science, McGill University), godfried@cs.mcgill.ca

**Abstract**

In the typical nonparametric approach to statistical pattern recognition, random data (the training set of patterns) are collected and used to design a decision rule. One of the most well known such rules is the $K$-nearest-neighbor decision rule in which an unknown pattern is classified into the majority class among its $K$ nearest neighbors in the training set. Several questions related to this rule have received considerable attention over the years. Such questions include the following. How should a value of $K$ be chosen? Should all $K$ neighbors be equally weighted when used to decide the class of an unknown pattern? How can the rule be made robust to outliers present in the training data? How can the storage of the training set be reduced without degrading the performance of the decision rule? Recently these problems have been tackled by computing geometric proximity graphs of the training set and designing algorithms that work on the graphs to obtain satisfactory solutions. Here we review and discuss recent work in this area.

*Thursday, 2:05 pm – 2:40 pm, Gold Room:*

### Class Cover Variants and Spherical Classifiers: Improved Results for Some Not So Special Cases

Adam H. Cannon (Department of Computer Science, Columbia University), `cannon@cs.columbia.edu`

**Abstract**

The class cover problem has been motivated by problems in facility location and high dimensional classification. Approximation algorithms for class cover were originally derived with the goal of obtaining results that were independent of dimension and inter-point cost function. This goal was achieved and general results have been established. Here we look at some special cases that are frequently relevant in classification. We refine the previous analysis to obtain improved approximation results for these settings.

*Thursday, 2:40 pm – 3:15 pm, Gold Room:*

### Random Walks and Catch Digraphs in Classification

Jason G. DeVinney (Department of Mathematical Sciences, Whiting School of Engineering, Johns Hopkins University), `devinney@mts.jhu.edu`

**Abstract**

We will review the applications of the class cover problem to statistical pattern classification. Applying the class cover problem to classification yields a classifier that is similar in spirit to a reduced nearest neighbor classifier. We also introduce a new data adaptive method for using the class cover problem in classification. We investigate the performance of our classifiers on simulation and experimental data.

## Statistics in Planetary Science (Invited Session)

Organizer and Chair: Amy Braverman

*Thursday, 1:30 pm – 2:05 pm, Grey Room:*

### Opportunities for Statistical Applications in Planetary Science

R. Stephen Saunders (Jet Propulsion Laboratory), `Ronald.S.Saunders@jpl.nasa.gov`

**Abstract**

Planetary science provides many potential applications of statistical techniques. This paper will discuss some of these applications. We identify some example areas where the results have been less successful or untried, as opportunities for collaborative research. As an example, one of the most common applications of statistics in planetary science is in the study of impact cratering. Craters provide an essential link for comparing the geologic histories of the planets. How can we distinguish between a random distribution of impact craters on a sphere and one that has real differences in population density? How do we relate impact crater density to absolute age on planetary surfaces? Answers require statistical understanding of the processes, material properties and the nature or evaluation of random distributions in time and space. Problems also include the nature of the impacting objects and how their distribution evolves with time. Statistical evaluation of observation error is also important. We will discuss problems where new understanding of statistical methods is needed in planetary science.

*Thursday, 2:05 pm – 2:40 pm, Grey Room:*

### Statistical Applications in Pattern Recognition in Planetary Images

Sanjay S. Limaye (Space Science and Engineering Center, University of Wisconsin-Madison), `SanjayL@ssec.wisc.edu`

**Abstract**

The availability of high quality digital images of planets and their atmospheres has revolutionized the planetary science. Through the use of simple statistical tools, the analysis of the data has yielded much new knowledge about our solar system neighbors in the geosciences. The statistical tools applied have ranged from purely cosmetic, e.g. managing the appearance of the images, to quantitative, such as learning something about the scattering properties of the planetary surfaces or atmospheres/clouds, atmospheric circulation and even data compression. The tools generally applied have included histograms, auto and cross correlation, and robust estimation. Recently, the increased computational speed is pointing the way toward space based on-board tools such as Principal Component Analysis for data compression as well as noise reduction in spacecraft observations. Finally, the ability to render densely mapped geographical digital data as images has the potential of learning a lot more about the planets, including the earth.

*Thursday, 2:40 pm – 3:15 pm, Grey Room:*

### Statistics on Venus: Craters and Catastrophes(?)

Steven A. Hauck, II (Department of Terrestrial Magnetism, Carnegie Institution of Washington), `hauck@dtm.ciw.edu`

**Abstract**

The nature of the spatial distribution of impact craters is a key to understanding the geologic history of Venus' surface. Based solely upon impact crater centers (points) and using a variety of tests, including $M$th nearest neighbor analysis, the distribution of craters cannot (with a standard level of confidence of 0.05) be distinguished from one that is completely spatially random (CSR). Within the planetary science community this has led to three interpretations: (1) the distribution of craters is random, and therefore the surface is a single age, implying a global, catastrophic resurfacing event (CRM), (2) the apparently random distribution is due to a unique competition leading to equilibrium between random crater emplacement and removal (ERM), and (3) the assumption that the distribution is indeed CSR is a non-unique interpretation, and other geologically based models may be at least as likely to be representative of the crater distribution. These interpretations and their implications will be discussed within the context of unraveling Venus' geologic evolution. The third idea will be a focus of discussion; supporting evidence includes statistically distinct crater densities for surface units defined with geologic criteria without utilizing variations in crater density. In addition, $M$th nearest neighbor analysis of Monte Carlo simulations of a planet with a range of ages defined by geologic units and their respective relative crater densities indicates that the hypothesis that such models are representative of Venus cannot be rejected either. These results suggest that distributions of impact craters that are indistinguishable from random may also have a subtle structure that belies a planet's surface history. These topics are inextricably linked to the roles that statistics and randomness play in investigations of planetary evolution which will be a theme of discussion.

## Best of the National Security Agency (Invited Session)

Organizer and Chair: William F. Szewczyk

*Thursday, 1:30 pm – 2:05 pm, Blue Room:*

### Multiple Spike Train Data Analysis

Satish Iyengar (Department of Statistics, University of Pittsburgh), `si@stat.pitt.edu`

**Abstract**

A common experimental method in neuroscience involves the recording of the activity of a single neuron. However, studies of the functional connectivity of collections of neurons and their behavior require the simultaneous recording of their activity. Current technology permits such recordings of over a hundred neurons. These recordings yield large data sets that present challenging problems in their analysis and their interpretation in biological terms. In this talk, we describe recent work on various techniques for detecting functional connections between neurons and describing the nature of those connections.

*Thursday, 2:05 pm – 2:40 pm, Blue Room:*

### Exploiting the Waiting Time Paradox: Applications of the Renewal Length Transformation

Mark Brown (The City College, CUNY), `cybergarf@aol.com`

**Abstract**

We consider the transformation which take a distribution $F$ into the distribution, $T(F)$, of the length of an interval covering a fixed point, for the stationary renewal process corresponding to $F$. The fact that $T(F)$ tends to be larger than $F$ is the source of the famous "waiting time paradox". Properties of the transformation, $T$, are derived as well as applications to several areas of probability and statistics.

*Thursday, 2:40 pm – 3:15 pm, Blue Room:*

### From Kernels to Mixtures to Clusters

William F. Szewczyk (National Security Agency), `wfszewc@afterlife.ncsc.mil` and
David W. Scott (Department of Statistics, Rice University)

**Abstract**

Most clustering algorithms are based on the heuristic: lots of points close together are more interesting than points far apart. Different algorithms will define "close together" differently. One reasonable way to find points close together is to search for regions of high density and define those as clusters. In this talk we will show how starting with an nonparametric density estimate one is able to identify clusters by a simple recursive simplification of the estimate.

## Computer Science in Statistics (Contributed Session)

Chair: Andrew Moore

*Thursday, 1:30 pm – 1:50 pm, Green Room:*

### A PostgreSQL/R Architecture for a Cognitive Assessment System

E. James Harner (Department of Statistics, West Virginia University), `jharner@student.stat.wvu.edu`,
Jun Tan, Hengyi Xue, and Lingyi Zheng

**Abstract**

The Department of Statistics has been developing the infrastructure for an Intelligent Distributed Environment for Adaptive Learning (IDEAL; `http://ideal.stat.wvu.edu`) to teach introductory statistics and algebra courses. An advanced cognitive assessment system is being built for IDEAL that derives from Tatsuoka's Q-matrix theory, i.e., cognitive skills will be associated with a large number of questions which the student must answer during the semester. Hierarchical Bayesian models for doing formative assessment will be built on Junker's DINA and NIDA conjunctive discrete cognitive-attribute models for task performance. Student scores will be stored in a PostgreSQL database. The assessment models will be built using R, which will be embedded within PostgreSQL. Cognitive assessment update requests to PostgreSQL will be passed to R for processing and the results will be returned to the database for student and instructor Web-based access.

*Thursday, 1:50 pm – 2:10 pm, Green Room:*

### Computer Systems That Learn: An Empirical Study of Noise on the Performance of Three Classification Methods

James R. Nolan (Siena College), `jnolan@siena.edu`

**Abstract**

Classification learning systems are useful in many domain areas. One problem with the development of these systems is feature noise. Learning from examples classification methods from statistical pattern recognition, machine learning, and connectionist theory are applied to synthetic data sets possessing a known percentage of feature noise. Linear discriminant analysis, the C5.0 tree classification algorithm, and a backpropagation neural network tool are used as representative techniques from these three categories. K-fold cross validation is used to estimate the sensitivity of the true classification accuracy to level of feature noise present in the data sets. Results indicate that the backpropagation neural network outperforms both linear discriminant analysis and C5.0 tree classification when appreciable (10 percent or more of the cases) feature noise is present. These results are confirmed when the same type of empirical analysis is applied to a real-world data set previously analyzed and reported in the statistical and machine learning literature.

*Thursday, 2:10 pm – 2:30 pm, Green Room:*

### Comparing Bayesian Neural Network Models and Multilayer Perceptrons for Survival Analysis

Quoc-Anh Trinh (Laboratoire de Statistique Medicale, Universite Rene Descartes),
`trinh@biomedicale.univ-paris5.fr`,
Thu Hoang (Laboratoire de Statistique Medicale, Universite Rene Descartes), and
Bernard Asselain (Institut Curie)

### Abstract

Neural Networks are powerful tools for nonparametric regression and prediction but may result in overfitting. Parameters shrinkage by penalized likelihood methods is used to overcome this problem. Alternatively the Bayesian approach allows to avoid overfitting. In this paper, we use the two approaches to model censored survival data. We compare the predictive accuracy of the models on simulated data and real data. In an application to metastatic breast cancer data, the Bayesian network models allow a better identification of patients with short term survival than the MLP. Regression tree on the output of the networks helps interpreting the results and allows further comparison between the models.

References:

[1] Faraggi D. and Simon R., "A Neural Network Model for Survival Data", *Statistics in Medicine*, 14, 173-82, 1995.

[2] Neal R. M., Bayesian Learning for Neural Networks. New York, NY: Springer-Verlag, 1996.

[3] Ripley B. D. and Ripley R. M., "Neural Networks as Statistical Methods in Survival Analysis", in R. Dybowski and V. Gant, Artificial Neural Networks: Prospects for Medicine, Cambridge University Press, Cambridge, 2001.

*Thursday, 2:30 pm – 2:50 pm, Green Room:*

## Interaction Graphs for Basket Data

J. Van Horebeek (CIMAT), horebeek@cimat.mx and
J. Navarro-Barrientos (CIMAT)

### Abstract

We introduce a class of Graphical Models that allows us to describe and visualize the interaction structure between binary multivariate characteristics at a finer scale than what one obtains with classical Graphical Models. We characterize the underlying class of probability distributions and propose a procedure for model estimation. An application to the analysis of groups of Basket Data in the context of Web-mining is extensively discussed.

*Thursday, 2:50 pm – 3:10 pm, Green Room:*

## Empirical Spectral Analysis of Random Number Generators

David Zeitler (Western Michigan University),
Siemens Dematic, Joe McKean, and John Kapenga (Western Michigan University)

### Abstract

Computer simulation procedures have become a staple of research and development in many fields including statistics. The generation of random number sequences in critical to these procedures. The validity of research results often depend on the underlying validity of the generator being used. In this work we develop the theory for the Empirical Spectral Test (EST). The EST is a class of tests of spatial uniformity based on a multi-dimensional Fourier transform of the empirical probability density function. The test can be applied to sequences from any random number generator, can be adapted to specific user requirements and has the added advantage that its computational complexity is relatively independent of the number of data points being tested.

*Thursday, 3:10 pm – 3:30 pm, Green Room:*

## On Scale Mixtures of Stable Distributions

Dhaif K. Al-Mutairi (Kuwait University)

### Abstract

This method of Shimizu and Fujikoshi (1997) is modified to obtain an estimate for the distribution of a scale mixture of any distribution so that a reduction in the error term is achieved. The use of the modified method to a scale mixture of stable distribution is given.

# KDD Featured Session: Massive Data Sets (Invited Session)

Organizer: Padhraic Smyth, Chair: Arnold F. Goodman

*Thursday, 3:45 pm – 4:20 pm, Gold Room:*

### Caching Statistics in Spatial Data Structures: Fast Real-time Massive Science Data Analysis

Andrew W. Moore (Department of Computer Science, Carnegie Mellon University), awm@cs.cmu.edu

#### Abstract

This talk is about the algorithmic challenges involved in allowing biologists and astrophysicists to continue using the modeling and inference tools they've been happily applying to megabytes of data, when they start drawing in terabytes of data.

We'll discuss new algorithms and data structures that fall into the class of "cached sufficient statistics". These are summary data structures that live between the statistical algorithm and the database, intercepting the kinds of operations that have the potential to eat up valuable time if they were answered by direct reading of the dataset. Some structures may be familiar (kd-trees and R-trees, for example) while some are new (All-dimensions trees, and the Anchors Hierarchy for high dimensions), but for all structures we introduce new search algorithms operating on the cached structures that have interesting properties which call for further development.

I will give some computer demonstrations showing various classes of accelerations broadly covering kernel density speedups (1000 fold), $k$-means, mixture and hierarchical clustering speedups (1000-10000 fold), anomaly detection (100-fold) and 2-, 3-, 4- and 5-point correlation function computation (100-fold up to about a trillion-fold). If time permits we will also discuss (i) "racing" methods to accelerate expensive model selection operations by early termination of models that have less than delta probability of being more than epsilon better than the best model and (ii) preliminary results of new optimization approach to non-linear regression of image morphology parameters for tens of millions of galaxy images.

In collaboration with: Brigham Anderson, Alex Gray, Dan Pelleg, Mary Soon Lee, Jeff Schneider, Bob Nichol, Andy Connolly (U Pitt), Alex Szalay (JHU), Larry Wasserman, Weng-Keen Wong. Related papers and software download information: www.autonlab.org.

*Thursday, 4:20 pm – 4:55 pm, Gold Room:*

### Empirical Bayes Methods for Massive Transaction Data Sets

William DuMouchel (AT and T Research), dumouchel@research.att.com, and
Daryl Pregibon (AT and T Research)

#### Abstract

This paper considers the framework of the so-called "market basket problem", in which a database of transactions is mined for the occurrence of unusually frequent item sets. In our case, "unusually frequent" involves estimates of the frequency of each item set divided by a baseline frequency computed as if items occurred independently. The focus is on obtaining reliable estimates of this measure of interestingness for all item sets, even item sets with relatively low frequencies. For example, in a medical database of patient histories, unusual item sets including the item "patient death" (or other serious adverse event) might hopefully be flagged with as few as 5 or 10 occurrences of the item set, it being unacceptable to require that item sets occur in as many as 0.1 percent of millions of patient reports before the data mining algorithm detects a signal. Similar considerations apply in fraud detection applications. Thus we abandon the requirement that interesting item sets must contain a relatively large fixed minimal support, and adopt a criterion based on the results of fitting an empirical Bayes model to the item set counts. The model allows us to define a 95 percent Bayesian lower confidence limit for the "interestingness" measure of every item set, whereupon the item sets can be ranked according to their empirical Bayes confidence limits. For item sets of size $J > 2$, we also distinguish between multi-item associations that can be explained by the observed $J(J-1)/2$ pairwise associations, and item sets that are significantly more frequent than their pairwise associations would suggest. Such item sets can uncover complex or synergistic mechanisms generating multi-item associations. This methodology has been applied within the U.S. Food and Drug Administration (FDA) to databases of adverse drug reaction reports and within AT and T to customer international calling histories. We also present graphical techniques for exploring and understanding the modeling results.

*Thursday, 4:55 pm – 5:30 pm, Gold Room:*

**Data Mining for Viral Marketing**

Pedro Domingos (Department of Computer Science and Engineering, University of Washington), `pedrod@cs.washington.edu`

**Abstract**

One of the major applications of data mining is in helping companies determine which potential customers to market to. If the expected profit from a customer is greater than the cost of marketing to her, the marketing action for that customer is executed. So far, work in this area has considered only the intrinsic value of the customer (i.e, the expected profit from sales to her). I propose to model also the customer's network value: the expected profit from sales to other customers she may influence to buy, the customers those may influence, and so on recursively. Instead of viewing a market as a set of independent entities, I view it as a social network and model it as a Markov random field. I show the advantages of this approach using a social network mined from a collaborative filtering database. Marketing that exploits the network value of customers – also known as viral marketing – can be extremely effective, but is still a black art. This work can be viewed as a step towards providing a more solid foundation for it, taking advantage of the availability of large relevant databases. (Joint work with Matt Richardson.)

# Geoinformatics (Invited Session)

Organizer and Chair: Sara Graves

*Thursday, 3:45 pm – 4:20 pm, Grey Room:*

**Supervised Learning from Very Large, High Dimensional Remote Sensing Data Sets**

Mark A. Friedl (Department of Geography and Center for Remote Sensing, Boston University), `friedl@bu.edu`, and
Carla Brodley (School of Electrical and Computer Engineering, Purdue University)

**Abstract**

In recent years machine learning and data mining methods have become increasingly common in remote sensing applications. One area in which such techniques are particularly useful is classification of remotely sensed data for land cover and vegetation mapping applications. In this paper we describe the techniques and algorithms being used to map global land cover using data from the Moderate Resolution Imaging Spectroradiometer (MODIS) onboard NASA's Terra spacecraft. Data provided by MODIS include global multispectral observations in seven wavelength bands acquired at 16-day intervals over 12 month periods. The spatial resolution of these data is 1 km. Thus, this classification problem involves very large data volumes with high dimensionality (roughly 100 GB and 180 features). The classification algorithm uses a supervised approach. Training data are provided by a database of over 1000 representative land cover sites that have been compiled from high-resolution satellite data globally. Because of the diversity of global land cover and the complexity of the feature space provided by MODIS, common classification methods do not work well for this problem. To provide robust, repeatable, and accurate maps of land cover at global scales a variety of data mining and machine learning approaches have been utilized. Specifically, we describe techniques to filter training data, include contextual domain information derived from existing land cover maps, and novel uses of ensemble classification methods for this problem domain. Sample results from these algorithms will be presented based on recently available data from MODIS.

*Thursday, 4:20 pm – 4:55 pm, Grey Room:*

**Ecosystem Forecasting**

Joseph Coughlan (Ecosystem Science and Technology, NASA/Ames Research Center), `jcoughlan@mail.arc.nasa.gov`,
Ramakrishna Nemani (School of Forestry, University of Montana), and
Petr Votva (School of Forestry, University of Montana)

**Abstract**

The satellites comprising NASA's Earth Observing System (EOS) Program can collect high quality satellite data to quantify the state and fluxes of the dynamic Earth System. Now those EOS satellites are being launched and have begun to collect data. The challenge is to improve our understanding and produce a predictive capability of the highly integrated Earth System. Most Earth science studies employing satellite data are retrospective analyses, however important predictions that impact decisions and actions must be made in real time. Investments in technology and technology integration are necessary to produce a capability to do on-line analysis of satellite data and make forecasts of important ecosystem conditions (snowpack, runoff, soil moisture and primary plant production) that are useful in resource management. We have developed a data assimilation system, TOPS, (Terrestrial Observation and Prediction System) that integrates satellite data, surface weather observations and weather/climate forecasts with a terrestrial ecosystem model. TOPS produces daily 1 $km^2$ estimates of carbon and water fluxes using MODIS derived leaf area index, land cover and gridded meteorological data created using over 2000 surface weather stations over the conterminous U.S. Daily outputs are expressed as anomalies from historical normals that were computed using 20 years (1982-2001) of satellite and surface weather data.

*Thursday, 4:55 pm – 5:30 pm, Grey Room:*

**Joint Exploration of 3-D Global Atmospheric Models and Related Remote Sensing Data Products with Temporal Displacements of Several Days**

G. D. Emmitt (Simpson Weather Associates, Inc.), gde.swa.com, and
S. Greco (Simpson Weather Associates, Inc.)

### Abstract

In support of proposed new remote sensors of the earth and its atmosphere, simulation experiments are conducted using global circulation models coupled with data obtained from satellite images. The model experiments are known as OSSEs (Observing System Simulation Experiments). These experiments involve long integrations (Nature Runs) of a 3-D numerical model (out to several days or months) which serve as the experimental truth. These files can be quite large, on the order of 4-5 Gbytes/day. In addition to model data, real data extracted from remote sensing products (e.g. multi-spectral images) are used to establish the basis for judging the realism of the Nature Run.

A typical exercise would be to use the Nature Run to simulate data from a new instrument (a wind lidar for example); add those observations to the conventional input data of another global model (the testbed model); and then assess the incremental impact of the new observations on the analyses and forecasts from the testbed model. The realism of the atmosphere is always an issue in the judging the credibility of the OSSE. To evaluate the realism, the remote sensing data can be used to supply statistics or phenomenological details appropriate to the questions being addressed.

Presently, the evaluation of the impacts and the assessment of the realism are limited to very simple inspections. What is really needed is the ability to search both the gridded model fields and the remote sensing data sets for complex patterns and relationships. The target patterns may be fuzzy descriptions of a phenomena (example: find and count all the “cyclones” resolved by the global model). The remote sensing data may be searched for the same phenomena, but with a differing set of fuzzy descriptors which are dependent upon the information content of the data set. A more demanding search is to look both at the forcings in the global model and at time delayed responses in the remote sensing imagery. This type of searching or data mining is currently beyond the capabilities of most researchers and may be a good candidate for a generalized relational search tool.

## Text Mining
## (Invited Session)

Organizer and Chair: Edward J. Wegman

*Thursday, 3:45 pm – 4:30 pm, Blue Room:*

**The Quest for Automated Serendipity**

Anna Tsao (Algotek, Inc.)

### Abstract

Science and engineering knowledge have experienced explosive growth in recent decades due to the fast pace of technological change fueled by revolutions in experimental and computational capability. This growth has made it difficult, if not impossible, for even large research groups or organizations to identify and leverage any but the most immediately accessible developments. Many of the most significant advances in science and technology have been the result of serendipity catalyzed by an insight juxtaposing two ideas or concepts in a novel or unexpected way. In this talk, we introduce a proposed paradigm we shall call Automated Serendipity, which is envisioned to be a high performance information processing system that aids strategic correlation, analysis, and human understanding of available scientific, technical, and engineering knowledge for the purpose of systematically uncovering timely opportunities for technological innovation. Of particular interest is the ability to identify opportunities arising from synergistic developments in disparate disciplines or domains that can be derived from the technical literature. The purpose of this talk is to pose this problem to the community because of both its potential for high impact and to incite interest in meeting the significant mathematical challenges, as part of a concerted, integrated, multidisciplinary research agenda.

*Thursday, 4:30 pm – 5:15 pm, Blue Room:*

**Text Stream Transformation for Semantic-Based Clustering**

Angel R. Martinez (Naval Surface Warfare Center and George Mason University), MartinezAR@NSWC.NAVY.MIL and

Edward J. Wegman (Center for Computational Statistics, George Mason University)

**Abstract**

Assuming a bounded domain of discourse, a transformation of the text stream, called a bigram proximity matrix, is proposed. The proximity matrix is used to classify documents using $K$ nearest neighbor discrimination, and various distances are evaluated. To use other types of classifiers, we reduce the dimensionality of the proximity matrix using the Isomap method. Classification results on the reduced data are also presented. Keywords: semantics, clustering, isomap, bigram, $k$-nearest neighbor.

## Statistical Methods (Contributed Session)

Chair: Godfried Toussaint

*Thursday, 3:45 pm – 4:05 pm, Green Room:*

**Partial Regression**

David W. Scott (Department of Statistics, Rice University), scottdw@stat.rice.edu

**Abstract**

Partial regression refers to the situation of fitting regression models when the “good” data are mixed not only with outliers but also with a separate unknown regression process. A few authors have discussed fitting mixtures of regressions, but assume models are known for each. Here, we describe how a partial mixture estimator, L2E, can be extended to partial regression, providing a technique which is also resistant to outliers.

*Thursday, 4:05 pm – 4:25 pm, Green Room:*

**Variable Subset Selection for Dimension Reduction**

Dennis D. Cox (Department of Statistics, Rice University), dcox@stat.rice.edu
E. Nealy Atkinson (Department of Biomathematics, University of Texas M.D. Anderson Cancer Center),
Iouri Boiko (Biomedical Engineering Center, University of Texas M.D. Anderson Cancer Center),
Calum MacAulay (Department of Cancer Imaging, British Columbia Cancer Research),
Rebecca R. Richards-Kortum (Biomedical Engineering Program, University of Texas), and
Michele Follen (Department of Gynecologic Oncology, University of Texas M.D. Anderson Cancer Center)

**Abstract**

In many modern data sets, the investigator is presented with a large number of variables of potential interest. In such cases it is often desirable to reduce the dimensionality of the problem before undertaking analysis. A dimension reduction permits easier examination of the data for model construction and validation and involves less computational expense. The most common method for dimension, principle component analysis, produces results which may not be easy to interpret, since it constructs new variables which are linear combinations of the original variables. We present a general framework for evaluating methods for dimension reduction. We focus on techniques which select a subset of the original variables rather than constructing new variables. We present several algorithms for dimension reduction and apply them to data sets drawn from biomedical and Earth science applications.

*Thursday, 4:25 pm – 4:45 pm, Green Room:*

### The Gradient Statistic

George Terrell (Department of Statistics, Virginia Polytechnic Institute), terrell@vt.edu

**Abstract**

Large-sample tests of hypotheses about one or more parameters beyond the classical case of the normal linear model usually must resort to normal theory approximations. Wilks (1938) proposed the log-likelihood ratio statistic. Other influential proposals include the Wald (1943) statistic and the Rao (1947) score statistic. All have asymptotically chi-squared distribution. They vary in how convenient a form they take, how readily they may be computed, and how accurate is their asymptotic distribution theory (see e.g. Severini (2000)). We will here propose a remarkably simple alternative test statistic, the gradient statistic, with the same asymptotic distribution. We will establish some properties, then develop a higher-order asymptotic theory in the single-parameter case. Our tool will be a two-point analog to an Edgeworth expansion.

*Thursday, 4:45 pm – 5:05 pm, Green Room:*

### Tree-Based Models for Fitting Stratified Linear Regression Models

William Shannon (Washington University in St. Louis School of Medicine) shannon@ilya.wustl.edu, and Maciej Faifer, Michael A. Province, and D. C. Rao

**Abstract**

This paper generalizes the methods developed in Shannon, Province, and Rao (2001) to use recursive partitioning to identify subsets of the aggregate data within each of which simple linear regression models give better fit. This method is proposed as an alternative to multivariate regression modeling when the analyst is primarily concerned with the regression of an outcome onto a single predictor and needs to control for other covariates. Splitting rules and pruning methods are derived, programmed in C, and linked to the public domain 'RPART' software providing a full software implementation of this methodology. Examples are presented to illustrate the methodology and software.

*Thursday, 5:05 pm – 5:25 pm, Green Room:*

### Bayesian Penalized Splines in Semi-Parametric Modeling

Naomi S. Altman (Department of Statistics, Pennsylvania State University), naomi@stat.psu.edu

**Abstract**

Penalized regression splines provide a useful tool for fitting complicated models with smooth components. Because they are sieve estimators, parametric tools such as likelihood and information criteria can be used for fitting. In this talk, I will demonstrate applications of Bayesian Penalized Splines to self-modeling regression and to varying coefficient models. We will also see that some statistics that are pivotal for the fixed knots (parametric) case do not appear to be pivotal for the sieve estimator. Also, knot placement, which has been shown to be of minimal importance for univariate smoothing, can have a large effect in more complicated settings.

*Thursday, 5:25 pm – 5:45 pm, Green Room:*

### Boosting Multivariate Regression Trees

Stephen R. Sain (Department of Mathematics, University of Colorado at Denver), ssain@math.cudenver.edu

**Abstract**

Boosting was introduced to improve weak classifiers but it has also been shown to be a powerful learning tool applicable to a wide variety of situations and methods, including regression trees. Using a very general mixture approach to motivate splitting rules for regression trees, a gradient boosting algorithm is developed for regression problems with multivariate response vectors. Details of the algorithm will be discussed as well as the results of a brief simulation study. An application based on predicting cholesterol and triglyceride levels will be also presented.

# Spatial Statistics
# (Invited Session)

Organizer: Noel Cressie, Chair: Timothy Haas

*Friday, 8:15 am – 8:50 am, Gold Room:*

### Spatial-temporal Nonlinear Filtering in Command and Control (C2)

Mark E. Irwin (Department of Statistics, The Ohio State University), irwin@stat.ohio-state.edu

#### Abstract

Assimilating data in a highly changing dynamic environment, which allows battle commanders to make timely, informed decisions, is a difficult and challenging problem. In this paper, a spatial-temporal approach to examining the battlespace, based on multiple noisy data signals, is taken. We examine the danger-potential field generated by the positions of an enemy's weapons in the battlespace. The incoming noisy data on the position of weapons is filtered to update the weapons' positions and the danger field. The approach that is taken is that of sequential importance sampling (SIS), or particle filtering, which is used to generate realizations from the posterior distribution of the spatial-temporal danger field. Given realizations of the danger field, non-linear questions such as the locations of maximum and minimum danger, the extent of regions exceeding certain danger thresholds, and changes in the danger field over time, can be addressed. To examine the properties of this SIS approach, data generated from an object-oriented, combat-simulation program is analyzed and compared to a Kalman-filtering approach that depends on just the first two posterior moments.

*Friday, 8:50 am – 9:25 am, Gold Room:*

### Statistical Analysis of fMRI Data

Keith J. Worsley (Department of Mathematics and Statistics, McGill University), keith.worsley@mcgill.ca

#### Abstract

There are some fascinating statistical problems with fMRI data both at the practical and theoretical level. The data itself is a time series at every one of up to 100,000 voxels in 3D space. An external stimulus, such as a cognitive task, is given to he subject during the course of the experiment, so design is an important issue. Since the experiment is repeated over subjects, random effects must be considered. To keep down computation time, only simple models and methods are feasible. The final result is a 3D image of effects, standard errors, and $T$ statistics for a contrast in the stimulus. However the most challenging theoretical problem is how to interpret images of test statistics. Here concepts from differential topology, integral geometry and random fields are important. The Euler characteristic of the (random) excursion set of the $T$ statistic image, first introduced by Robert Adler, plays a vital role in determining which regions of the brain are 'activated' by the stimulus. Recent results of David Siegmund, Jiayang Sun, Satoshi Kuriki, Akemichi Takemura and Jonathan Taylor link this to the volume of tubes and Steiner's formula. Thus brain mapping data is a fascinating blend of mathematics and statistics, with applications beyond neuroscience to other areas such as astrophysics.

*Friday, 9:25 am – 10:00 am, Gold Room:*

### Parametric and Semiparametric Frailty Models for Spatio-Temporally Correlated Multivariate Survival Data

Bradley P. Carlin (Division of Biostatistics, School of Public Health, University of Minnesota), brad@biostat.umn.edu, and
Sudipto Banerjee (Division of Biostatistics, School of Public Health, University of Minnesota)

#### Abstract

Survival models have a long history in the literature (see e.g. Cox and Oakes, 1984), and are enormously popular in the analysis of time-to-event data. Very often these data will be grouped into strata, such as clinical sites, geographic regions, and so on. Such data will often be available over multiple time periods and for multiple diseases. In this paper, we consider hierarchical spatial process models for multivariate survival datasets which are spatio-temporally arranged. Such models must account for correlations between survival rates in neighboring spatial regions, in adjacent time periods, and for similar diseases (say, two different forms of cancer). We investigate both parametric (e.g. Weibull) and semiparametric (e.g. Cox) survival modeling approaches, adding temporal effects in a hierarchical structure. Due to data limitations and computational complexity issues, we avoid geostatistical (kriging) models, and instead handle spatial correlation by placing a particular multivariate generalization of the conditionally autoregressive (MCAR) distribution on the region-specific frailties. Exemplification is provided using time-to-event data for various cancers from the Surveillance, Epidemiology, and End Results (SEER) database.

## Statistics in Earth Science (Invited Session)

Organizer and Chair: Ralph Kahn

*Friday, 8:15 am – 8:50 am, Grey Room:*

### Analysis of Intra-seasonal, Tropical Variability in TRMM Precipitation and Outgoing Longwave Radiation Data

Gerald R. North, (Department of Atmospheric Sciences, Texas A and M University), g-north@tamu.edu, Kenneth P. Bowman, and Hye-Kyung Cho

**Abstract**

Data from the Tropical Rainfall Measuring Mission (TRMM) Microwave Imager and Precipitation Radar are used to study the wavenumber-frequency characteristics of the precipitation and outgoing longwave radiation (OLR) signals in the tropics. At the lowest resolvable frequencies, interannual, annual and semi-annual variability is readily apparent. Not surprisingly, the characteristics of the higher-frequency intra-seasonal waves are strongly affected by the El Nino event that ended in 1998. On intra-seasonal timescales, both eastward and westward moving waves can be detected in both the precipitation and OLR data. The effects of sampling and aliasing on the results are discussed, and the relationship between the wavenumber-frequency spectra of the precipitation and OLR data is described.

*Friday, 8:50 am – 9:25 am, Grey Room:*

### Reconstructing Precipitation Fields from the Historical Record: Think Globally Act Locally

Doug Nychka (Geophysical Statistics Project, National Center for Atmospheric Research), nychka@ucar.edu

**Abstract**

The historical precipitation record for the coterminous US consists of approximately 12,000 stations over 100 years. Although the data is highly irregular in both spatial and temporal sampling, the goal is to create high resolution monthly spatial fields that are useful for forcing ecological models or as inputs for other experiments. In this work we describe statistical methods for generating valid ensembles, or in Bayesian terms, approximate sampling from the posterior distribution. Traditional spatial statistics methods break down when brought to bear on such a large and nonstationary problem. The key to our approach is to efficiently blend sample covariance information from station pairs with large scale parametric models for the precipitation fields. Also the spatial analysis is done locally to reduce the computational order but unlike windowed Kriging we accumulate local results to give a single unified spatial model. One surprising result is that a first infilling missing station data turns out to be a productive intermediate step in computing an ensemble member evaluated on an arbitrary grid.

*Friday, 9:25 am – 10:00 am, Grey Room:*

### Satellite Sampling of Climate

Murry Salby (Department of Atmospheric and Oceanic Sciences, University of Colorado), Murry.Salby@colorado.edu

**Abstract**

Satellite observations provide uniform global coverage of the Earth, unrivaled in studies of climate. Their application, however, requires the gridding or mapping of global structure, which can be used in scientific interpretation and validation of climate models. It is limited by the asynoptic nature of satellite data: Different sites are observed at different times. This feature of satellite data fuses variations in space and time, limiting the scales which can be unambiguously determined. Sampling considerations for asynoptic data are reviewed. They are then applied to address two major issues surrounding their application in studies of climate.

Many of the important issues facing the study of climate involve convection and related properties, like cloud, humidity, and diabatic heating. Mapping the global structure of these properties is complicated by

small-scale and diurnal variations that comprise the global convective pattern. They leave much of the variance undersampled, which is then misrepresented in asynoptic measurements.

The diurnal cycle of convection introduces a bias into time-mean behavior. This source of systematic error has been evaluated in high-resolution Global Cloud Imagery (GCI), which has been composited from 6 satellite platforms simultaneously observing the global convective pattern. Having synoptic coverage of the Earth (all sites observed at the same time), with space-time resolution of 50 km and 3 hrs, the GCI resolves the dominant scales of organized convection. It has been sampled asynoptically according to orbital and viewing geometries of several satellites. Comparing retrieved structure against that actually present in the GCI establishes the bias due to diurnal aliasing. It is significant for all of the orbital geometries, even for a precessing platform that drifts through local time. This source of systematic error is shown to be alleviated with measurements from multiple platforms, which push back the Nyquist limits of asynoptic sampling.

Random variance that is undersampled can be treated through a cancellation of incoherent error. A procedure has been developed to identify small-scale undersampled variance and reject it. This leaves a more accurate representation of large-scale variance that is adequately sampled. Comparing the space-time behavior retrieved against that actually present in the GCI shows that the error variance is reduced to 10representation of large-scale coherent variations, which can then be mapped synoptically on periods as short as 2 days. By recovering the organization of convective properties, this opens the door to a wide range of scientific studies into climate and its interaction with the global circulation.

## Naval Applications of Statistics and Computing (Invited Session)

Organizer and Chair: Wendy Martinez

*Friday, 8:15 am – 8:50 am, Vice Royal 1 Room:*

### Assessing Uncertainty in Numerical Weather Prediction

Adrian Raftery (Department of Statistics, University of Washington), raftery@stat.washington.edu,
Montserrat Fuentes (North Carolina State University),
Tilmann Gneiting (University of Washington), and
Yulia Gel (University of Washington)

**Abstract**

We will briefly review the goals of a new Multidisciplinary University Research Project aimed at developing methods for assessing and communicating uncertainty in numerical weather prediction. Our goals are to develop methods for evaluating the uncertainty of mesoscale meteorological model predictions, and to create methods for the integration and visualization of multisource information derived from model output, observations and expert knowledge. We take several approaches to this, including one based on the recently developed Bayesian melding approach (Poole and Raftery, 2000, JASA). Direct application of Bayesian melding is not feasible in this problem because of the very high dimensionality, and we will outline an alternative based on generating ensembles of initializations from a posterior distribution of the initial state of the atmosphere. The project also aims to develop tools and methods for visualizing predictions of quantities of interest and the uncertainty about them by (i) choosing appropriate quantities of interest for display based on cognitive factors, and (ii) developing appropriate plots, maps, three-dimensional displays, and video displays for decision support. This is joint work with Montserrat Fuentes, Tilmann Gneiting and Yulia Gel.

*Friday, 8:50 am – 9:25 am, Vice Royal 1 Room:*

### Combining Classifiers Based on Kernel Density Estimates and Gaussian Mixtures

Edgar Acuna-Fernandez (Department of Mathematics, University of Puerto Rico, Mayaguez Campus), edgar@math.uprm.edu

**Abstract**

Previous work has shown that combining techniques such as Bagging and Boosting are very effective to reduce the misclassification error of unstable classifiers. This paper reports experimental results of applying both methods to classifiers where the class conditional density is estimated by using a) Kernel density estimation and b) Gaussian mixtures. The experiment was carried out using a collection of machine-learning benchmarks. Keywords: Bagging, Boosting, Kernel density estimates, Gaussian mixtures, classification, ensembles learning.

*Friday, 9:25 am – 10:00 am, Vice Royal 1 Room:*

**Multifractal Modeling of Computer Network Traffic**

Patricia H. Carter (Naval Surface Warfare Center), carterph@nswc.navy.mil

**Abstract**

Network traffic analysis is the study of the flow of packets across a network. Two measurements of the aggregate flow at a choke point are the interarrival process (the times between the successive arrivals of packets) and the packet rate process (the number of packets arriving per unit time). One objective of analysis is to characterize the state of network traffic by modeling the aggregate interarrival or packet rate processes.

The packet rate process is naturally thought of as a (coarse-grained) probability density or measure. Its observed high degree of burstiness across many (time) scales suggests a multifractal model. The multiplicative cascade is one method of generating a multifractal measure. The use of two explicit cascade models, the semi-random multiplicative cascade model (Gilbert, Willinger and Feldman) and the multifractal wavelet model (Riedi, Crouse, Ribeiro and Baraniuk) will be explored. The computation of both analysis and synthesis steps will be described, and the modeling results on a set of live network traffic data analyzed. A new visualization of the results of the semi-random cascade model will be presented.

## Estimation and Prediction (Refereed Session)

Chair: Hugh Chipman

*Friday, 8:15 am – 8:50 am, Vice Royal 2 Room:*

**Approximate Likelihood Procedures for the Boolean Models Using Linear Transects**

John C. Handley (Xerox Corporation), jhandley@crt.xerox.com

**Abstract**

The Boolean model is a random closed set process consisting of a Poisson point process (producing germs) coupled with an independent random shape process (grains). Origins of grains are translated to germs to produce an arrangement of overlapping (or interpentrating) shapes. An accurate discrete approximation to the continuous linear Boolean model offers computationally efficient likelihood procedures including maximum likelihood estimation and likelihood ratio tests. The discrete approximation allows continuous densities as the rate increases. Inference for higher dimensional Boolean models is handled by linear transects. Two two-dimensional estimation examples demonstrate the efficacy of this method.

*Friday, 8:50 am – 9:25 am, Vice Royal 2 Room:*

**A Study on Prediction of Spatial Binomial Probabilities with an Application to Spatial Design**

Hao Zhang (Program in Statistics, Washington State University), zhanghao@wsu.edu, and
H. Holly Wang (Department of Agricultural Economics, Washington State University)

**Abstract**

This work studies some issues that are related to interpolation of spatial binomial probabilities. In some situation, binomial counts are observed at some spatial sampling locations and binomial probabilities are interpolated at un-sampled locations based on the sample data. Such an example in precision farming is considered in this work. A natural practical question is how the number of sampling locations and the sampling sizes at the locations affect the interpolation. This question is studied in this work through simulations. The model-based geostatistics approach (Diggle et al., 1998) is employed in which the binomial counts are modeled through a spatial generalized linear mixed model. The minimum mean-squared error prediction is carried out for the binomial probability.

*Friday, 9:25 am – 10:00 am, Vice Royal 2 Room:*

**Regularized Wavelet Estimation in Partially Linear Models**

Leming Qu (Department of Statistics, Purdue University), quleming@stat.purdue.edu

**Abstract**

The estimates in Partially Linear Models have been studied previously in traditional smoothing methods such as smoothing spline, kernel and piecewise polynomial smoothers. Here, we apply the regularized wavelet estimators by penalizing the L1 norm of the wavelet coefficients of the nonparametric function. The regularization parameter can be chosen by universal threshold or by cross validation. Simulation results show that regularized wavelet approach performs well. But wavelet method makes no strong assumptions about the smoothness of the underlying function of nonparametric part. The computational time is linear.

# Modeling, Computation and Spatial Applications (Invited Session)

Organizer and Chair: Ehsan Soofi

*Friday, 10:30 am – 11:15 am, Gold Room:*

### Spatial Modeling of Atmospheric Pollutants

Richard L. Smith (University of North Carolina, Chapel Hill), rls@email.unc.edu,
Petrutza Caragea (University of North Carolina, Chapel Hill),
Stanislaw Kolenikov (University of North Carolina, Chapel Hill), and
David M. Holland (Environmental Protection Agency)

**Abstract**

The 1990 Clean Air Act Amendments mandated reductions in emissions of a number of atmospherically transported pollutants. One of the responses to that was the setting up of the Clean Air Status and Trends Network (CASTNet), a network of rural stations on which, among other things, concentrations of sulfur dioxide and nitrogen species are measured. It is therefore of considerable interest to monitor long-term trends across this network, but the modeling of temporal trends should also reflect spatial inhomogeneities, requiring spatio-temporal analysis. More recently, in 1997 the Environmental Protection Agency proposed new standard for atmospheric particulate matter, based on the level PM2.5 (airborne particles of diameter 2.5 microns or less), which had not previously been monitored everywhere. Following this a new network of PM2.5 monitors was set up, from which it is of interest to establish where the proposed new standards are violated. The talk will describe ongoing research in the analysis of these pollution networks, including hierarchical models and nonstationary models for the underlying spatial processes.

*Friday, 11:15 am – 12:00 pm, Gold Room:*

### Nonlinear Spatio-Temporal Statistics via Monte Carlo Methods Implemented in a JavaSpaces Distributed Computer

Timothy C. Haas (University of Wisconsin at Milwaukee), haas@uwm.edu

**Abstract**

Coarse grained parallel computing has the potential for allowing a variety of computationally intensive spatio-temporal statistical calculations to be performed by anyone with access to a network of 50 or more PCs. These calculations include robust estimation of nonlinear spatio-temporal trend models, Monte Carlo assessment of model goodness-of-fit and parameter estimate reliability, optimal prediction of a spatio-temporal random field at many locations under asymmetric loss, and the construction of and high speed access to a distributed spatio-temporal variate database. A distributed computer that performs these calculations can be constructed through a transaction space protocol using either JavaSpaces from Sun Microsystems or TSpaces from IBM. As an example, a nonlinear spatio-temporal trend model is estimated with Minimum Distance (a robust statistical parameter estimator) followed by a Monte Carlo computation of parameter estimate standard errors with a JavaSpaces program running on the 50 PCs contained in an instructional computer laboratory during hours that the laboratory is closed.

*Friday, 12:00 pm – 12:15 pm, Gold Room:*

### Floor Discussion

Ehsan Soofi (University of Wisconsin, Milwaukee), esoofi@uwm.edu

# Forecasting the Weather (Invited Session)

Organizer and Chair: Doug Nychka

*Friday, 10:30 am – 11:05 am, Grey Room:*

**The Origin and Modeling of Uncertainty in Numerical Weather Prediction**

Peter Houtekamer (Environment Canada), `Peter.Houtekamer@ec.gc.ca`

### Abstract

For many aspects of numerical weather prediction it is important to have good error statistics. Here one can think of applications such as data assimilation, model improvement and forecasting.

The data-assimilation procedure updates an estimate of the state of the atmosphere with new observations. The current estimate of the atmospheric state has itself been obtained from a short integration with a numerical prediction model of an estimate valid at an earlier time. It is described using of order 10 000 000 coordinates. Every six hours we have of order 100 000 observations of different aspects of the state of the atmosphere. To combine the vector of observations and the state vector we need an estimate of the error covariances of both vectors.

The high dimensionality of the estimation problem, and the significant cost of operations with the prediction model, suggests that Monte-Carlo methods be used to obtain error statistics. This leads to an ensemble of about 100 states that can be used to describe the uncertainty in the short integration that is required by the data-assimilation procedure. A longer integration of the same ensemble can provide the error statistics for the forecast that is issued to the public.

In the Monte-Carlo procedure, one has to sample both the uncertainty in the observations and the weaknesses of the forecast model. If the weaknesses of the model were limited to a few parameter values that are known only with limited precision, these parameter values could themselves be obtained from the data assimilation procedure. However, the forecast model consists typically of 100 000 lines of code and its behavior differs from the atmosphere in ways that are currently only partially understood. This makes it particularly difficult to describe (and sample) the weaknesses of the model.

The approach used to deal with these various issues in the Canadian ensemble prediction system will be presented.

*Friday, 11:05 am – 11:40 am, Grey Room:*

**An Ensemble Kalman Filter for Estimation and Prediction of Severe Thunderstorms**

Chris Snyder (Mesoscale and Microscale Meteorology Division, National Center for Atmospheric Research), `chriss@ucar.edu`

### Abstract

Severe thunderstorms are of great interest not only to fans of the Weather Channel but also to homeowners, businesses and travelers in the path of such storms. Copious, real-time observations of these storms are available in the US from the network of Doppler radars installed in the 1980's. Because the radars provide measurements only of the power of the returned signal (which is related to the amount and size of precipitation and other particles in the beam path) and the component of velocity along the beam, it has proven difficult to use these observations to initialize numerical weather forecasts, where estimates of temperature, pressure, moisture, and all three components of velocity are typically required on a grid of $O(10^6)$ points. I will review recent progress in developing sequential Monte-Carlo methods for state estimation in this problem and discuss several unresolved issues of statistical interest.

*Friday, 11:40 am – 12:15 pm, Grey Room:*

**Particle Filters, Covariance Localization and Numerical Weather Prediction**

Thomas Bengtsson (Geophysical Statistics Project, National Center for Atmospheric Research), `tocke@ucar.edu`

### Abstract

Although forecasting and data assimilation for geophysical processes are common tasks, the application to numerical weather prediction (NWP) poses difficult statistical problems. Characteristic of the NWP problem are nonlinear dynamics and non-Gaussian forecast distributions in high dimensions, making the need for on-line prediction and assimilation particularly challenging. This talk will focus on how Bayesian ideas in nonlinear filtering theory (i.e., particle filters) may be applied in context of NWP. We make use of recent advances in data assimilation research pertaining to localization of covariance structures. Our methods are illustrated using simulations. This is joint work with Doug Nychka and Chris Snyder.

## Computational Statistics in the Army (Invited Session)

Organizer and Chair: Robert Launer

*Friday, 10:30 am – 11:05 am, Vice Royal 1 Room:*

### An Exploitation of Tufte's Small Multiples

Carl T. Russell (Colorado Springs), russellcarl@earthlink.net

**Abstract**

The Theater Missile Defense System Exerciser (TMDSE) is a large geographically-distributed simulation developed by the Missile Defense Agency (MDA). TMDSE is used to investigate Joint Data Network (JDN) interoperability between the members of the theater missile defense Family of Systems (FoS) as they develop. In this paper, SAS and PowerPoint are used to exploit Tufte's notion of small multiples ('graphics can be shrunk way down') for analyzing the communication time delays (communications latencies) inherent to the simulated JDN message traffic. This paper shows how simple statistical graphics implemented on a modern laser printer can produce comprehensive, easily understood characterizations of such communications latencies for 100,000 or more data points.

*Friday, 11:05 am – 11:40 am, Vice Royal 1 Room:*

### Knowledge Based Engineering Assessment of C4SI Digitized Battlefield Experiments

Jock Grynovicki (Human Research and Engineering Directorate, U.S. Army Research Laboratory), grynovi@arl.army.mil

**Abstract**

The U.S. Army Research Laboratory (ARL) has undertaken a research program aimed at better understanding the distributed, non-linear decision-making process at the brigade level and above, as it is shaped by time, stress, team structure, staff experience, the environment, and the introduction of computer technology. Critical battle command tasks were quantified using response data from key Tactical Operations Control (TOC) staff during several Advanced Warfighting Experiments based on ARL structured Knowledge Based System metrics (KBSM). The KBSM is a data collection instrument designed to facilitate recording key battle command data elements through surveys and interviews related to battle command computer assisted tasks made by commanders, staff officers and operators during U.S. Army experiments and exercises. The metrics incorporate recent theories of cognitive science and organizational psychology. Thus, a framework for assessing digital staff performance is established that considers, hardware, battle command functions, soldier operator capabilities as well as staff and leader dynamics. The KBSM was also designed, in part, as the data collection complement to cognitive and knowledge based engineering models of the decision making process, as a major component of the "Cognitive Engineering of the Human-Computer Interface for Army Battle Command Systems (ABCS)" project. The purpose of this paper is to present an approach to data collection, analysis, and classification of critical decision events using multivariate and non-parametric statistical techniques. Key findings from past Advanced Warfighting Experiments and field studies are presented.

Key Words: Knowledge Based Engineering, Army's Battle Command System, performance metrics, behavior anchor scales, soldier system interface, cluster analysis, categorical data.

*Friday, 11:40 am – 12:15 pm, Vice Royal 1 Room:*

### Classification Trees in Army Applications

Barry A. Bodt (U.S. Army Research Laboratory), babodt@arl.army.mil

**Abstract**

The classification tree is one of the more widely recognized tools for uncovering structure in large data sets. As is the case in many arenas, at the U.S. Army Research Laboratory greater use of computing, for example, to assist in data collection from on-line sensors or to generate data directly through simulation has broadened the amount of information gathered. Change is both in terms of the number of cases and the number of potential independent variables to be considered. This presentation focuses on a few recent examples of application where classification trees were helpful, if not essential, in recognizing useful data patterns.

# Computational and Graphical Statistics (Contributed Session)

Chair: Michael Friendly

*Friday, 10:30 am – 10:50 am, Vice Royal 2 Room:*

### Graphical Analysis of High-Dimensional Classifiers

Michael C. Minnotte (Department of Mathematics and Statistics, Utah State University), friendofchaos@yahoo.com
Adele Cutler (Department of Mathematics and Statistics, Utah State University)

#### Abstract

Modern algorithmic classification methods such as trees, forests, and neural networks tend to share two common traits. They can often have far greater predictive power than classical model-based methods. And they are frequently so complex as to make interpretation difficult, often leading to a "black box" appearance. We propose a graphical tool to facilitate investigation of the inner workings of such classifiers. Expansion of the ideas of the data image of Minnotte and West (1999) and the color histogram of Wegman (1990) allows simultaneous examination of dozens to hundreds of variables across similar numbers of observations. Additional information can be visually incorporated as to true class, predicted class, and casewise variable importance. Careful choice of orderings across cases and variables can clearly indicate clusters, irrelevant or redundant variables, and other features of the classifier, leading to substantial improvements in interpretation of both classifier mechanisms and the underlying relationships of class and feature variables in nature.

*Friday, 10:50 am – 11:10 am, Vice Royal 2 Room:*

### Visualization and Analysis of Massive Internet Traffic Data

Don X. Sun (Bell Labs, Lucent Technologies), dxsun@research.bell-labs.com

#### Abstract

Data collection on high speed Internet links can result in large databases. To exploit very large databases we need to do more than computing simple summary statistics; we need to study the data in detail and in its full complexity. To help achieve this we developed S-Net, a traffic measurement and analysis system that begins with packet header collection on network links, and ends with data analysis on a cluster of linux PCs running S, Splus, or R. The system provides tools for the study of packet header data with detailed comprehensive characterization, visualization, statistical modeling and simulation. When used in real-time, the system enables network performance monitoring and anomaly detection for network traffic. Currently, the system has been used at two network locations, one at Bell Labs Research network and the other at the NCNI Research facility as part of the DARPA Helios project. (Joint work with Jin Cao and William S. Cleveland)

*Friday, 11:10 am – 11:30 am, Vice Royal 2 Room:*

### Local Generalized Linear Models as a Framework for Change Detection

Sandra E. Thompson (Pacific Northwest National Laboratory), Sandra.Thompson@pnl.gov,
Don S. Daly (Pacific Northwest National Laboratory), and
Eileen M. Perry (Pacific Northwest National Laboratory)

#### Abstract

Within the context of remote sensing, change detection refers to the characterization of changes in the earth's surface relative to some target or targets of interest. There are many applications using satellite imagery to detect changes in land cover in agriculture, urban planning, and the military. Two major problems with current approaches are 1) the methods are effective in identifying scene changes only if viewing, illumination, atmosphere conditions and other changes not of interest are negligible or corrected for, and 2) the results are typically a binary response, i.e., change or no change. We propose a method which utilizes locally applied linear models that account for image viewing changes, while estimating scene changes of interest through the identification of scene change probabilities at every pixel. In this method we apply generalized linear modeling techniques to sub-regions of images, creating a variety of statistical images that characterize potential landcover or topographical changes over time. By focusing on sub-regions of images, this method is more robust to changes in viewing the scene that other methods might identify as scene changes, and provides greater assessment potential. Extensions to this framework include models that can be formulated to emphasize specific types of changes.

*Friday, 11:30 am – 11:50 am, Vice Royal 2 Room:*

### *k*-Nearest Neighbor and Kernel-Based Credal Classifiers: Theory and Development

Mark Ducey (Department of Natural Resources, University of New Hampshire), `mjducey@cisunix.unh.edu`

#### Abstract

Credal classifiers extend probabilistic classification schemes by employing *credal sets*, or sets of probability measures, to describe imprecision in classification. Credal classifiers have previously been developed as classification trees using discrete or discretized attributes. Here, I extend this family of techniques to continuous classification variables using *k*-nearest neighbor, kernel-based, and weighted *k*-nearest neighbor methods. These methods can tolerate imprecision in both training and validation data, and return classification results having logically defensible precision. A test example, employing hemispherical photography for the description of forest architecture, shows results competitive with a traditional voting *k*-nearest neighbor kernel classifier when the training and validation data are precise, and returns useful results even when the classes in the training set are imprecise. Parameter selection rules based on the Total Uncertainty Measure (TUM) suggest the need for post-classification analysis may be reduced with credal classifiers, but in our examples the TUM indicates parameter values that are too conservative. However, further research is needed on accuracy measures, representation invariance properties, and parameter selection for the IDM itself. Additional pragmatic testing is also required to establish whether, and under what circumstances, credal classifiers offer improvements over more traditional methods.

*Friday, 11:50 am – 12:10 pm, Vice Royal 2 Room:*

### Fast and Stable Bootstrap Methods for Robust Estimates

Matias Salibian-Barrera (School of Mathematics and Statistics, Carlton University), `matias@math.carleton.ca`

#### Abstract

The standard error and sampling distribution of robust estimates can, in principle, be estimated using the bootstrap. However, two problems arise when we want to use bootstrap with robust estimates on moderately large data sets: (a) the bootstrap estimates may be unreliable because the proportion of outliers in the bootstrap samples could be higher than that in the original data set; (b) the high computational demand of robust regression estimates may render the method unfeasible for moderately high-dimensional problems. Problem (a) has been considered by Singh (*Ann. Statist.* 1998) for location estimates. Feasibility considerations related to (b) above have been studied by Schucany and Wang (*J. Roy. Statist. Soc. Ser. B*, 1991) Hu and Kalbfleisch (*Can. J. Statist.*, 2000).

Recently, Salibian-Barrera and Zamar (*Ann. Statist.*, 2002) have proposed a new bootstrap method called "robust bootstrap" to estimate the asymptotic distribution and asymptotic variance of MM-estimates. This method overcomes the problems mentioned above, namely: it is fast and stable (it can resist large proportion of outliers on the bootstrap samples). Unfortunately, its convergence seems to be only of order $O_p(1/\sqrt{n})$.

Another way to estimate the asymptotic variance of robust estimates on large datasets is to bootstrap a one-step Newton-Raphson iteration of their estimating equations. This method will typically be fast (and hence feasible on moderately large datasets). In this paper we compare the performance of this method with that of the "robust bootstrap".

*Friday, 12:10 pm – 12:30 pm, Vice Royal 2 Room:*

### No Need to Talk to Strangers: Cooperation of Interactive Software with R as Moderator

Simon Urbanek (Department of Computer Oriented Statistics and Data Analysis, University of Augsburg), `Simon.Urbanek@math.uni-augsburg.de`

**Abstract**

Interactive software is very helpful for exploratory analysis of data, but most packages are optimized only for a specific field of work. Interaction between the various projects is rather limited. On the other hand very flexible statistical software, such as R, S or S-plus exist covering the needs of almost all fields, but without support for interactive graphics. This paper describes our experience with different approaches linking R with interactive software using flat files and the Omegahat interface. Such linking directly extends capabilities of the interactive software without the need for special communication layers such as CORBA. We also show how such interface can be used to implement cooperation between different interactive software packages with R as a moderator. Practical examples are given based on our current projects Klimt and Mondrian.

# Computer Vision
# (Invited Session)

Organizer and Chair: Yingnian Wu

*Friday, 1:30 pm – 2:05 pm, Gold Room:*

### Towards Prior Models for the Local Geometric Structures in Natural Images

Ann Lee (Department of Computer Science, Brown University), Ann_Lee@Brown.edu

**Abstract**

Recently, there has been a great deal of interest in the statistics of natural images and many investigations of these from both the biological and computational vision perspectives. Despite the many advances in sparse coding and multi-resolution analysis, we are still missing a description of the full probability distribution (as opposed to marginals) of small neighborhoods of pixels or filter responses.

In this talk, I will start by exploring the state space of 3-by-3 high-contrast patches from images of natural scenes. We will see that the distribution of natural data is extremely “sparse” with the majority of data points concentrated along a low-dimensional manifold that correspond to edge structures. Furthermore, I will show evidence from a scale-space study of natural images that the results generalize to general filter responses and larger scales.

*Friday, 2:05 pm – 2:40 pm, Gold Room:*

### Parsing Images into Region and Curve Processes

Zhuowen Tu (Department of Computer Science, The Ohio State University), ztu@cis.ohio-state.edu

**Abstract**

Natural scenes consist of a wide variety of stochastic patterns. Many of these patterns are represented well by statistical models in two dimensional regions as most image segmentation work assume, and some other patterns are fundamentally one or even zero dimensional. We call these three basic categories of patterns as region, curve, and point processes respectively. Image parsing is then referred to the task of parsing (decomposing) natural images into their natural constituents such as region processes, curve processes, point processes, object processes etc. We present a computational paradigm called Data-Driven Markov Chain Monte Carlo (DDMCMC) for parsing images into region and curve processes in the Bayesian statistical framework.

*Friday, 2:40 pm – 3:15 pm, Gold Room:*

### Statistical Modeling and Conceptualization of Visual Patterns

Song Chun Zhu (Department of Computer Science, The Ohio State University), szhu@cis.ohio-state.edu

**Abstract**

The objective of perceptual organization (grouping, segmentation and recognition) is to parse generic natural images into their constituent components which are respectively instances of a wide variety of visual patterns in our visual environment. These visual patterns are fundamentally stochastic processes governed by probabilistic models which ought to be learned from the statistics of natural images. In this paper, we divide existing models into four categories: *descriptive models, causal models, generative models, discriminative models*, and review the objectives, principles, theories, and typical models in each category along the progress in studying natural image statistics. The central theme of this epistemological paper is to study the relationships between the four types of models and to pursue a unified mathematical framework for the conceptualization and modeling of various visual patterns. Indeed, many vision concepts can be rigorously defined only in the context of explicit mathematical models. It is also desirable that under this mathematical framework statistical models for various visual patterns form a “continuous” spectrum – in the sense that they belong to a serial of probability families and are learned under one principle. These statistics models and concepts should amount to a visual language which is essential for building effective, robust, and generic vision systems. There is still a long way to go before such a general visual language can be established, nevertheless, the mathematical framework becomes increasingly clear in recent years. This paper is an attempt to summarize our current understanding of the framework, for the purpose of effective communication between various image search streams.

## Time Series Applications in Seismology (Invited Session)

Organizer and Chair: Edward J. Wegman

*Friday, 1:30 pm – 2:05 pm, Grey Room:*

### Time-Frequency Clustering and Discriminant Analysis

Robert H. Shumway (Department of Statistics, University of California, Davis), rshumway@ucdavis.edu

**Abstract**

A fundamental problem faced in monitoring a potential comprehensive nuclear test ban treaty (CTBT) is that of discriminating between seismic records originating from nuclear explosions and those generated by other seismic events such as earthquakes and mining explosions. In areas where no nuclear testing has occurred, it is also of importance to be able to identify new events of suspicious origin that are substantially different from previously encountered events from that area. Hence, classification and clustering become important approaches to analyzing potential violations of a CTBT.

Most current discriminants depend on measurements of the power spectrum read over specific frequency bands. The inherent non-stationarity of the data is accounted for by extracting spectral components corresponding to primary and secondary arrival phases. Hence, the time-varying spectrum is paramount and this observation leads naturally to considering discriminant and cluster analysis for locally stationary processes. We consider here the application of locally stationary versions of K-L discrimination information measures that give optimal time-frequency measures for measuring the discrepancy between two non-stationary time series. We show that time frequency profiles for earthquakes and nuclear explosions differ in important ways and that the K-L discrepancy measures, integrated over frequency and time, discriminate as well or better than many of the standard measures.

Key Words: Spectral analysis, Kullback-Leibler, Seismology, Nuclear testing.

*Friday, 2:05 pm – 2:40 pm, Grey Room:*

### Application of Regularized Discrimination Analysis to Regional Seismic Event Identification

Dale N. Anderson (Pacific Northwest National Laboratory), Dale.Anderson@pnl.gov

**Abstract**

We present multivariate seismic event identification methods that can be applied to a large number of highly correlated regional discriminants extracted from a seismogram. The methods employ Ridge Discrimination techniques first proposed by Smidt and McDonald (1976). Ridge discrimination was developed to address high-dimensional, co-linear class features. Ridge discrimination was generalized to Regularized Discrimination Analysis (RDA) by Friedman (1989). RDA includes linear and quadratic discrimination in it parameterization. We propose a new approach to optimal selection of RDA parameters by application of the Kullback-Liebler information index.

*Friday, 2:40 pm – 3:15 pm, Grey Room:*

### Eigenpattern Analysis of Geophysical Data Sets– Applications to Southern California

K.F. Tiampo (University of Colorado), kristy@caldera.colorado.edu,
J.B. Rundle, W. Klein, and S. McGinnis

### Abstract

Earthquake fault systems are now thought to be an example of a complex nonlinear system (Bak, 1987; Rundle, 1995). Under the influence of a persistent driving force, the plate motions, interactions among a spatial network of fault segments are mediated by means of a potential that allows stresses to be redistributed to other segments following slip on another segment. The slipping segment can trigger slip at other locations on the fault surface whose stress levels are near the failure threshold as the event begins. In this manner, earthquakes occur that result from the interactions and nonlinear nature of the stress thresholds. This spatial and temporal system complexity translates into a similar complexity in the surface expression of the underlying physics, including deformation and seismicity. Specifically, the southern California fault system demonstrates complex space-time patterns in seismicity that include repetitive events, precursory activity and quiescence, as well as aftershock sequences. Our research suggests that a new pattern dynamic methodology can be used to define a unique, finite set of seismicity patterns for a given fault system (Tiampo et al., 2002). Similar in nature to the empirical orthogonal functions historically employed in the analysis of atmospheric and oceanographic phenomena (Preisendorfer, 1988), the method derives the eigenvalues and eigenstates from the diagonalization of the correlation matrix using a Karhunen-Loeve expansion (Fukunaga, 1990, Rundle, et al., 1999).This Karhunen-Loeve expansion (KLE) technique may be used to help determine the important modes in both time and space for southern California seismicity as well as deformation (GPS) data. These modes potentially include such time dependent signals as plate velocities, viscoelasticity, and seasonal effects. This can be used to better model geophysical signals of interest such as coseismic deformation, viscoelastic effects, and creep. These, in turn, can be used for both model verification in large-scale numerical simulations of southern California and error analysis of remote sensing techniques such as InSar.

## Applications (Refereed Session)

Chair: Adrian Raftery

*Friday, 1:30 pm – 2:05 pm, Vice Royal 1 Room:*

### Microarray Gene Expression Analysis: Data Transformation and Multiple-Comparison Bootstrapping

David R. Bickel (Office of Biostatistics and Bioinformatics, Medical College of Georgia), bickel@mailaps.org

### Abstract

A simple transform function is proposed to preprocess the intensity of gene expression, where the intensity can be that of a colored dye for cDNA microarrays or a gage of probe matching for oligonucleotide arrays. A new measure of skewness is introduced to show that the transform function effectively reduces the asymmetry of intensity values for Affymetrix data of Golub et al. (1999). This transform approaches a logarithmic transform for large intensities, but approaches a linear transform for small intensities, so that the effect of spurious ratios of small intensities is avoided. When the intensity is the average difference (AD) score, the suggested transform function preserves the stochastic nature of AD values rather than resetting negative values to an arbitrary positive value. A conservative estimator of the fold-change based on this transform is proposed. After the B-cell and AML data of Golub et al. (1999) was transformed, a nonparametric bootstrapping method found that the number of genes considered differentially expressed is 48 when controlling the family-wise error rate at the 5 percent level and 572 when controlling the false-discovery rate at the 1 percent level.

*Friday, 2:05 pm – 2:40 pm, Vice Royal 1 Room:*

### Subsampling-based Inference for the Parameters of the Atmospheric Boundary Layer

Alexander Gluhovsky (Purdue University), aglu@purdue.edu, and
Dimitris N. Politis (University of California, San Diego), politis@math.ucsd.edu

### Abstract

Employing the computer-intensive subsampling methodology can considerably improve the statistical validity of atmospheric data analysis. In particular, it makes possible statistically significant comparisons between statistical characteristics of the atmospheric boundary layer computed from observational and large-eddy simulation (LES) data sets. This work illustrates these possibilities by examining the vertical velocity variance estimates from Project LESS (Lake-Effect Snow Studies) aircraft measurements and from LES of the Project LESS event, where subsampling also helps to explain some discrepancies between observations and LES results.

*Friday, 2:40 pm – 3:15 pm, Vice Royal 1 Room:*

### Gaussian Mixture Discriminant Analysis and Sub-Pixel Land Cover Classification in Remote Sensing

Junchang Ju (Department of Geography, Boston University)
Eric D. Kolaczyk (Department of Mathematics and Statistics, Boston University), and
Sucharita Gopal (Department of Geography, Boston University)

**Abstract**

Mixture analysis is a necessary component for capturing sub-pixel heterogeneity in the classification of land cover from remotely sensed images. Mixture analysis approaches in remote sensing vary from simple, conventional linear mixture models to nonlinear neural network mixture models. Linear mixture models are fairly simple and generally result in poor classification accuracy. Neural network models can achieve much higher accuracy, but typically lack interpretability. In this paper we present a mixture discriminant analysis (MDA) model for inferring fractional land covers within forest stands from Landsat Thematic Mapper images. Specifically, individual classes are modeled as mixtures of Gaussian distributions, and classification is performed based on the appropriate posterior distribution. Compared to a benchmark study on mixture models with Plumas National Forest data, this MDA model easily outperforms traditional linear mixture models and parallels the performance of the ARTMAP neural network mixture model. In other words, the MDA model is observed to successfully combine the performance characteristics of more complex neural network models (due to the nonlinear nature of its classification rules), with the ease of interpretation associated with linear mixture models (due to its relatively simple structure). MDA models therefore offer a viable alternative for addressing the mixture modeling problem in remote sensing.

## Spatial Statistics (Contributed Session)

Chair: Bradley Carlin

*Friday, 1:30 pm – 1:50 pm, Vice Royal 2 Room:*

### Estimating Map Accuracy Without a Spatially Representative Training Sample

David Patterson (Department of Mathematical Sciences, University of Montana) `DAPatterson@mso.umt.edu`

**Abstract**

A land cover map is constructed by partitioning a geographic area of interest into a finite set of map units and assigning a land cover class label to each unit. Land cover maps covering millions of acres consisting of millions of units are often constructed from satellite remotely-sensed data. A classification rule is constructed from a training sample of ground-truthed map units. Because of the expense of collecting a spatially representative training sample for such a large map, the training sample is often drawn from a variety of existing data collected for purposes other than mapping land cover. The spatial distribution of the training sample tends therefore to be highly irregular. It is crucial to estimate the accuracy of the resulting map both overall and on a smaller scale since accuracy may vary spatially and by land cover type. Traditional methods of assessing accuracy, such as cross-validation, may be biased because of the spatial irregularity of the training sample if the classification rule uses spatial information. To reduce bias, we suggest methods of estimating overall map accuracy and unit-by-unit accuracy by using calibrated estimates of the posterior probability of correct classification for each map unit.

*Friday, 1:50 pm – 2:10 pm, Vice Royal 2 Room:*

### Ensemble Generation for Large Incomplete Spatial Data Sets

Craig Johns (University of Colorado, Denver), `johns@math.cudenver.edu`

**Abstract**

Geostatistical data sets are frequently incomplete and irregularly spatially distributed. However, many geophysical numerical models, such as those used to predict climate change, require data that are spatially and temporally regular and complete. Computationally efficient methods of interpolation frequently oversmooth the field. Thus, interpolated fields used as inputs for these numerical models lack the variability of the climatological process that they were designed to receive as inputs. Geoscience researchers are recognizing the need to force numerical models with ensembles of inputs, rather than a simple mean field. Recently, an approximate Bayesian framework was used to infill incomplete precipitation records. These methods were shown to be computationally reasonable for large data sets. We show how to use the method to generate ensembles of random observations which are consistent with observed data.

*Friday, 2:10 pm – 2:30 pm, Vice Royal 2 Room:*

**Estimating a Regression Parameter from Spatially Mismatched Data**

Lisa Madsen (Cornell University), `ljm9@cornell.edu`

**Abstract**

Suppose $X(s)$, $Y(s)$, and $\epsilon(s)$ are spatially autocorrelated Gaussian processes that are related as $Y(s) = X(s)\beta + \epsilon(s)$ where $\beta$ is univariate and unknown. Assume the processes are stationary with known covariance structures, and that $E(\epsilon(s)) = 0$ and $E(X(s)) = \mu$ with $\mu$ unknown. The problem is to estimate $\beta$, but the catch is that $Y$ and $X$ are not observed in the same location. This situation may arise when the observed $X$'s and $Y$'s are recorded by different agencies, or when one of the samples is more expensive to obtain. A natural but naïve approach is to predict ("krige") the missing $X$'s at the locations $Y$ is observed, and then use weighted least squares (WLS) regression to estimate $\beta$ as if the predicted $X$'s were actually observed. Call this estimator $\hat{\beta}_{KR}$ (Krige and Regress). Unfortunately, the standard errors obtained by substituting the predicted $X$'s into the variance formula for the WLS estimator of $\beta$ are much to small. A more accurate standard error can be computed using the sandwich formula. An alternative to $\hat{\beta}_{KR}$ is the maximum likelihood estimator $\hat{\beta}_{ML}$. With $\beta$ univariate and $X$ and $Y$ Gaussian, the joint likelihood is easy to maximize with respect to $\beta$. A simulation study suggests that $\hat{\beta}_{ML}$ is much less variable than $\hat{\beta}_{KR}$, performing nearly as well as the WLS estimator of $\beta$ were the missing $X$'s known.

*Friday, 2:30 pm – 2:50 pm, Vice Royal 2 Room:*

**Two-step Algorithm for Spatial Sampling Design**

Zhengyuan Zhu (Department of Statistics, University of Chicago), `zhu@galton.uchicago.edu`

**Abstract**

Gaussian random fields (GRFs) can be used to model many physical processes in space. We study spatial sampling design for prediction of stationary isotropic GRFs with estimated parameters of the covariance function. The key issue is how to incorporate the parameter uncertainty into the design criteria. Several possible design criteria are discussed. A simulated annealing algorithm is used to search for optimal designs of small sample size and a two-step algorithm is proposed for moderately large sample sizes. Simulation results are presented for the Matérn class of covariance functions.

*Friday, 2:50 pm – 3:10 pm, Vice Royal 2 Room:*

**Correction for Interferogram Asymmetry in Hyperspectral Signals**

Mark Fitzgerald (University of Colorado, Denver), `perfabe@email.msn.com`

**Abstract**

The Michelson interferometer can be used to collect an infrared signature over a wide band of wavelengths. Light reflection within the sensor material leads to destructive interference in the signal that is produced. However, the structure of the interference can be estimated, and phase corrections based on these estimates can be implemented to reconstruct the symmetric interferogram.

# Computer Security<br>(Invited Session)

Organizers and Chairs: David Marchette and Jeffrey Solka

*Friday, 3:45 pm – 4:20 pm, Gold Room:*

**Testing Intrusion Detection Systems: Issues Without Answers**

John McHugh (CERT/CC) `jmchugh@cert.org`

### Abstract

In 1998 (and again in 1999), the Lincoln Laboratory of MIT conducted a comparative evaluation of Intrusion Detection Systems (IDSs) developed under DARPA funding. While this evaluation represents a significant and monumental undertaking, there are a number of issues associated with its design and execution that remain questionable. The difficulties associated with this evaluation have been the subject of several papers and a number of presentations. As a result of our investigations of Lincoln's efforts, we have been attempting to develop an appropriate framework in which similar, but meaningful and useful, evaluations can be performed. This talk will contrast our proposed approach with the work that Lincoln performed (and is continuing to perform). Our primary conclusion for signature based systems are that the we simply do not know enough to generate appropriate artificial background data for false alarm evaluation, but that there are a systematic approaches to measuring true positive and negative performance, under both ideal and appropriate environmental stress conditions. The situation is much less clear for with respect to anomaly based systems since the relationships between anomalous and intrusive behavior are poorly understood. In both areas, there is a paucity of theory that can be applied to the problem and we feel that the ad hoc and intuitive approaches that characterize todays efforts may be nearing their limits.

*Friday, 4:20 pm – 4:55 pm, Gold Room:*

## Spatio-temporal Analysis of Internet Routing: Discovery of Global Routing Instabilities Due to Worm Attacks and Other Events

James H. Cowie (Renesys Corporation), and
Andy T. Ogielski (Renesys Corporation), `ato.renesys.com`

### Abstract

Analysis of BGP routing message traffic from many Internet locations provides an accurate spatio-temporal picture of the dynamic state of the global connectivity. BGP (Border Gateway Protocol) is the Internet standard providing a global routing infrastructure.

When a BGP router's best route to a given network address has changed (for better or worse), it sends out route update messages to each neighboring peer router. Therefore, by establishing BGP peering connections with a large number of important BGP routers worldwide, analysis of collected message streams can provide a great deal of information about instantaneous connectivity of the entire Internet, about routing dynamics over a wide range of time scales, and about connectivity failures and "routing storms".

We will present our recent results obtained from real-time multiresolution analysis of live, time-stamped streams of BGP routing messages collected from about 200 routers worldwide, from the multiple monitoring locations of RIPE RIS, RouteViews, and Renesys data collection centers.

We will discuss techniques used to discover the emergence of high-rate, long-lived global routing instabilities, which cause significant, widespread degradation in the end-to-end utility of the global Internet, and we will analyze some of their mechanisms. Case studies of our discovery of unexpectedly large routing instabilities triggered by the spread of Internet worms (Code Red II and Nimda) will illustrate the presentation.

This work leads to new research areas in multiresolution analysis on graphs, statistics of path spaces, and related fields.

*Friday, 4:55 pm – 5:30 pm, Gold Room:*

## Where Are the Nuggets in System Audit Data?

Wenke Lee (Georgia Tech) `wenke@cc.gatech.edu`

### Abstract

Intrusion detection, the process of identifying malicious activities in network and systems, is a very important area of research. Data mining approaches can be applied to network and system audit data to learn normal usage profiles and attack patterns, and to construct intrusion detection models. These (semi-)automated approaches have many advantages over the traditional hand-coding approaches. In this talk, I will first give an overview of current techniques in mining audit data. I will then discuss the research challenges and opportunities in data mining-based intrusion detection.

# Distributed Data Systems for Earth Science (Invited Session)

Organizer and Chair: Menas Kafatos

*Friday, 3:45 pm – 4:20 pm, Grey Room:*

### The ESIP Federation: Building a Distributed DAAC-to-Desktop Data and Information Stream

Howard Burrows (Formerly of the ESIP Federation)

#### Abstract

This talk will cover the history, vision, accomplishments and major current plans of the Federation of Earth Science Information Partners (ESIP Federation), a community-based earth-data and information association serving earth scientists and the larger earth data user community. Sponsored by NASA on the recommendation of the National Research Council, the ESIP Federation has been building the distributed middleware and data service capacities that are needed to realize the value of the latest generation of in-situ and remote-sensing instruments. The ESIP Federation is a democratic, self-governing, public-service organization with partners from government, university, and private industry.

*Friday, 4:20 pm – 4:55 pm, Grey Room:*

### Distributed Web Mapping Tools for Resource Managers

Stanley A. Morain (Earth Data Analysis Center, University of New Mexico), `smorain@edac.unm.edu`

#### Abstract

A web-based Resource Information Management System has been developed by the Earth Data Analysis Center to provide geospatial data and products to resource managers. The system is a web mapping service that is compliant with Open GIS Consortium WMS specifications, providing online data visualization and mapping capabilities. Users can access toolkits and distributed geospatial vector and raster data and products that focus on range management, hydrology, and transportation safety hazards applications. Toolkits that are customized for specific applications are possible via subscription or contract arrangements.

*Friday, 4:55 pm – 5:30 pm, Grey Room:*

### TRMM Rain Rate Statistics

Long S. Chiu (Center for Earth Observing and Space Research, School of Computation Sciences, George Mason University and Goddard Earth Science Enterprise Distributed Active Archive Center, NASA/Goddard Space Flight Center), `lchiu@gmu.edu`

#### Abstract

The complex temporal and spatial structure of rainfall poses major challenges to operational weather forecast and hydrology. The Tropical Rainfall Measuring Mission (TRMM) is a joint U.S.-Japan satellite mission to monitor tropical and subtropical (40 S - 40 N) precipitation and to estimate its associated latent heating. The TRMM satellite carries the first space-borne precipitation radar (PR), a microwave imager (TMI), and a visible Infrared Scanner and provides the first detailed dataset on the four dimensional distribution of rainfall over vastly undersampled tropical and subtropical oceans and continents. TRMM data are processed by the TRMM Science Data and Information System (TSDIS) and archived and distributed by the NASA Goddard Earth Science Enterprise Distributed Active Archive Center (DAAC). To facilitate access and user interactions, GDAAC develops Remote Sensing Information Partners (RSIP) that act as secondary data distribution sites. The first RSIP is the Earth Data Analysis Center (EDAC) at the University of New Mexico.

In this presentation, we will describe the data and access tools for TRMM. Rain rate statistics derived from the TRMM algorithms will be presented, with emphasis for New Mexico. In collaboration with EDAC RSIP, we computed rain rate statistics such as mean rain rate, rain frequency and rain probability over 0.5 × 0.5 degree grids and examined diurnal, seasonal and inter-annual variability over New Mexico based on three years of TRMM data. These satellite estimates are compared with the gage data collected by a network of rain gages maintained by EDAC and with the monthly 0.5x0.5 degree gage analyses produced by Willmott and Matsuura of the University of Delaware.

The boost of TRMM satellite to a higher altitude extends the mission life to 2007. This ten-year data set will advance our knowledge of the space and temporal structure of rainfall.

# Best of the International Association for Statistical Computing (Invited Session)

Organizer and Chair: Anthony Unwin

*Friday, 3:45 pm – 4:20 pm, Vice Royal 1 Room:*

**Analyzing High Dimensional On-Line Monitoring Data**

Ursula Gather (Department of Mathematical Statistics and Industrial Applications, University of Dortmund),
Gather@statistik.uni-dortmund.de

### Abstract

In critical care an abundance of data is generated during the process of care: this data can be stored in real time in clinical information systems which comprise databases with more than 2000 separate patient related variables. Modeling and analysis of the underlying process is a central task that needs to be solved to develop clinical decision support for intensive care medicine. Important requirements for these methods are the ability to handle the data online, the detection of change points, the necessity for the monitoring of the individual patient, good interpretability of results, clinical applications with fast algorithms, as well as the ability to handle multivariate time series.

*Friday, 4:20 pm – 4:55 pm, Vice Royal 1 Room:*

**The PLS approach to Generalised Linear Models and Causal Path Modeling: Algorithms and Applications**

Vincenzo Esposito Vinzi (Department of Mathematics and Statistics, University of Napoli),
vincenzo.espositovinzi@unina.it

### Abstract

PLS (Partial Least Squares or Projection to Latent Structures) methods represent a new generation of statistical procedures. They cover a very broad area of statistical methods, from regression to generalised linear models, from data analysis to causal and path modeling, with several theoretical and statistical properties being demonstrated. The success of PLS methods is widely recognized in chemistry, oil industry, food industry, medicine, biology. These methods are now getting the same level of success in business and industry, especially in the areas of marketing and strategic management. PLS regression (PLS1 for the univariate case and PLS2 for the multivariate one) fits particularly well to situations where classical OLS regression is unstable or not feasible at all (high degree of multicollinearity, small number of observations compared to the number of variables, missing data). There exist several versions of the PLS algorithm within the regression framework. At first, the paper will deal with some extensions of these algorithms to the case of ordinal response variables (PLS logistic regression) as well as to the wider framework of generalised linear models. Then, it will be shown how the PLS principle allows to develop a distribution-free approach to path modeling as an alternative to LISREL for the study of causal relationships. Comparisons between LISRLER and PLS path modeling will be shown with respect to their objectives, statistical properties and performances. In particular, the paper will focus on the algorithmic/computational aspects of the presented methods as well as on the related problems of validation and variables selection within a non parametric context. An outline of the open problems and of the most interesting research perspectives will be also sketched. Examples on real data will be finally shown in order to provide insights on the interpretation rules of the proposed methods.

*Friday, 4:55 pm – 5:30 pm, Vice Royal 1 Room:*

**Models for Three-dimensional Objects**

Adrian Bowman (Department of Statistics, The University of Glasgow), adrian@stats.gla.ac.uk,
Mitchum Bock and Shola Ajayi

### Abstract

Stereophotogrammetry uses pairs of cameras and sophisticated computing algorithms to collect information defining the surface shapes of objects in three dimensions. Surface representation can be in the form of dense point clouds or triangular meshes. Collaborative work among dentists, computer scientists and statisticians at the University of Glasgow is using a system of this type to study the growth of the faces of young children. A quantitative understanding of normal shape and growth is particularly helpful in assessing the effectiveness of surgical repair of cases of cleft lip and palate.

Statistical models for data of this type most naturally begin with landmark data for which a well developed set of tools is available in statistical shape analysis, as described by Dryden and Mardia (1998; Wiley). A much richer representation is expressed in three-dimensional curves, selected to correspond to meaningful anatomical features. Here, functional data analysis, as described by Ramsay and Silverman (1997; Springer-Verlag) provides suitable techniques which can be adapted to the surface anatomy setting. Other aspects of analysis include a local version of the standard technique of Procrustes analysis to assess smooth changes in shape with relevant covariates, including object size. Assessment of symmetry also needs to be considered. All of these aspects will be illustrated on the baby faces study.

# Geoscience (Contributed Session)

Chair: Jean Thiebaux

*Friday, 3:45 pm – 4:05 pm, Vice Royal 2 Room:*

### Relationships Between Land Cover and Spatial Statistical Compression Capabilities in High-Resolution Imagery

James A Shine (George Mason University), jshine@gmu.edu and
Daniel B. Carr (George Mason University)

**Abstract**

Current remote sensing technology offers resolution of 1 meter per pixel or better on satellite or airborne imagery. This results in very large data sets and significant computational challenges in classifying and mapping such imagery. Compression of these data sets without loss of important information continues to be a relevant research issue. Previous work has shown that the use of spatial correlation models can be used to compress multispectral imagery data with a resolution of 1 meter per pixel by two orders of magnitude over an Army fort in central Virginia (Shine, 2001). This paper will extend the study of spatial statistical compression over a variety of 1-meter multispectral imagery, including densely populated urban areas such as New York City. Results will be used to describe relationships between the type of landcover in an image and the ability to compress that image using spatial statistical approaches.

References:

[1] 1. Shine, J.A., "Compression and Analysis of Very Large Imagery Data Sets Using Spatial Statistics", proceedings of the 33rd Symposium on the Interface, Costa Mesa, CA, June 2001.

*Friday, 4:05 pm – 4:25 pm, Vice Royal 2 Room:*

### Multifractals and Resolution Dependence in Remote Sensing: The Example of Ocean Colour

S. Lovejoy (Physics Department and GIROQ, McGill University), lovejoy@physics.mcgill.ca,
D. Schertzer (Laboratoire de Modelisation en Mecanique, U. Pierre et Marie Curie), and
H. Gaonac'h (GEOTOP, Sciences de la Terre, Universite du Quebec a Montreal)

**Abstract**

We argue that geophysical and geographical fields are generally characterized by wide range scaling implying systematic, strong (power law) resolution dependencies when they are remotely sensed. The corresponding geometric structures are fractal sets, the corresponding fields are multifractals. Mathematically, multifractals are measures which are singular with respect to the standard Lebesgue measures, they therefore are outside the scope of many of the methods of classical geostatistics. Because the resolution of a measurement is generally (due to technical constraints) much larger than the inner scale of the variability/scaling, the observations will be fundamentally observer dependent and hence standard remote sensing algorithms which do not explicitly take this dependence into account will depend on subjective resolution dependent parameters. We argue that on the contrary the resolution dependence must be systematically removed so that scale invariant algorithms independent of the observer can be produced. We illustrate these ideas in various ways with the help of eight channel, 7m resolution remote ocean colour data (from the MIES II sensor) over the St. Lawrence estuary. First we show that the data is indeed multiscaling over nearly four orders of magnitude in scale, and we quantify this using universal multifractal parameters. With the help of conditional multifractal statistics we then show how to use multifractals in various practical ways such as for extrapolating from one resolution to another or from one location to another, or to correcting biases introduced when studying extreme, rare phenomena. We also show how the scaling interrelationship of surrogate and in situ data can be handled using vector multifractals and examine the resolution dependence of principle components in dual wavelength analyses. Finally, we indicated why the standard ocean colour algorithms have hidden resolution dependencies, and we show how they can (at least in principle) be removed.

*Friday, 4:25 pm – 4:45 pm, Vice Royal 2 Room:*

### Nonlinear Regression for Describing Glaciated Valley Profiles

Mark C. Greenwood (Department of Statistics, University of Wyoming), markg@uwyo.edu, and
Neil Humphrey (Department of Geology and Geophysics, University of Wyoming)

#### Abstract

The cross valley profile of bedrock valleys that have been extensively glaciated are generally U-shaped; questions of interest include estimation of the actual shape from the data, ability to discriminate between shapes and comparison of these profiles across valleys. Data for such profiles are obtained from high-resolution digital elevation maps via sampling transects across valleys. Current methodology as in Pattyn and Heule (1998) employs "curve fitting" without the use of any diagnostic measures. We consider improvements which include the use and comparison of various nonlinear regression models; illustrated with data from Alaska and the Himalayas.

*Friday, 4:45 pm – 5:05 pm, Vice Royal 2 Room:*

### Comparing Five Modeling Techniques for Mapping Forest Characteristics in the Interior Western US

Gretchen G. Moisen (US Forest Service), moisen@fs.fed.us, and
Tracey S. Frescino (US Forest Service)

#### Abstract

Broad-scale maps of forest characteristics are needed throughout the United States for a wide variety of forest land management applications. Inexpensive maps can be produced by modeling forest class and structure variables collected in nationwide forest inventories as functions of satellite-based information. But little work has been directed at comparing modeling techniques to determine which tools are best suited to mapping tasks given multiple objectives and logistical constraints. Consequently, five modeling techniques were compared for mapping forest characteristics in the Interior Western United States. The modeling techniques included linear models (LMs), generalized additive models (GAMs), classification and regression trees (CARTs), multivariate adaptive regression splines (MARS), and artificial neural networks (ANNs). Models were built for two discrete and four continuous forest response variables using a variety of satellite-based predictor variables within each of five ecologically different regions. All techniques proved themselves workable in an automated environment. When their potential mapping ability was explored through simulations, tremendous advantages were seen in use of MARS and ANN for prediction over LMs, GAMs, and CART. However, much smaller differences were seen when using real data. In some instances,a simple linear approach worked virtually as well as the more complex models, while small gains were seen using more complex models in other instances. In real data runs, MARS and GAMS performed (marginally) best for prediction of forest characteristics.

*Friday, 5:05 pm – 5:25 pm, Vice Royal 2 Room:*

### Modeling Sea Ice Concentrations With the Biased Voter Model

Theodoro Koulis (Department of Statistics and Actuarial Science, University of Waterloo), tkoulis@math.uwaterloo.ca

#### Abstract

Many researchers now agree that climate change is a real threat to our survival and to our ecosystem. The role of seasonal sea ice formation at the poles is complex and closely linked to the Earth's climate. It is thought that the amount of sea ice can have a significant effect on the energies transferred between the atmosphere and the ocean. Understanding the seasonal sea ice process at the poles is therefore of great interest to scientists. Sea ice concentration data sets derived from Earth orbiting satellites are readily available and contain observations that span decades. This data which is both spatial and temporal in nature can be quite difficult to analyze. The methods of analysis for this type of data can be computationally intensive. We present the biased voter model as a candidate for describing the sea ice process. The model, which is borrowed from biology, is a Markov process on a lattice and can be controlled through two parameters. These parameters give some insight on the term behavior of the process. We will discuss various methods for estimating these parameters. The methods are based on differential equations associated with the biased voter model. It is hoped that these methods will be helpful in analyzing multi-temporal spatial data and to make inferences on global climate change.

*Friday, 5:25 pm – 5:45 pm, Vice Royal 2 Room:*

**A Distributed Computing Approach for Remote Sensing Data**

Gregg M. Petrie (Pacific Northwest National Laboratory), gregg.petrie@pnl.gov, and
G. Fann, E. Jurrus, B. Moon, K. Perrine, C. Dippold, and D. Jones
(Pacific Northwest National Laboratory)

**Abstract**

Processing image data generated by new remote sensing systems can severely tax the computational limits of the classic single processor systems that are normally available to the remote sensing practitioner. Operating on these large data sets with a single computer system sometimes means that simplifying approximations are used that can limit the precision of the final results. For instance in supervised classification it is often necessary to assume a gaussian structure for the data. While this assumption has the advantage of greatly reducing the amount of pixels that must be processed this abstraction can also mask important structures in raw data. Recent work at Pacific Northwest National Laboratory strongly suggests that a distributed network of inexpensive PCs can be designed that is optimal to deal with the type of computationally intensive problems encountered in processing remotely sensed images. The paper will briefly describe the basic approach to efficient parallel programming methodology and an associated implementation for processing large images using the latest parallel scientific computing methods. Under the assumption that this new type of distributed computing will remove computational constraints new image processing algorithms for remote sensing images are now being considered. Specific examples where consideration of the entire training set, instead of abstracting the training set to a few representative parameters, can significantly improve classification algorithms will be presented.

# Graphics for Epidemiology and Health (Invited Session)

Organizer and Chair: Dan Carr

*Saturday, 8:15 am – 8:50 am, Vice Royal 1 and 2 Room:*

**Visualization Tools Used to Explore and Disseminate Cancer Statistics**

B. Sue Bell (National Cancer Institute), sb401e@nih.gov
Linda Williams Pickle (George Mason University), and
Daniel Carr (National Cancer Institute and George Mason University)

**Abstract**

The National Cancer Institute (NCI) is developing a website to improve the dissemination of cancer statistics to the policy makers who plan for and prioritize cancer control efforts. To accomplish this goal, NCI is collaborating with the National Science Foundation's Digital Government, Quality Graphics (dgQG) initiative. NCI is now working to deploy several products of that research. The goal is to communicate complex health statistics in a way that makes them understandable to and useful for a diverse audience that includes state epidemiologists and health journalists. We will present the latest versions of these interactive graphics and discuss usability test results.

Specifically, we will present templates from NCI prototypes that include linked micromap (LM) plots, change point regression plots, bar plots, and conditioned choropleth (CC) maps. LMplots, implemented as a JAVA applet, allow users to interact with their data in various ways including sorting and drilling down. The template allows a policy maker to view simultaneously an outcome statistic for a cancer such as the mortality rate, a risk factor statistic such as the prevalence of a behavior associated with that cancer, and the Healthy People 2010 targets for the outcome and risk factor statistics. A second template presents the results of a JoinPoint regression that summarizes a long-term trend as a series of linear segments. A third template answers the question of how much each major cancer site is contributing to the recent trend in the overall cancer rate using bar plots. Finally, we will demonstrate how dynamically conditioned CCmaps can be used to interactively evaluate areas for targeting cancer control efforts. Using CCMaps, a cancer control

planner can explore relationships among an outcome variable, a risk factor variable, a demographic variable, and the spatial distribution of each variable.

*Saturday, 8:50 am – 9:25 am, Vice Royal 1 and 2 Room:*

### Using Landsat Satellite Imagery for Estimating Agricultural Chemical Exposure in an Epidemiological Study

S.K Maxwell (USGS EROS Data Center), maxwell@usgs.gov,
J.R. Nuckols, L. Small, and M.H. Ward

#### Abstract

Knowledge of the spatial distribution and historical change of land cover and water quality in agricultural regions is essential for exposure assessment studies in human health research. Exposure to agricultural chemicals (pesticides and nitrates) has been associated with increased risk of adverse health affects including cancer and birth defects. Traditional epidemiological methods of collecting exposure data by questionnaires are of limited use because the general population lacks knowledge of pesticides sprayed on fields and chemical contaminants in their water supply. Biological and environmental sampling have limitations because samples may only reflect recent exposures. Therefore alternative methods need to be developed to estimate historical agricultural chemical exposure to human populations.

We propose to use historical Landsat satellite imagery to derive crop type maps that will be combined with historical pesticide chemical use data to determine the likelihood that an individual was exposed to a particular pesticide. Landsat satellite imagery has been collected since 1972 and successfully used to classify general land cover types as well as specific crop types. A major problem with using imagery to derive land cover information is the time and expense using traditional classification methods. Cost-effective classification methods are essential in our case, because rare diseases like cancer must cover large geographical regions (e.g., entire states) and span many years (e.g., 5-10 years) in order to determine statistically significant relationships between environmental exposures and disease occurrence. We propose to develop an automated approach to classification of crop types using Landsat satellite imagery that will be applicable to the U.S. Midwest region. We will present our progress to date in developing an automated classification methodology.

*Saturday, 9:25 am – 10:00 am, Vice Royal 1 and 2 Room:*

### LandScan USA: A High Resolution Population Distribution Model

Budhendra Bhaduri (Oak Ridge National Laboratory), bhaduribl@ornl.gov,
Edward Bright (Oak Ridge National Laboratory), and
Phil Coleman (Oak Ridge National Laboratory)

#### Abstract

High resolution population distribution data is critical to the success of epidemiological research. Oak Ridge National Laboratory, as part of its LandScan global population project for the Department of Defense, has developed a high resolution (1 km cell) population distribution model (LandScan 1998 and 2000) for the entire world. LandScan, the finest global population data ever produced, is the first of its kind to use satellite imagery in population distribution data. As an expansion to global LandScan, ORNL is currently developing very high-resolution (90m cell) population distribution data that includes nighttime (residential) as well as daytime distributions. The potential benefit of LandScan USA has been demonstrated for 29 counties covering coastal Texas and Louisiana including the Houston metropolitan area. The LandScan model uses Census data in combination with many other geospatial data, such as land use/cover, topography, slope, roads and nighttime lights, in order to improve the estimation and prediction of the spatial distribution of residential population. LandScan USA provides timely and more spatially precise population and demographic information (compared to information currently available from the Census) to support geographic analysis anywhere in the United States. For example, Census data is routinely used to estimate the number of people living within specified distances from point sources or within areas contaminated by ambient pollutants. Given the format in which Census data are reported, there often is great uncertainty about exact locations of residents. This is particularly appropriate in suburban and rural areas, where the population density is much lower than urban areas. Because of this uncertainty, there is a great potential to misclassify people with respect to their residential distance from pollution sources, and consequently it becomes challenging to determine if certain sub-populations are actually more likely than others to live closer to polluting facilities. LandScan USA helps to reduce this misclassification error.

# Modern Statistical Computing (Invited Session)

Organizer and Chair: Kerby Shedden

*Saturday, 8:15 am – 8:50 am, Canadian 224 Room:*

**Learning Treed Generalized Linear Models**

Hugh Chipman (University of Waterloo), hachipma@icarus.math.uwaterloo.ca

### Abstract

Tree models can be an effective and interpretable tool for supervised learning problems (i.e., regression and classification). A recent variation on trees is the "treed model", which includes a more sophisticated model in each terminal node of the tree, such as a linear regression. This talk considers generalized linear models as a broader class of terminal node models. Specific examples include binary and Poisson regression. One of the main challenges in this area is effective algorithms for learning these models. This talk will explore a Bayesian approach which offers several advantages, including regularization through careful specification of prior distributions, a stochastic search in the tree space, and the potential to improve predictions by model averaging. Data mining applications in areas such as marketing, insurance, and drug discovery will be discussed. Connections with other methods such as Boosting and Bagging will also be explored.

*Saturday, 8:50 am – 9:25 am, Canadian 224 Room:*

**Shift and Scale Coupling Methods for Perfect Simulation**

Jem Corcoran (Department of Applied Mathematics, University of Colorado) Jem.Corcoran@colorado.edu

### Abstract

The development and application of algorithms that enable "perfect" sampling of the invariant measure of a Markov chain, following work in the seminal 1996 paper of Propp and Wilson, provides an important and ever growing set of tools for simulation approaches to inference. Given the availability of recent Markov chain Monte Carlo methodology, which allows many problems of interest in Bayesian and frequentist settings to be couched in terms of such invariant measures, perfect sampling is of particular interest in a statistical context.

These perfect simulation, or "coupling-from-the-past" algorithms rely on the investigators ability to couple sample paths of a Markov chain. This is often a non-trivial task, and, in the case of a continuous state space, it may depend, for example, on the development of tedious minorization conditions. As an alternative, in this talk we describe and develop a variation on a layered multishift coupler due to Wilson that allows one to obtain potentially common draws from two different distributions. The coupler is based on slicing density functions and we describe a "folding" mechanism as an attractive alternative to the accept/reject step commonly used in slice sampling algorithms. Applications will be given to storage models and model selection.

*Saturday, 9:25 am – 10:00 am, Canadian 224 Room:*

**Face Detection and Synthesis Using Markov Random Field Models**

Sarat C. Dass (Department of Statistics and Probability, Michigan State University) sdass@stt.msu.edu

### Abstract

The spatial distribution of gray level intensities in an image can be naturally modeled using Markov Random Fields (MRFs). We develop and investigate the performance of face detection algorithms derived from MRF considerations. For enhanced detection, the MRF models are defined for every permutation of site indices (pixels) in the image. We find the optimal permutation that provides maximum discriminatory power to identify faces from nonfaces using certain types of metrics. These metrics avoid parameter estimation when finding the optimal permutation from the training data base. A maximum pseudolikelihood criteria is proposed for subsequent estimation of the MRF parameters. We investigate the performance of the estimated MRF models for face detection and synthesis. Some detection and synthesis results based on the first and second order neighborhood systems will be shown.

# Data Visualization
# (Invited Session)

Organizer and Chair: Edward J. Wegman

*Saturday, 10:30 am – 11:05 am, Vice Royal 1 and 2 Room:*

### Using Self-Similar Geometric Structures to Represent Letter-Sequence-Indexed Statistical Summaries from Gene Regulation and Peptide Docking Studies

Daniel B. Carr (George Mason University), dcarr@gmu.edu

#### Abstract

The paper addresses the challenge of representing statistics that are indexed by sequences of letters in a way that has the potential of revealing structure in the space of all combinations. The approach develops coordinate systems based on simple geometric structures: tetrahedrons in the case of 4 nucleotides and icosahedron face centers in the case of 20 amino acids. The paper demonstrates two self-similar coordinate generating mechanisms that help to provide cognitive accessibility, self-similarity at different scales and at the same scale. The coordinate systems directly represent short sequences of say 6 nucleotides or 3 amino acids and extend to longer sequences by connecting points with line segments. Variations can modify the space to produce simpler appearance. New visualization software will illustrate applications to gene regulation and peptide docking studies.

*Saturday, 11:05 am – 11:40 am, Vice Royal 1 and 2 Room:*

### New Applications of Image Grand Tour

Juergen Symanzik (Department of Mathematics and Statistics, Utah State University),
symanzik@sunfs.math.usu.edu,
Edward J. Wegman (Center for Computational Statistics, George Mason University), and
Amy Braverman (Jet Propulsion Laboratory)

#### Abstract

The image grand tour is a method for visualizing multispectral images or multiple registered images. Such images can be described as having a multivariate pixel vector associated with each pixel. Examples are the RGB values of a photograph or the multiple spectral and infrared readings from remote sensing satellites. The image grand tour projects a linear combination of the pixel vectors into one-dimensional space (for each pixel) and then renders these projected values as a gray-scale image. An animation of the projected images is obtained by using different linear combinations of the pixel vectors. While the image grand tour was initially used to highlight mines in a minefield, in this talk we present new applications of the image grand tour such as finding hidden petroglyphs in photographs from sites in state parks in Texas and satellite imagery from the Multi-angle Imaging SpectroRadiometer, one of the instruments aboard NASA's Earth Observing System satellite Terra.

*Saturday, 11:40 am – 12:15 pm, Vice Royal 1 and 2 Room:*

### Visualization of Massive Volumetric Data Sets

Bradley C. Wallet (Chroma, Inc and George Mason University), brad@chroma-corp.com,
Robert Wentland (Chroma Energy), and
Jawad Mokhtar (Chroma Energy)

#### Abstract

Visualization of seismic data for hydrocarbon exploration presents difficult challenges in terms of hardware, algorithms, and human computer interfaces. Data sets are often several gigabytes in size, and the volumetric nature of the data forces the operator to explore the data via successive 2-D slices. These tasks are further complicated in the case of derived data where the data is generally multivariate volumetric data. Furthermore, the increasing use of 4-D seismic and amplitude versus offset (AVO) methods present challenges in the visualization of volumetric time series data.

In this paper, we discuss our work in the area of visualizing seismic data. This work includes the use of multiple methods of false color mapping, interactive color map adjustments, alpha channel mappings, and geobody extraction. We also discuss the possible application of these methods to other fields that utilize volumetric data such as medical imaging and nondestructive inspection.

# Data Mining
# (Refereed Session)

Chair: Greg Ridgeway

*Saturday, 10:30 am – 11:05 am, Canadian 224 Room:*

### Novelty Detection in Mass Spectral Data Using a Support Vector Machine Method

Christopher Tong (Department of Statistics, Purdue University), ctong@purdue.edu and
Vladimir Svetnik (Biometrics Research Department, Merck Research Laboratories),
vladimir_svetnik@merck.com

**Abstract**

Outlying samples are sought in a very high-dimensional data set, a library of mass spectra. Such samples are considered novel from the chemical structure point of view and are identified for further investigation of their potential biological activity. The support vector machine algorithm for domain description (Tax and Duin 1999; Scholkopf et al. 2000, 2001) is used to generate a list of potential outliers. The results are compared to those found by a sequential clustering procedure (Svetnik and Liaw, 2001). The results are quite reasonable.

*Saturday, 11:05 am – 11:40 am, Canadian 224 Room:*

### Active Learning in Discrete Input Spaces

Jeff Schneider (School of Computer Science, Carnegie Mellon University), and
Andrew Moore (School of Computer Science, Carnegie Mellon University)

**Abstract**

Traditional design of experiments (DOE) from the statistics literature focuses on optimizing an output parameter over a space of continuous input parameters. Here we consider DOE, or active learning, for discrete input spaces. A trivial example of this is the $k$-armed bandit problem, which is the case of having a single input attribute of parity $k$. We address the full problem of many attributes where it is impossible to test every combination of attribute-value pairs even once within the given number of experiments, but we expect to be able to generalize on the results of experiments. We further pose the problem of active learning on fixed experiment sets where we can not choose any possible setting of input variables, but instead must choose from a fixed set of available experiments. We discuss discrete DOE and fixed experiments sets in marketing and pharmaceutical domains. We propose several active learning algorithms based on the idea of building a function approximator for the experiments taken so far and using its prediction and confidence intervals to select future experiments. The algorithms are tested using commonly available data sets. We conclude with our ideas for extending these algorithms.

*Saturday, 11:40 am – 12:15 pm, Canadian 224 Room:*

### Efficient Algorithms for Non-Parametric Clustering with Clutter

Weng-Keen Wong (Department of Computer Science,Carnegie Mellon University), wkw@cs.cmu.edu and
Andrew Moore (Department of Computer Science, Carnegie Mellon University), awm@cs.cmu.edu

**Abstract**

Detecting and counting overdensities in data is a common problem in the physical and geographic sciences. One of the most successful of recent algorithms for the counting version of the problem was introduced by Cuevas, Febrero, and Fraiman [Cuevas et al., 2000], which will be referred to as the CFF algorithm.l This algorithm first determines the subset of data points that are in high density regions using a non-parametric density estimator. A clustering step follows where such high density points are agglomerated. While this algorithm was originally intended to estimate the number of clusters, it can also be used to perform non-parametric clustering against a noisy background. However, the algorithm proposed by CFF is too computationally expensive to work on large datasets with greater than two dimensions. We propose an alternative implementation of the CFF algorithm producing exactly the same results but addressing the computational problems in both the density estimation and in the agglomeration step. We will then illustrate the effectiveness of our approach on large multi-dimensional astrophysics datasets.