Big Data in Environmental Science

May 11 to 15, 2015
University of British Columbia

Speakers, titles, abstracts.

1. **Speaker**: James Balamuta, University of Illinois

   Title: Powering Up the Computation for Big Data

   Abstract: *On top of the statistical methods that have been demonstrated to largely reduce the computation for large data sets, the efficient usage of computing resource can further speed up the computation. I will introduce several ways to improve the computing. The improvements will largely focus on effectively utilizing Rs C/C++ API with Rcpp and by applying parallelization techniques afforded to R using explicit parallelism for OpenMP, MPI, and parallel as well as implicit parallelism through a parallelized BLAS. Finally, we illustrate these techniques using two spatial data examples*

2. **Speaker**: Soutir Bandyopadhyay, Lehigh University

   Title: Spatial Methods for Nonstationary Fields Using Compact Basis Functions

   Abstract: *Kriging is a non-parametric regression method used in geostatistics for estimating curves and surfaces and forms the core of most statistical methods for spatial data. In climate science these methods are extremely useful for estimating how climate varies over a geographic region when the observational data is sparse or the computer model runs are limited. A statistical challenge is to implement spatial methods for large sample sizes and also account for the heterogeneity in the physical fields, both common features of many geophysical problems. Equally important is to provide companion measures of uncertainty so that the estimated surfaces can be compared in an objective way and are suitable for decision making. Here we present a statistical method that expands the spatial field in a large number of basis functions of varying resolution but all with compact support. Parsimonous models for the precision matrix of the basis coefficients are able to approximate standard covariance models but also scale to large numbers of spatial locations.*

3. **Speaker**: Sudipto Banerjee, Professor and Chair, Department of Biostatistics, UCLA Fielding School of Public Health

   Title: On Gaussian Process Models for High-Dimensional Geostatistical Datasets

   Abstract: *With the growing capabilities of Geographical Information Systems (GIS) and user-friendly software, statisticians today routinely encounter geographically referenced datasets containing observations from a large number of spatial locations. Over the last decade, hierarchical spatial process models have become widely deployed statistical tools for researchers to better understanding the complex nature of spatial variability. However, fitting hierarchical spatial models often involves expensive matrix decompositions whose computational complexity increases in cubic order with the number of spatial locations. This renders such models infeasible for large spatial data sets. In this talk, I will present two approaches for constructing well-defined spatial stochastic processes that accrue substantial computational savings. Both these processes can be used as "priors" for spatial random fields. The first approach constructs a low-rank process operating on a lower-dimensional subspace. The second approach constructs a Nearest-Neighbor Gaussian Process (NNGP) that can be exploited as a dimension-reducing prior embedded within a rich and flexible hierarchical modeling framework to deliver exact Bayesian inference. Both these approaches lead to Markov chain Monte Carlo algorithms with floating point operations (flops) that are linear in the number of spatial locations (per iteration). We compare these*

methods and demonstrate its use in inferring on the spatial distribution of forest biomass from the US Forest Inventory database spanning the continental US.

Joint work with Abhirup Datta, Andrew O. Finley and Alan E. Gelfand.

4. **Speaker**: Luke Bornn, Simon Fraser University

   Title: Lost Moments: The Effect of Pre-processing on Environmental Data.

   Abstract: *Monitoring networks for environmental and meteorological data are typically vast and irregularly spaced with significant missingness across space and time. As a result, this data is often aggregated both spatially and temporally to ease analysis. In this talk, I will explore the biases introduced and information lost when using this aggregated data, and propose methods for mitigating these impacts.*

5. **Speaker**: Jonathan Bradley, University of Missouri

   Title: Efficient Parameterizations for Multiscale Multivariate Spatio-Temporal Data

   Abstract: *Many data sources report related variables of interest that are also referenced over multiple geographic regions and time; however, there are relatively few general statistical methods that one can readily use that incorporate dependencies over different spatial locations, spatial scales, time points, and variables. Additionally, many multivariate spatio-temporal areal datasets are extremely high-dimensional, which leads to practical issues when formulating statistical models. We use the multivariate spatio-temporal mixed effects model (MSTM) in a fully Bayesian framework to analyze data of this type. Moreover, we introduce the use of Obled and Creutin eigenfunctions within this framework to allow for multivariate spatio-temporal data observed on different spatial scales. We illustrate our method through an empirically motivated simulation study based on Quarterly Workforce Indicators (QWI) published by the US Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program. We also provide a demonstration of our approach using an environmental dataset.*

6. **Speaker**: Mike Brauer, University of British Columbia

   Title: Coming soon.

   Abstract: *Coming soon.*

7. **Speaker**: Jenny Bryan, University of British Columbia

   Title: Visualization with ggplot2

   Abstract: *I will give a brief overview of some recent, exciting additions to the toolkit for statisticians and data analysts who work in R. Examples include the RStudio IDE and the dplyr package for data manipulation. Then we will go directly into a hands-on tutorial on ggplot2, a package written by Hadley Wickham, implementing the Grammar of Graphics from Leland Wilkinson.*

   Please bring your laptop! Make sure you are running a fairly recent version of R (R version 3.2.0 was released 2015-04-16). Install the ggplot2 and gapminder packages (run install.packages(c("ggplot2", "gapminder"))). If you haven't already, this might be a great time to install and try out the RStudio IDE (http://www.rstudio.com/products/rstudio/download/).

8. **Speaker**: Peter Craigmile, The Ohio State University

   Title: Heteroscedastic asymmetric spatial processes (HASP)

   Abstract: *The Gaussian stochastic process is commonly used for modeling time series and geostatistical data. The Gaussianity assumption, however, is known to be insufficient or inappropriate in many settings. In this talk, I discuss the development of specific non-Gaussian models to capture the asymmetry and heavy tails of many real-world datasets indexed in space and time.*

   Introducing a general framework for constructing non-Gaussian spatial processes using transformations of a latent multivariate Gaussian process, we develop a heteroscedastic asymmetric spatial process (HASP) for capturing the non-Gaussian features of environmental or climatic data, such as the heavy tails and skewness. The conditions for this non-Gaussian spatial process to be well defined is discussed at length. The properties of the HASP, especially its marginal moments and covariance structure, are established along with a Markov chain Monte Carlo (MCMC) procedure for sampling from the posterior distribution.

   The HASP model is applied to the study of a US nitrogen dioxide concentration dataset, demonstrating the ability of HASP to capture asymmetry and heavy tails benefits its predictive performance.

   This research project is joint with Jiangyong (Matt) Yin, Ph.D., Google Inc.

9. **Speaker**: Marco Ferreira, Virginia Tech

   Title: Dynamic Multiscale Spatiotemporal Models for Gaussian and Poisson Processes

   Abstract: *We discuss classes of dynamic multiscale models for Poisson and Gaussian spatiotemporal processes. Specifically, we use multiscale spatial factorizations to decompose the process at each time point into spatiotemporal multiscale coefficients. We then connect these spatiotemporal multiscale coefficients through time with state-space-like evolutions. In the case of Gaussian processes we use a Gaussian state-space evolution, whereas in the case of Poisson processes we use a novel Dirichlet evolution. Further, we propose simulation-based full Bayesian posterior analysis. In particular, for both Gaussian and Poisson processes we develop filtering equations for updating of information forward in time and smoothing equations for integration of information backward in time, and use these equations to develop forward filter backward samplers for the spatiotemporal multiscale coefficients. Because the multiscale coefficients are conditionally independent a posteriori, our full Bayesian posterior analysis is scalable, computationally efficient, and highly parallelizable. We present results on the spatial and spatiotemporal dependence structure. Finally, we illustrate the usefulness of our multiscale spatiotemporal methodology with two applications. The first application examines mortality ratios in the state of Missouri, and the second application considers tornado reports in the American Midwest.*

10. **Speaker**: Dorit Hammerling, IMAGE NCAR

    Title: A new ensemble-based consistency test for the Community Earth System Model

    Abstract: *Climate simulations codes, such as the National Center for Atmospheric Researchs Community Earth System Model (CESM), are especially complex and continually evolving. Their on-going state of development and installation on new systems require frequent software verification in the form of quality assurance to both preserve the quality of the code and to instill confidence in the model in situations where bit-to-bit reproducibility is not feasible. To formalize and simplify this previously subjective and computationally-expensive verification process, we have developed a new tool for evaluating climate consistency. Because an ensemble of simulations allows us to gauge the natural variability of the models climate, our new tool uses an ensemble approach for consistency testing. In particular, an*

ensemble of CESM climate runs is created, from which we obtain a statistical distribution that can be used to determine whether a new climate run is statistically distinguishable from the original ensemble. The CESM Ensemble Consistency Test, referred to as CESM-ECT, is objective in nature and accessible to CESM developers and users. The tool has proven its utility in detecting errors in software and hardware environments and in providing rapid feedback to model developers. We will also discuss ongoing research to extend the current tool. This is joint work with Allison Baker and Daniel Milroy.

11. **Speaker**: Mevin Hooten, Colorado State University

    Title: Fringe benefits: The hidden utility of constraints in telemetry studies

    Abstract: *Advances in animal telemetry data collection techniques have been a catalyst for the creation of statistical methodology for analyzing animal movement data. While the technology for data collection is improving dramatically over time, we are left with massive archives of historical animal telemetry data that are subject to measurement error (i.e., location uncertainty). This form of error arises as a combination of factors due to accuracy of the telemetry device and system, animal behavior, atmospheric interference, and landscape features. Furthermore, the measurement error varies with both location and time and the information available about the accuracy is not easily incorporated into statistical models and is often in flux due to ongoing manufacturer findings. Thus, there is a need for 1.) approaches to better estimate the telemetry error distribution and 2.) improved methods to incorporate it into animal movement models. Using both simulations and real data, we describe how certain forms of auxiliary information can be surprisingly useful for learning about telemetry error. We show how hierarchical models, with appropriate constraints on the data and process components, can increase the utility of common forms of telemetry data for understanding animal space use and resource selection.*

12. **Speaker**: Timothy Johnson, University of Michigan

    Title: Analysis of Point Pattern Imaging Data using Log Gaussian Cox Processes with Spatially Varying Coefficients

    Abstract: *Log Gaussian Cox Processes (LGCP) are used extensively to model point pattern data. In these models, the log intensity function is modeled semi-parametrically as a linear combination of spatially varying covariates with scalar coefficients plus a Gaussian process that models the random spatial variation. Almost exclusively, the point pattern data are a single realization from the driving point process. In contrast, our motivating data are lesion locations from a cohort of Multiple Sclerosis patients with patient specific covariates measuring disease severity. Patient specific covariates enter the model as a linear combination with spatially varying coefficients. Our goal is to correlate disease severity with lesion location within the brain. Estimation of the LGCP intensity function is typically performed in the Bayesian framework using the Metropolis adjusted Langevin algorithm (MALA) and, more recently, Riemannian manifold Hamiltonian Monte Carlo (RMHMC). Due to the extremely large size of our problem — 3D data ($64 \times 64 \times 64$) on 240 subjects — we show that MALA performs poorly in terms of posterior sampling and that RMHMC is computationally intractable. As a compromise between these two extremes, we show that posterior estimation via Hamiltonian Monte Carlo performs exceptionally well in terms of speed of convergence and Markov chain mixing properties. To speed up posterior estimation, critical parts of the HMC algorithm are ported to a GPU and run in parallel. Further, we compare the statistical efficiency of the fully Bayesian approach with two deterministic approximations: variation Bayes and integrated nested-Laplacian approximation (INLA) in a simple LGCP model.*

13. **Speaker**: Matthias Katzfuß, Texas A & M

    Title: A multi-resolution approximation for big spatial data

Abstract: *Automated sensing instruments on satellites and aircraft have enabled the collection of big spatial data over large and inhomogenous spatial domains. If these kinds of datasets can be efficiently exploited, they can provide new insights on a wide variety of issues. However, traditional spatial statistical techniques such as kriging are not computationally feasible for big datasets. We propose a multi-resolution approximation (M-RA) of Gaussian processes observed at irregular (i.e., non-gridded) locations in space. The M-RA process is specified as a linear combination of basis functions at multiple levels of spatial resolution, which can capture inhomogenous spatial structure from very fine to very large scales. The basis functions are chosen to optimally approximate a given covariance function, and no restrictions on the covariance function are necessary. All computations involving the M-RA, including fully Bayesian parameter inference and prediction, are highly scalable for massive datasets. Crucially, the inference algorithms can also be parallelized to take full advantage of distributed computing environments.*

14. **Speaker**: Will Kleiber, University of Colorado

    Title: Coherence for Random Fields

Abstract: *Multivariate spatial field data are increasingly common and whose modeling typically relies on building cross-covariance functions to describe cross-process relationships. An alternative viewpoint is to model the matrix of spectral measures. We develop the notions of coherence, phase and gain for multidimensional stationary processes. Coherence, as a function of frequency, can be seen to be a measure of linear relationship between two spatial processes at that frequency band. We use the coherence function to give natural interpretations to cross-covariance parameters of the Matérn class, where the smoothness indexes dependence at low frequencies while the range parameter can imply dependence at low or high frequencies. Estimation follows from smoothed multivariate periodogram matrices. We illustrate the estimation and interpretation of these functions on two datasets, forecast and reanalysis sea level pressure and geopotential heights over the equatorial region. Examining these functions lends insight that would otherwise be difficult to detect using standard cross-covariance formulations.*

15. **Speaker**: Bo Li, University of Illinois

    Title: Evaluating Climate Field Reconstructions in Reduced Dimension

Abstract:

The large scale climate field reconstructions (CFRs) of the common era target hemispheric or global patterns of temperature change. They provide important spatial information that in turn can be used to infer dynamical insights about the causes of past climate variability and change. An important tool for assessing the performance of CFRs is the pseudoproxy experiments, which are controlled and systematic experiments based on millennium-length, forced transient simulations with fully coupled general circulation models. All CFRs, modeled climate and real climate are high dimension and correlated random processes. Evaluating the difference between CFRs as well as their performance relative to their target climate fields presents statistical challenges. We propose to assess the first and second moment structures and the trend of climate fields in a reduced dimension via functional data analysis approach. The evaluation is through a sequence of formal hypothesis tests that are constructed based on the functional principal components or dominant modes of variation inherent in the data. Our method well respects the properties of the climate fields and is robust by being nonparametric.

16. **Speaker**: Finn Lindgren, University of Bath.

    Timing Constraint: Monday afternoon & Tuesday Afternoon

Title: Stochastic partial differential equations and numerical methods for large scale spatial statistics

Title: Towards realistic stochastic modelling of global temperatures

17. **Speaker**: Robert Lund, Clemson University

    Title: Changepoints and Associated Climate Controversies

    Abstract: *This talks overviews changepoint issues in climatology. Changepoints (inhomogeneities) are ubiquitous features in climatic time series, arising, for example, when stations relocate or instrumentation is changed. Changepoints confound many inference problems and are very important data features. Here, we show why changepoint information is essential in making accurate trend conclusions. Examples are given where inferences are questionable when changepoints are ignored. The talk will delve into two recent contentious climate issues: 1) the recent increase in Atlantic Basin hurricanes; and 2) the "warming hole" (lack of warming) seen in the Eastern United States.*

18. **Speaker**: Renjun Ma, University of New Brunswick

    Title: Spatiotemporal Analysis of Environmental Health Risk

    Abstract: *Big data with complex spatiotemporal structures are common in environmental studies. In order to account for such spatiotemporal structures, spatially and temporally correlated random effects are often incorporated into generalized linear models for such data. The estimation of these models often poses theoretical and computational challenges. We propose an orthodox best linear unbiased predictor (BLUP) approach to these models. Our approach is illustrated with application to Ohio lung cancer data where the annual lung cancer deaths for 88 counties were obtained from 1968-1988. With estimated spatial and temporal random effects, we will also discuss the identification of high/low risk areas, spatial clustering as well as temporal trend. An adaptation of this approach to analysing spatially correlated survival data will also be illustrated with application to American Cancer Society study of air pollution and mortality where over half million of subjects were followed every half month since 1982.*

19. **Speaker**: Doug Nychka, NCAR

    Title: Extremes in regional climate: What to do with 8000 histograms.

    Abstract: *As attention shifts from broad global summaries of climate change to more specific regional impacts there is a need for the data sciences to quantify the uncertainty in regional predictions. A regional climate model (RCM) is a large and complex computer code based on physics that simulates the detailed flow of the atmosphere in a particular region from the large scale information of a global climate model. Part of the value of these simulations is to explore the potential extremes in weather that are due to natural variation and also to climate change. Here we present an application that combines logspline density estimates to discern tail behavior in a distribution with spatial methods for large data sets (LatticeKrig). This is applied to estimate return levels for daily precipitation from a subset of the North American Regional Climate Change and Assessment Program. Here the regional models comprise about 8000 grid locations over North America and so pose challenges for the statistical analysis of functional data.*

20. **Speaker**: Bruno Sanso, University of California at Santa Cruz

    Title: Using MARS for functional computer model emulation and sensitivity analysis

Abstract: *Sophisticated computer programs that produce realistic simulations of, for example, physical or biological systems, often require large amounts of computer power and take long time to produce results. Such programs usually depend on inputs that are subject to uncertainty. Due to the computational burden of running the code for each set of inputs, a full exploration of the impact of input uncertainty on the model output may be unfeasible. A popular alternative is to develop a surrogate model, based on statistical methods, that provides a fast to compute approximation to the computer model output. This is referred to as a statistical emulator. The most popular emulators are based on Gaussian processes (GP). GPs are very flexible but become computationally unfeasible when the number of inputs is very large. We implement, as an alternative, a Bayesian version of a multivariate adaptive regression splines model (B-MARS). We focus on a case study consisting of a deformation experiment for the protecting plate of a X-ray imager. We are interested in finding how sensitive the deformation in the plate is to configuration variables like plate width and spacer width, using simulations that output the profile of the deformed plate. We compare the results and the Sobol sensitivity indexes for functional emulators based on GPs, dynamic trees and B-MARS. We explore the potential for scalability of the B-MARS approach.*

In collaboration with Devin Francom, PhD student at UCSC.

21. **Speaker**: Gavin Shaddick, University of Bath

    Title: Incorporating large scale exposure modelling into environmental epidemiological studies.

    Abstract: *Coming soon.*

22. **Speaker**: Andrew Trites, University of British Columbia

    Timing Constraint: Friday morning

23. **Speaker**: Hao Zhang, Purdue University

    Title: Modeling the Complexity of Data Structure in Environmental Sciences

    Abstract: *Data in environmental sciences are not only huge in size, but may be of different spatial and temporal scales. This coupled with the spatio-temporal correlation makes the analysis of such data a challenging task. Statistical models should capture the complex structure that exists in the data. Otherwise, contradictory and misleading results may occur and this is well documented in literature that I will review. As models become complex, so should be the statistical theory that provides a basis for comparing models and methods. I will illustrate this point through an example in ecology.*

24. **Speaker**: James V. Zidek and Seagle Liu, University of British Columbia

    Title: Interpolating the high dimensional track record of the fur seal: fusing a physical model with data.

    Abstract:

    Many biologging studies deploy biologgers equipped with magnetometers and accelerometers to record animal movements at infra-second frequencies, thus allowing their tracks to be to be reconstructed at high-resolution by dead reckoning (DR). But stochastic factors limit the accuracy of the DR paths. So a conventional (but ad hoc) method was developed, which uses irregularly observed GPS points and simply shifts the DR paths to pass through them. While appealing simple, the conventional method lacks the stochastic foundation that enable quantitative uncertainty statements about the true path to be made.

    The Bayesian melding (BM) approach provides such a foundation for melding model (the DR path) with data (the GPS measurements). However that approach is computational intensive at the best of times

*and here the challenges are daunting, due the high dimensional data records. Our implementation of the BM uses a Brownian Bridge process to combine the fine-resolution (but seriously biased) DR path and the sparse (but precise) GPS measurements. But several key approximations and a conditional independence property of the Brownian Bridge process were needed to make it work. A cross-validatory assessment of the method will be described and show that the BM works pretty well, when applied to data obtained from northern fur seals (*Callorhinus ursinus*) foraging in the Bering Sea. The GPS corrected high-resolution path also revealed that the total distance traveled by the fur seals was much greater than that calculated by simply joining the dots (linear interpolation of the GPS observations)! The method has been published in a CRAN package.*

Co-authors: Brian Bataille, and Andrew Trites