Interdisciplinary Workshop for Undergraduate Students
May 17-22, 2015
Titles, Bios, and Abstracts
(in order of appearance in the schedule)

Paul Brooks: "What Causes Shifts in the Human Microbiome".

Paul Brooks is an associate professor in the Department of Statistical Sciences and Operations Research and the Department of Supply Chain Management and Analytics at Virginia Commonwealth University (VCU). He is also a fellow of the Center for the Study of Biological Complexity at VCU. He received a B.A. in Mathematics and a B.A. in Physics from the University of Virginia, and an M.S. in Operations Research and a Ph.D. in Operations Research from Georgia Tech. He serves on the Ph.D. Steering Committee for the Systems Modeling and Analysis program at VCU, a Ph.D. program with research areas in mathematical biology, operations research, and statistics. His research involves using mathematical optimization to develop new data analysis methods, and applying them to clinical and translational data.

J. Paul Brooks, Ph.D., CAP
Associate Professor,
Department of Statistical Sciences and Operations Research
and Department of Supply Chain Management and Analytics
Fellow, Center for the Study of Biological Complexity
Virginia Commonwealth University

Research Fellow, SAMSI
Email:jpbrooks@vcu.edu
Web: http://www.people.vcu.edu/~jpbrooks

Abstract:
The human microbiome is the community of micro-organisms that reside in various body habitats. Advances in sequencing technology allow us to take a ``DNA census" of these living counterparts without having to culture them to keep them alive. Surveys of the vaginal microbiome indicate that the microbiome profiles tend to cluster into distinct states, called community state types (CSTs) or vagitypes. Certain CSTs are clinically associated with higher risk for disease and adverse pregnancy outcomes. Samples from the same subject taken over time can indicate a change in CST, representing a major shift of the microbiome. We will examine published datasets containing repeated measures from subjects to understand changes in CST. The questions we seek to answer include:

1. How well can we predict a change in CST between time points?
2. Are there bacteria that signal an upcoming change?
3. Does clinical data help to improve predictions?
4. Are there patterns across the datasets?
5. What is the nature of particular CST transitions?

Daniel Taylor: "*A method for site-occupancy model estimation and objective Bayesian selection*" and "Hands-on Introduction to R".

Daniel Taylor is a postdoctoral fellow in the Mathematical and Statistical Ecology program at SAMSI and Duke University. He received his PhD in Interdisciplinary Ecology with concentration in Statistics from the University of Florida in 2014. His research has mainly focused on Bayesian variable selection problems, developing priors on the space of models to account for test multiplicity arising in increasingly large model spaces. He has also worked on estimation and selection problems related to ecological models using hierarchical Bayesian methods.

Daniel Taylor-Rodriguez
Postdoctoral Fellow
SAMSI / Department of Statistical Science Duke University
taylor-rodriguez@samsi.info / 919-685-9341

Abstract: Using presence-absence data, site-occupancy models are used to estimate the proportion of area occupied by a biological species. In surveys, observed zeroes can occur because the species of interest is truly absent from a site, or because it was present but remained undetected. By surveying repeatedly each site, occupancy models resolve the ambiguity in an observed zero (non-detections), separating detection and occupancy probabilities. In spite of the popularity of these models in the ecological literature, variable selection in this context is mostly limited to using AIC, which requires enumerating and fitting every model in the model space. Other Bayesian alternatives are available, but these rely either on parameter priors that require substantial previous knowledge --commonly unavailable for all parameters if the number of parameters is large-- or priors that, in spite of attempting to be "objective", are not suitable for model comparison. First, we present a formulation of the occupancy model with probit links and use data-augmentation to make the parameter posterior probabilities tractable. This specification suggests a formulation of the Bayesian selection procedure in terms of the data-augmented variables, conveniently helping derive "objective" intrinsic priors for this problem. Additionally, to enable the algorithm for large model spaces we propose a fast stochastic search strategy.

Lea Jenkins: "The Strawberries of Wrath: Farming Under the Realities of Drought".

Dr. Jenkins is an associate professor in the Department of Mathematical Sciences at Clemson University. She received her doctorate in mathematics from North Carolina State University. Her work is motivated by physical problems of interest, including filtration and separations processes and water resources management. In particular, she likes to help scientists, engineers, and farmers develop and use simulation tools and to enhance their product designs and operations management decisions.

Lea Jenkins
Associate Professor
Dept. of Mathematical Sciences
Clemson University
Clemson, SC 29634-0975
lea@clemson.edu / 864.656.6907

Abstract:
California is the source of 80% of the fruits and vegetables we consume in the U.S. The current drought crisis has put farmers, and the agricultural industry they support, in peril. As part of an effort to help berry farmers mitigate the effects of a forced reduction in irrigation, our team has developed and analyzed a "virtual farmer" software tool. The objective is to allow farmers to remain profitable while operating under water resource limits. In this talk, I will describe our team's use of mathematical modeling and optimization in support of conservation efforts in the Pajaro Valley region of California. I will also discuss future directions for this work which will allow for a more holistic approach to resolving problems motivated by sustainability issues.

Kimberly Kaufeld: "Hands-on Introduction to R".

Kimberly Kaufeld is a postdoctoral researcher at the Statistical and Applied Mathematical Sciences Institute and North Carolina State University. She graduated from University of Northern Colorado in Applied Statistics Spring 2014 and received my masters in Mathematics and Statistics at Minnesota State University, Mankato. She has worked at the National Center for Atmospheric Research modeling climate change and the United States Forest Service modeling beetle outbreaks. Her interest is using statistics to model ecological and environmental problems such as generalized linear models, spatio-temporal statistics, generalized method of moments, and Bayesian analysis.

Postdoctoral Fellow
SAMSI / Department of Statistics North Carolina State University
Kimberly.kaufeld@gmail.com 919-513-2445

Abstract:
We will provide a hands on approach to using R, the free statistical data analysis tool. In the tutorial we will go over how to handle your data and different types of exploratory data analysis techniques. We will also provide a short introduction to some ecological tools for presence/absence data using Occupancy models commonly seen in ecological data.

Jyotishka Datta: "A Gentle Introduction to Statistical & Probabilistic concepts with R".

Jyotishka Datta is a postdoctoral fellow in the Beyond Bioinformatics program at SAMSI and Duke University working under the supervision of Prof. David Dunson. He obtained his Ph.D. in Statistics from Purdue University in 2014 under the guidance of Prof. Jayanta K. Ghosh. His dissertation focused on some theoretical and methodological aspects of Bayesian methods for high dimensional problems, such as multiple testing and estimation of sparse precision matrices. Currently, he works on methodological problems motivated by massive datasets originating from large scale genomic studies (collaborative effort with Duke Genomic and Computational Biology) as well as some foundational problems in Bayesian statistics.

Postdoctoral Fellow
SAMSI / Department of Statistical Science Duke University
jyotishka.datta@gmail.com 765-398-2914.

Abstract:
We will cover a few basic statistical and probabilistic concepts and discuss their application in the area of Bioinformatics with a focus on implementation using R. We will start with regression and classification and extend to high dimensional applications where the number of variables exceed the number of observations. We shall also discuss a few basic concepts of Markov chain and calculation of important quantities using R.

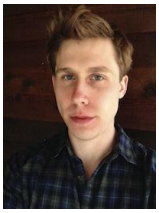
Yize Zhao: "A Gentle Introduction to Statistical & Probabilistic concepts with R" and "How to Give an Effective Presentation, Creating Slides and Posters".

Yize Zhao received her BSc degree in Statistics from Zhejiang University, China in 2010 and her PhD degree in Biostatistics
from Emory University in 2014. She is now a post-doctoral research at Statistical and Applied Mathematical Sciences Institute (SAMSI) and Biostatistics Department at University of North Carolina at Chapel Hill. Her research interests include Bayesian modeling, high dimensional data analysis and functional data analysis.

SAMSI and UNC
yzhao@samsi.info

Abstract:
This talk is focusing on how to give an oral presentation and how to create a poster presentation for research work.

Neal Grantham: "Fungi Identify the Geographic Origin of Dust Samples".



Neal Grantham is a Statistics Ph.D. student at North Carolina State University. He received a B.S. in Mathematics and a B.S. in Statistics from California Polytechnic State University in 2012, and a Master of Statistics from NCSU in 2014. His research with Brian Reich on Bayesian hierarchical modeling and machine learning seeks to capture complex relationships among variables correlated over space and time. In particular, he has worked on incorporating satellite-derived aerosol data into the modeling of harmful air pollution over the Southeastern U.S. and, more recently, joint work on forensic biology with collaborators at NCSU and CU Boulder has leveraged dust-associated fungal communities to aid in spatial source identification. This summer, Neal will intern at NASA's Langley Research Center where he will apply machine learning algorithms to eye-tracking data collected on airplane pilots to characterize human flight deck interaction in critical scenarios.

Neal S. Grantham
Graduate Student
Dept. of Statistics
North Carolina State University
5109 SAS Hall
2311 Stinson Dr.
Raleigh, NC 27695-8203
Email: ngranth@ncsu.edu
Web: nsgrantham.github.io

Abstract:
There is a long history of archaeologists and forensic scientists using pollen found in a dust sample to identify its geographic origin or history. Such palynological approaches have important limitations as they require time-consuming identification of pollen grains, a priori knowledge of plant species distributions, and a sufficient diversity of pollen types to permit spatial or temporal identification. We demonstrate an alternative approach based on DNA sequencing analyses of the fungal diversity found in dust samples. Using nearly 1,000 dust samples collected from across the continental U.S., our analyses identify up to 40,000 fungal taxa from these samples, many of which exhibit a high degree of geographic endemism. We develop a statistical learning algorithm via discriminant analysis that exploits this geographic endemicity in the fungal diversity to correctly identify samples to within a few hundred kilometers of their geographic origin with high probability. In addition, our statistical approach provides a measure of certainty for each prediction, in contrast with current palynology methods that are almost always based on expert opinion and devoid of statistical inference. Fungal taxa found in dust samples can therefore be used to identify the origin of that dust and, more importantly, we can quantify our degree of certainty that a sample originated in a particular place. This work opens up a new approach to forensic biology that could be used by scientists to identify the origin of dust or soil samples found on objects, clothing, or archaeological artifacts.

Bo Zhang & Yue Qi: "Leverage Big Data to Fight Fraud".

Bo Zhang is a data scientist at EMC, with more than 5 years of experience working on data sampling, data profiling, data modeling, data warehouseing, predictive analytics and reports, which included industry working knowledge in the education, financial services, health care, telecommunications and retail arenas. With a master in Mathematics and a Ph.D. in statistics, he has conducted research and customer-based projects, based on data mining, mathematics, statistics, and machine learning algorithms.

Bo Zhang
EMC
bo.zhang@emc.com
9192470166

Dr. Yue Qi works as a data scientist in SAS Institute Inc. His work focuses on modeling and interactive visualization of big data, providing innovative insights for big data analytics, and parallel and distributed computing platform development. Dr. Qi?s research interests include predictive modeling and machine learning for big data and high dimensional data, distributed and parallel computing, and anomaly detection. He is based in Cary, North Carolina.

Yue Qi
SAS
yue.qi@sas.com,
9197933775

Abstract:
In this talk we will introduce applications of advanced statistical and machine learning methods in fraud detection. Fraud incurs tremendous loss every year in the fields such as banking, credit card application and usage, insurance claims, trader surveillance, health care claims, and government funding and allowance management. Successful practices show that advanced statistical and machine learning models are powerful tools to fight fraud. Specially designed models are needed for fraud detection since fraud data is usually very imbalanced due to the nature that fraudulent activities are very rare compared with legitimate activities. Big data brings both challenges and opportunities to fraud detection - helps building more accurate models to fight fraud but requires very fast and large scale computing abilities. Analytics architecture for big data is also covered.

Christopher Strickland: Introduction of Python and Mathematics behind some of the data.

I am a Postdoctoral Fellow in the Mathematical and Statistical Ecology program at SAMSI and the University of North Carolina, Chapel Hill. I received my PhD in mathematics at Colorado State University in December 2013. My research is broadly focused on modeling, analyzing, and optimizing systems in ecology, with my main projects at the moment concerning the dispersal of parasitoid wasps as biocontrol for agricultural pests and modeling the spread of white nose bat syndrome between caves in the United States. Mathematically, I am broadly interested in both applications and mathematical methods for dynamical systems, modeling and data analysis. My background in applied mathematics is diverse, including mathematical modeling, numerical analysis, dynamical systems theory, and mathematical biology, and I enjoy collaborating directly with scientists and engineers to obtain practical, data driven results.

SAMSI and UNC
cstrickland@samsi.info