Abstracts
Spring Meeting – April 15, 2016
Colorado / Wyoming Chapter of the American Statistical Association

Caitlyn Cole, Ian Greenwald, and Cody Griffith
Metro State University of Denver

"Diving into Bear Creek"

In this presentation, we address the growing E. coli problem in our local Bear Creek through the use of statistical modeling. Bear creek runs from Morrison and meets with the South Platte River in Englewood, so many residents in the Denver and Jefferson Counties are directly affected by contamination in the creek. The E. coli populations in the creek have exceeded the EPA recommended safety level and we have focused our attention to discovering why. Is it a human source? Weather and climate change? We will get to the bottom of this problem.

The approach we have taken is to establish a multiple regression model to factor in many potential contributing sources. This also allows us to make predictions about when the E. coli levels are unsafe based on a few key factors. We have used kriging to establish a spatio-temporal model for visual representation of the E. coli population, which can show the spatial dependency as well as the time dependency of the population. This will help identify where and when the populations reach a relative maximum!

Ekaterina Smirnova, Farhad Jafari, and Snehalata Huzurbazar
University of Wyoming

"Microbiome Data Normalization Methods"

Human Microbiome Project (HMP) is a large scale nationwide study that utilizes next generation sequencing technology (NGS) to investigate the relationships between the human microbiota composition, diet and health status. Fragments of DNA sequences obtained in these experiments are classified at a species level, and typically referred to as species or taxa. One particular characteristic of these studies is that the data are often quite sparse but collected on a large number of variables, many of which are possible contaminants. To remove possible contaminants, a data normalization step, known in microbiome literature as filtering is applied prior to analysis. Currently there is neither any consensus on filtering criteria used, nor is there an evaluation of loss due to filtering done. We propose a taxa co-presence network based data normalization method that removes extremely rare taxa, evaluate loss due to filtering, and take a step towards understanding the algorithmic dimension reduction methods such as nonnegative matrix factorization.

Eric Gilleland
National Center for Atmospheric Research

"Evaluating high-resolution forecast performance"

With the advent of increasingly higher resolution forecast models, the challenge of verifying those forecasts has multiplied. Often, high-resolution models are found to be more useful than coarse-scale ones by human forecasters, but they would seldom verify as well. Reasons for this behavior include the double penalty problem (when the same error, e.g., a timing/spatial displacement error, is penalized twice, e.g. once for misses and again for false alarms), as well as more numerous small-scale errors that add up. Moreover, a need for more diagnostic information about forecast performance (e.g., how did the forecast do well or poorly) has recently been demanded. All of these criteria have led to a multitude of new verification procedures, referred to as spatial forecast verification. The methods come from several disciplines, such as image analysis, computer vision, spatial statistics, shape analysis, and others. This talk will give a flying arm-chair view of most of the variously proposed methods, and attempt to distill their main features.

Philippe Naveau, Alexis Hannart, Aurelien Ribes and Francis Zwiers
Visiting Scientist – National Center for Atmospheric Research

"How To Revise Return Periods For Record Events In A Changing Climate"
Breaking a record simply means that the current observation exceeds all past measurements. Such a type of an event is regularly followed to a media frenzy and, in such instances, climatologists are often asked if the frequency of record has increased with climate change.

This leads to the question of Detection and Attribution (D&A)   ("Detection" is the process of demonstrating that climate has changed in some defined statistical sense, without providing a reason for that change and "Attribution" is the process of establishing the most likely causes for the detected change with some defined level of confidence, see the IPCC definition).

The field of statistics has become one of the mathematical foundations in D&A studies because computing uncertainties represent   difficult inferential challenges when analyzing small probabilities.

Pratyaydipta Rudra
University of Colorado – Health Sciences

"Model based heritability scores for high-throughput sequencing data“

(A joint work with Brian Vestal, Wen Shi, Laura Saba and Katerina Kechris)

Heritability of gene expression as a molecular trait has been of interest due to the fact that it helps us understand the nature of the relationship between expression and genetics. Previous studies have performed the analysis of heritability of gene expressions using analysis of variance techniques. Use of intra-class correlation coefficient from a linear mixed effects model as a measure of heritability was common. However, next generation high-throughput sequencing data presents a challenge in properly modelling and estimating the heritability scores. The models required for such count data are usually generalized linear mixed models where the computation and interpretation of intra-class correlation is not straightforward. We propose a measure of heritability score based on variance partitioning for a class of dispersion models. The method is tested against alternatives such as using linear mixed effects model after performing a variance stabilizing transformation. Extensive simulations as well as analysis of real data shows the efficiency of our approach.

Xiyue Liao and Mary C. Meyer
Department of Statistics, Colorado State University

"Change-Point Estimation using Shape-Restricted Regression Splines"

We consider estimating a regression function fm and a change-point m, where m is a mode or an inflection point. For a given m, the least-squares estimate of fm is found using constrained regression splines, then the set of possible change-points is searched to find the overall least-squares ˆm. Convergence rates are obtained for each type of change-point estimator, and simulations show that for small and moderate sample sizes, these methods compare well to existing methods. The extension to correlated errors is given, and the methods are available in the R package ShapeChange.
Keywords: mode estimation, inflection point, monotone, convex, convergence rate.

Carolyn P. Johnston, Ph.D.
Sr. Manager, Image Mining; DigitalGlobe

"Accuracy metrics for automatically detected settlement boundaries"

I will describe a project in which settlement boundaries were extracted automatically from large quantities of DigitalGlobe satellite imagery, followed by false alarm reduction via an online crowdsourcing campaign by Tomnod. I will also discuss our approach for estimating the recall of the automated process.

Branden Olson
University of Colorado - Boulder

"Stochastic Weather Generators"

Stochastic weather generators (SWGs) are designed to create simulations of synthetic weather data and are frequently used as input into physical models throughout many scientific fields. In particular, the simulation of spatially coherent precipitation occurrence over a spatial domain presents a major challenge, in part due to the difficulty of estimating model parameters. We propose a technique to fit SWGs for precipitation occurrence based on an emerging set of methods called Approximate Bayesian Computation (ABC), which bypass the evaluation of the likelihood function. We demonstrate the viability of our technique through a case study of historical precipitation data from the state of Iowa, and show that the simulations exhibit statistical characteristics similar to the observations.

Peter DeWitt
Department of Biostatistics and Informatics
Colorado School of Public Health, University of Colorado Denver

B-spline transformations of continuous predictors are commonly used in regression models to estimate a smooth non-linear relationship with the response. The quality of the regression fit is subject to a knot sequence ξ. Selection of a knot sequence is traditionally achieved by choosing between regression models with a varying number of knots, which are placed at the predictor quantiles. AIC or BIC are then used for model selection. It is well known that AIC and BIC can result in big models (i.e. models with a large number of internal knots). If parsimony, minimizing the number of interior knots $n(\xi)$, is important this approach is not desirable. Our goal is to develop an efficient knot selection algorithm that selects models with smaller $n(\xi)$ without sacrificing goodness of fit.

Instead of focusing on likelihood maximization, we present a knot selection method based on the geometry of the b-spline control polygons (CP). CPs have been used extensively in computer aided graphic design and numeric analysis; primarily for deriving and evaluating B-spline approximations to fit complex shapes measured with little to no noise. Changes in CP provide a useful metric for assessing the influence of a particular knot, which we demonstrate can then be used for smart removal of knots.
Our control polygon reduction (CPR) algorithm starts with a CP based on an initial ξ with large $n(\xi)$ and knot positions on a fine partition of the predictor. Inspired by Lyche and Morken (1988), we assess the influence of each knot on CP geometry and omit the knot exerting the least-influence on the CP shape. After a knot omission, the model is refit with the coarsened knot vector. The process continues until all internal knots are removed. The final regression model is selected as the model with the smallest $n(\xi)$ such that a single additional knot has negligible effects on CP geometry.

We show that for a wide range of functional shapes, including complex longitudinal hormone data, the CPR algorithm results in a final model with fewer internal knots than models selected via traditional approaches and with negligible differences in the sum of squared residuals. CPR is computationally efficient and provides high quality fits built on low-rank design matrices. The CPR algorithm is an attractive solution for knot selection in a wide range of applications.

E. Jackson, A. Menon, E. Smirnova, S. Huzurbazar
University of Wyoming

"Exploratory visualization of 16s rRNA microbiome data"

Microbiome data obtained from sequencing the 16S rRNA gene consists of a large number of variables (taxa) and usually a small number of samples. Ordination techniques such as principal component analysis achieve dimension reduction using only the first two or three principal components to explain a large proportion of variation in the data. However, in microbiome data sets, many more than 3 principal components are required to account for a reasonable amount of the variation in the data. We introduce several new perspectives on visualizing data on more than 3 principal components, along with R code designed to facilitate this process. Biological implications will also be discussed.

Sean Lopp - RStudio

"Increasing Impact: Adding Interactivity to Your Workflow"

We will highlight the increasing importance of interactive graphics in both exploratory analysis and final presentations. We will then cover how to get started using the R package Shiny and hosting service shinyapps.io.