NCCC-170 Annual Meeting
June 20-21, 2019
University of Illinois, Champaign-Urbana, IL

Technical Program

Thursday, June 20, 2019

8:00-8:30 am
Registration, check-in, coffee, juice, snacks

8:30-8:45 am
Welcome
CHRIS HARBOURT, CEO, AirScout, Inc.

8:45 – 9:20 am
"Hierarchical modeling of structural coefficients for heterogeneous networks with an application to animal production systems"
K. Chitakasempornkul[1], G. J. M. Rosa[2], A. Jager[1], N. M. BELLO[1], [1]Kansas State University, Manhattan, [2] University of Wisconsin-Madison, Madison.

Understanding the interconnections between performance outcomes in a system is increasingly important for integrated management. Structural equation models (SEM) are a type of multiple-variable modeling strategy that allows investigation of directionality in the association between outcome variables, thereby providing insight into their interconnections as putative causal links defining a functional network. A key assumption underlying SEM is that of a homogeneous network, whereby the structural coefficients defining functional links are assumed homogeneous and impervious to environmental conditions or management factors. This assumption seems questionable as systems are regularly subjected to explicit interventions to optimize the necessary trade-offs between outcomes. Using a Bayesian approach, we propose methodological extensions to hierarchical SEM that explicitly specify structural coefficients as functions of systematic and non-systematic sources of variation, thus allowing for hierarchical heterogeneity in the network links and recognizing design structure in the data. We validate our proposed method using a simulation study and show that hierarchical sources of heterogeneity on structural coefficients can be estimated and inferred upon accurately. Further, we show that networks can be consistently identified as homogeneous or heterogeneous based on model fit statistics that compare competing SEMs with flexible specifications of structural coefficients. We apply the proposed methodological extensions to a dataset from a designed experiment in swine production consisting of six interrelated reproductive performance outcomes to explore physiological links that differed by parity while accounting for experimental design. Overall, our results indicate that explicit hierarchical SEM-based modeling of heterogeneous functional networks can be used to advance understanding of complex networks of performance outcomes in an animal production system.

9:20-9:55 am
"Generalized Additive Mixed Modeling of Total Transport Losses of Market-Weight Pigs"

GUILHERME J. M. ROSA, Dept. Animal Sciences, and Dept Biostatistics & Medical Informatics
University of Wisconsin-Madison.

The objective of this study was to investigate factors associated with total transport losses (TTL) of market-weight pigs using a Generalized Additive Mixed Model (GAMM). The dataset was provided by Iowa Select Farms (Iowa Falls, IA), and included the information on 26,819 shipments delivered to 2 abattoirs. TTL was fitted as an overdispersed binomial process using a GAMM which included the fixed effects of abattoir, type

of driver (i.e. truck owner or employee), distance traveled, average market-weight, wind speed, precipitation, and temperature-humidity index (THI), as well as the random effects of truck company and the combination of farms and quarter of the year. The estimated risk of losses was 0.867 (95% CI: 0.865-0.870) times lower when trucks were driven by owners instead of by employees, suggesting that truck owners have more vested interest. THI was associated with TTL ($P < 0.0001$, Figure 1A), displaying greater risk of losses at its extremes, indicating that additional care must be considered at these levels. The interaction between wind speed and precipitation was associated with TTL ($P = 0.0209$, Figure 1B), indicating a complex relationship between both explanatory variables and TTL. The distance traveled was associated with TTL ($P = 0.0034$, Figure 1C), with increased losses at distances up to 125 km and decreased risk of losses afterwards. This result suggests that long trips may give extra time to pigs to recover from the prior stress incurred at loading. The interaction between average market weight and abattoir was positively associated with TTL ($P < 0.0001$, Figure 1D), indicating a faster increment in the risk of losses in one facility relative to the other. In conclusion, TTL of market-weight pigs are caused by a complex system involving multiple interacting factors. Furthermore, GAMM is shown to be a flexible predictive tool capable of modeling non-linear relationships and which can be used to support decision-making in swine industry.

9:55-10:10 am
Moring break

10:10-10:45 am
"Generalized Linear Mixed Model Approach to Time-to-Event Data with Censored Observations"
KATHLEEN YEATER[1], George Yocum[2], Joseph Rinehart[2], Arun Rajamohan[2], Julia Bowsher[3], Kendra Greenlee[3], [1]USDA-ARS-Plains Area Office of the Director, [2]USDA-ARS-Insect Genetics and Biochemistry Research Unit, [3]Dept. of Biological Sciences, North Dakota State University.

The time-to-event response is commonly thought of as survival analysis, and typically concerns statistical modeling of expected life span. In the example presented here, alfalfa leafcutting bees, *Megachile rotundata*, were randomly exposed to one of eight experimental thermoprofiles or two control thermoprofiles, for one to eight weeks. The incorporation of these fluctuating thermoprofiles in the management of the bees increases survival and blocks the development of sub-lethal effects, such as delayed emergence. The data collected here investigates the question of whether any experimental thermoprofile provides better overall survival, with a reduction of sub-lethal effects. The study design incorporates typical aspects of agricultural research; random blocking effects. All *M. rotundata* prepupae brood cells were randomly placed in individual wells of 24-well culture plates. Plates were randomly assigned to thermoprofile and exposure duration, with three plate replicates per thermoprofile x exposure time. Bees were observed for emergence for 50 days. All bees that were not yet emerged prior to fixed end of study were considered to be censored observations. We fit a generalized linear mixed model (GLMM), using the SAS® GLIMMIX Procedure to the censored data and obtained time-to-emergence function estimates. As opposed to a typical survival analysis approach, such as Kaplan-Meier curve, in the GLMM we were able to include the random model effects from the study design. This is an important inclusion in the model, such that correct standard error and test statistics are generated for mixed models with non-Gaussian data.

10:45-11:20 am
"Evaluating the comparative performance of popular gene set test methods"
JOHN R. STEVENS and Richard M. Lambert, Utah State University.

Gene set testing is a method of characterizing differential expression between sample groups by grouping functionally related genes together into gene sets. Statistical tests for differential expression are performed on the gene sets rather than on individual genes. In recent years, a number of gene set test methods have been developed and independently tested, but little has been done to do a wide scale comparison of these gene set methods. We use a custom simulation framework to compare the statistical power and false discovery rate of

a number of current gene set test methods (including mvGST, ROAST, CAMERA, ROMER, GlobalTest, GSA, PAGE, SAFE, sigPathway, and GSEAlm) over several combinations of realistic and relevant biological parameters. In total, 100 unique test scenarios were performed using 8,700 distinct simulated data sets. We found that all of the methods are subject to the classic trade-off between the power and the FDR. Some of the methods were more powerful but failed to maintain the FDR in some cases, while other methods maintained the FDR at the expense of power. In general, the gene set test methods either did not hold power or maintain the FDR in the presence of realistic conditions with sample sizes of 2 to 4 that are commonly used in real gene expression experiments. We also found that none of the methods performed well under realistic inter-gene correlation (dependent gene expression values within gene set) and conclude that further research and development is needed in this area.

11:20-11:50
"Evaluation of strategies for multi-trait association studies in maize architectural traits"
BRIAN RICE and Alex Lipka, Department of Crop Sciences, University of Illinois.

Genomic selection and genome-wide association studies (GWAS) are typically performed univariately on a single trait. However, multi-trait approaches that use information from correlated traits measured within specific plant and animal species are an emerging focus of quantitative genetics. Although promising, this area of study requires further evaluation across a wide variety of genetic architectures and species. Therefore, we studied leaf and inflorensce maize architectural traits that have been shown to be associated with putatively pleiotropic genomic loci. Using the eigenvalues from a principal component analysis (PCA) is a computationally efficient and available alternative to true multivariate GWAS model. Preforming GWAS on the PCs of related traits has the potential to uncover genomic markers with pleiotropic effects. To aid in distinguishing between pleiotropic and non-pleiotropic loci, we are currently developing an analytical pipeline that utilizes the simultaneous application of multivariate (PCs approach) and univariate statistical models to associate genomic markers with these traits. Finally post hoc tests using the *p* values from each association test will be implemented to obtain a single significance value for each genomic marker. In addition to facilitating the identification of genomic regions that potentially harbor pleiotropic loci, we hypothesize that analysis of this data will help calibrate accurate modeling of genetic and non-genetic sources of trait variability.

11:50 am – 12:20 pm
"Training Population Optimization for Genomic Selection in *Miscanthus*"
MARCUS O. OLATOYE and Alex Lipka, Department of Crop Sciences, University of Illinois.

Miscanthus is a C_4 perennial grass with great potential for lignocellulosic ethanol biofuel production. However, there is a significant drawback for the further development of this biofuel crop due to lack of sufficient winterhardiness in northern latitudes. Abundant genetic diversity exits for different traits in this crop species that can be introgressed to improve current cultivars. In this study we explored the use of diversity panels for training genomic selection (GS) models to predict phenotypically optimal lines. Also, we examined the challenges associated such methodology by (1) evaluating the impact of population structure in *Msi* and *Msa* diversity panels and (2) quantify the advantages and disadvantages of using both *Msi* and *Msa* panels as a training set for fitting GS models. Discriminant analysis of principal components, principal component analysis, and coefficient of determination methodologies were used to assess the impact of population structure on the prediction accuracy of the GS models. Our results demonstrated that population structure did have an effect on the performance of GS models, where the performance varied among the methods employed to account for population structure. GS prediction accuracies varied across training sets from low to moderate and were trait dependent. The implications of the results of this study, in particular the considerations one needs to take when fine tuning training sets and GS models for optimal predictions of breeding values, will be discussed.

12:20-1:15 pm
Lunch

1:15-1:50 pm

"Using a binomial GLMM for partially paired categorical data"

PHILIP DIXON, Department of Statistics, Iowa State University.

The motivation for this talk is a comparison of two models to predict invasiveness of a horticulturally interesting tree or shrub. The data are a list of trees and shrubs introduced for horticultural purposes to a state and, for each species, whether it is invasive (i.e., has established new populations in non-horticultural sites, and 11 biological characteristics. Random Forest models predicted the probability of invasiveness from the biological characteristics. We developed two models, one for species in the Chicago area and one for species in Iowa. The biologists wanted to know whether one model made more accurate predictions than the other. The complication is that the species in the Chicago data set partially overlap those in the Iowa data set. If there was no overlap, the analysis is a comparison of proportions from independent observations. If all species were on both lists, the analysis is a comparison of proportions from paired observations. Our experience with these models is that some species (e.g., Berberis thunbergii) are simply hard to predict accurately, so the pairing can not be ignored. To use all the data, we need an analysis that allows for partial overlap of lists. Hence, the model needs to account for heterogeneity among species. One option is a Bernoulli GLMM with a random effect for species. I compare four approximately likelihood estimators (SAS/PQL, SAS/PQL with KR adjustments, SAS/Laplace, and R/Laplace) and a Bayesian estimator implemented using Rstanarm.

1:50-2:25 pm

"The all-zero treatment problem with conditional binomial data and GLMMs"

LAURENCE V. MADDEN, Department of Plant Pathology, Ohio State University.

There are numerous advantages in using generalized linear mixed models (GLMMs) for the analysis of non-normal data, especially discrete data. Yet there are also multiple challenges with using GLMMs that are not of relevance, or of less relevance, when fitting linear mixed models to data. It is usually easy to fit a GLMM to data when the response variable has a binomial distribution conditional on the random effects. This is of high relevance in plant pathology, where disease incidence (number of diseased plants, y, out of n assessed plants in each experimental unit) is typically analyzed. With a GLMM fitted to such data, one can estimate the probability of a plant being diseased, π, through the inverse-link function for each treatment, of direct interest in plant epidemiological research. A not uncommon situation occurs when all of the plants are disease-free for all blocks for one (or more) of the treatments. This all-zero treatment creates a quasi-complete separation problem, making it very challenging to fit a GLMM to the data. Essentially, the model is ill-conditioned with such a dataset. Work to date shows that pseudo-likelihood estimation will not converge with the all-zero treatment problem. Likelihood approximation methods (e.g., Gauss-Hermite quadrature) typically do converge, but parameter estimates for one or more of the treatment effects (e.g., τ_i in a model with an intercept) are unreliable, even for treatments without all zeroes. Standard errors (SEs) are typically greatly inflated, as are the treatment-effect parameter estimates (τ_i). However, estimated expected values for some of the treatments, and for some treatment differences, may be reliable. The problematic treatment-effect parameters may not always correspond to the treatment with all zeroes, and it is not always easy to identify difficulties. Parameter estimates and their estimated SEs for a given dataset can depend on the optimization method used with integral approximation (e.g., quasi-Newton vs. Newton-Raphson vs. conjugate gradient optimization). The presentation will discuss implications of some of the ad-hoc ways of adjusting the data to allow for successful model fitting with integral approximation, and seek advice on moving forward when datasets have this all-zero property for some treatments. Bayesian analysis with informative priors for treatments with all zeroes is one potential approach to deal with this challenge.

2:25-3:00 pm

"Exploring longitudinal data with historical random forests"

SUSAN DURHAM[1], Leila Shultz[1], James Long[1], Wanda Lindquist[1], and Douglas Johnson[2], [1]Utah State University, [2]Utah National Guard

Ecological data sets often are observational, rather than experimental, with a variety of characteristics that present analysis challenges in the traditional statistical framework. The number of observations may be small while the number of predictors is large (i.e., the small *n*, large *p* problem). The nature of relationships between the response variable and predictor variables may be unknown and non-linear. Predictor variables may exhibit multicollinearity among themselves or interaction with the response. The response variable may be "distributionally challenged". Observations may not be independent, instead being clustered or measured over time. And yet, the researcher wants to know which predictor variables are "important" and how those predictors are related to the response. The random forest model is an appealing modeling tool that addresses some (but not all) of the challenges inherent in ecological data and has seen applications in the last decade. However the standard software does not accommodate lack of independence among observations and so is of limited value for longitudinal or otherwise clustered data.

I introduce the R package htree (Sexton 2018) which produces a nonparametric estimate of the relationship between a response variable and its previous history as well as the histories of time-varying predictor variables; it also provides measures of variable importance. The htree package fits random forest and gradient boosted ensemble models. My focus here is on the historical random forest as fitted by the hrf function. The advantages and disadvantages of this model will be illustrated by an analysis of longitudinal vegetation data collected sporadically over 20 years on 96 plots at Camp W.G. Williams, a training site located south of Salt Lake City and operated by the Utah National Guard. The goals of the analysis are to document shifts in species abundance over time and to identify and characterize the predictor variables that may drive these changes.

3:00 – 3:15 pm
Break

3:15-3:45 pm
"A Review and Discussion of Residuals for Mixed Models"
KATHERINE GOODE, Department of Statistics, Iowa State University.

Residuals are a key tool used to diagnose models. As a statistical consultant for researchers in many areas, I often find myself reminding my clients to visualize residuals to assess model assumptions. Many of my clients are working with mixed models, and I recently realized that I often recommend the use of certain residual types without a full understanding of the implications of selecting one type over another. This led me to have an interest in better understanding the many residuals types for mixed model. In this talk, I will provide a review of the residual types available for linear mixed models (marginal, conditional, studentized, etc.). I will explain how the residuals are computed and how these computations differ between R and SAS. I will also discuss what I have learned from the literature about how to select a residual type when assessing a model. Lastly, I will briefly touch on residual types for generalized linear mixed models and list some unanswered questions. If time permits, I will pose these remaining questions to the attendees to discuss as a group.

3:45-4:30 pm
Group Discussion: Potential replacement for the now defunct Kansas State Conference on Applied Statistics in Agriculture

4:30 pm
Adjourn

Friday, June 21, 2019

8:30-9:00 am

"Ordination plots for model-based analysis of species composition data: Connecting two very different sets of methods"

PHILIP DIXON, Department of Statistics, Iowa State University.

Traditional analysis of species composition data starts by computing a dissimilarity between each pair of samples. These dissimilarities are the basis for ordination plots, which explore patterns, and ANOVA-like inference. The more recent model-based analysis of species composition starts with an explicit probability model for the data. This provides a principled approach for inference but it has been difficult to produce an ordination plot. Often, a traditional ordination method is combined with a model-based analysis, but these make different assumptions about the data. I develop an approach to visualizing multivariate species composition that is consistent with a model-based analysis, in the sense that the plot and the analysis make the same assumptions. This approach retains the data distribution from a model-based analysis and makes no additional assumptions about the structure of the data.

The probability model is used to define a likelihood-based dissimilarity between each pair of observations. I derive these for Bernoulli (i.e., presence/absence) data, four models for count data, and Gaussian data. An ordination plot is constructed by multidimensional scaling. The approach can be extended to overlay model-based predictions of species composition on the ordination plot or compute a community-level residual. I illustrate these using nesting bird counts on Skokholm Island and tropical tree stem counts on Barro Colorado Island.

9:00-9:30 am

Group Discussion: Bayesian analysis book

9:30-11:00 am

NCCC-170 Business meeting

In addition to the usual agenda item, there will be an additional discussion on Project impact importing based on the updated North Central Project Handbook

Adjourn Meeting at 11:00 am.

11:00 am

Visit to Morrow Plots of the University of Illinois (***optional***)