Conference on Applied Statistics in Agriculture and Natural Resources

May 11-15, 2025

Gainesville, Florida, USA

ABSTRACT COMPILATION

Abstracts are listed in alphabetical order by **presenter last name.**

[This page left intentionally blank]

Quantifying the Sensitivity of Land Use Land Cover Metrics Through Simulation Techniques

Haley Burger[1] *and* ***Brennan Bean***[1]
[1]Department of Mathematics and Statistics, Utah State University, Logan, UT, USA

Accurate methods for characterizing land-use-land-cover (LULC) change are critical for addressing challenges in agriculture, such as food security and environmental health. A leading data source for analyzing LULC change is the USDA Cropland Data Layer (CDL), produced annually by the National Agricultural Statistics Service (NASS). With a classification accuracy of 85-95% for major crops, researchers sometimes use the CD to calculate various LULC metrics for use in environmental, agricultural, and ecological studies. However, limitations in CDL data, including temporal bias, pixel misclassification, and challenges distinguishing vegetation types, raise questions about the sensitivity of LULC metrics to the published inaccuracies in the CDL. To address these issues, we create the R package *cdlsim*, which simulates spatially aware perturbations of CDL data at the patch (i.e., a collection of adjacent pixels of the same class) level using confusion matrices provided by NASS and used to validate the accuracy of the CDL. The result of the simulations is the production of "sensitivity intervals," which represent the plausible range of landscape metric values based upon each calculated metric, given the random perturbations. We intentionally avoid the use of the term "confidence interval" since the simulations do not necessarily produce a distribution for some true and unknown LULC metric but rather illustrate the potential variability in the calculation of an LULC metric that can be attributed to published or potential misclassification rates. We demonstrate the utility of *cdlsim* with two case studies, one in an ag-dominated portion of South Dakota and another in a non-ag-dominated watershed in Utah/Idaho. The results show how *cdlsim* quantifies the sensitivity of landscape metrics based upon the published misclassification rates provided by NASS, offering a more nuanced representation of land use change, and improving the reliability of CDL-based studies for policy and land management.

Correlated Binomial Data: A Confluence of Analytic Challenges

Nora M. Bello*[1]*, *Clark Kogan*[2]*, Bruce A Craig*[3]*, Daniel G. Palmer*[4]*, Susan L. Durham*[5]*, and Walter Stroup*[6]
[1]United States Department of Agriculture – Agricultural Research Service, USA
[2]Washington State University, Spokane, WA, USA
[3]Purdue University, West Lafayette, IN, USA
[4]University of North Dakota, Grand Forks, ND, USA
[5]Utah State University, Logan, UT, USA
[6]University of Nebraska-Lincoln, Lincoln, NE, USA

Generalized linear mixed model (GLMM) analyses of repeated-measures binomial data pose a unique confluence of challenges. In the context of correlated binomial data, seemingly minor analytic decisions have critical consequences with cascading effects through estimation and inference. Specific challenges involve use of research objectives to inform decision on inference space and inference target prior to model specification. Said decision then dictates the choice between non-equivalent model specifications, namely conditional (G-side) vs. marginal (R-side), which in turn, poses non-trivial constraints on methods of estimation and bias correction.

Estimation of binomial GLMMs can proceed either by integral approximations (G-side) or linearization through penalized quasi-likelihood (PQL; R-side) or pseudo-likelihood (PL; both G- or R-side models). In particular, PL enables separation of the (pseudo) residual likelihood, thus enabling REML-like optimization in both R-side and G-side models. In its REML-like form, PL enables use of traditional mixed models capabilities for estimation of degrees of freedom and bias reduction in variance estimates and test statistics.

As with other repeated measures studies, model specification further involves deciding amongst competing covariance structures. For binomial GLMMs, these comparisons can be done for G-side models via standard model fit comparisons. With R-side models, covariance structures cannot formally be compared, so estimation by PL or PQL linearization is complemented by sandwich estimators robust to covariance specification and, in the case of small sample sizes, by Morel-Bokassa-Neerchal bias correction.

In this talk, we illustrate the thought process through these challenges for repeated measures binomial data, followed by a series of analytic decisions in the context of a completely randomized design. We further discuss the current state for GLMM implementations for correlated binomial data using R (lme4, glmmTMB, glmmPQL) and SAS (PROC GLIMMIX). We highlight important misalignments and technical gaps in the current analytic capabilities of each software platform. We end with a description of custom-made solutions from our group and recommendations for future software development.

Best Practices for GLMM Implementation: Setting the Stage

Nora M. Bello[1], *Walter W. Stroup*[2], *Julia Piaskowski*[3], *Josefina Lacasa*[4], *Reka Howard*[2], *Daniel G. Palmer*[5], *Quentin D. Read*[1], *Susan L. Durham*[6], *Conor Fair*[7], *Clark Kogan*[8], *Raúl E. Macchiavelli*[9], *Bruce A Craig*[10].
[1]United States Department of Agriculture – Agricultural Research Service, USA
[2]University of Nebraska-Lincoln, Lincoln, NE, USA
[3]University of Idaho, Moscow, ID, USA
[4]Kansas State University, Manhattan, KS, USA
[5]University of North Dakota, Grand Forks, ND, USA
[6]Utah State University, Logan, UT, USA
[7]University of Georgia, Griffin, GA, USA
[8]Washington State University, Spokane, WA, USA
[9] University of Puerto Rico, Mayagüez, PR, USA
[10]Purdue University, West Lafayette, IN, USA

Our objective in this session is to articulate best practices for the implementation of Generalized Linear Mixed Models (GLMM) in the context of agricultural and biological sciences. We present the progress of a joint effort by members of the North Central Coordinating Committee 170 "Research Advances in Agricultural Statistics" (NCCC170.org).

Motivated by a broad array of real-data examples (and a few realistically simulated), the talks and accompanying posters in this session illustrate critical conceptual and implementation challenges commonly encountered with mixed modeling. We further highlight renewed challenges conspicuously to GLMMs, despite their common misperception as settled matters based on lessons learnt from Gaussian data.

For each data example, we describe specific issues and illustrate implementation of best GLMM practices in the software platforms R and SAS, to the extent possible. Admittedly, modern GLMM implementation is heavily reliant on advanced software. Yet, we are emphatic in focusing further discussions on best practices users should know and address appropriately regardless of software of choice, rather than any software battles. Throughout, we emphasize direct inferential implications for statistical practice. When applicable, we delineate implementation gaps in need of further development. We conclude with practical recommendations on best GLMM practices for the statistical practitioner, intended to ensure proper GLMM specification, implementation and interpretation. Finally, we share on-going progress on web-based educational resources for practicing statisticians, instructors and associated continuing education efforts. These resources are intended as living documents aligned with our evolving understanding of best GLMM practices.

EoA Models: Is Partially Combining Data from Multiple Sites Useful for Estimating Bat Mortality Rates in Wind Farms?

Natalia Berberian[1] *and Philip M. Dixon*[2]

[1]Departamento de Biometría, Estadística y Computación, Facultad de Agronomía, Udelar, Montevideo, Uruguay

[2]Statistics Department, Iowa State University, IA 50011, USA

The number and size of wind farms generating renewable energy are increasing globally. In agricultural landscapes, wind farms are a major source of wildlife mortality, especially for bats due to barotrauma. Estimating true mortality rates is challenging due to low detection probability. Only a fraction of the area is surveyed, scavengers may remove evidence before searches, and search teams often fail to detect fatalities. As a result, fatalities may occur without being recorded.

The Evidence of Absence (EoA) model estimates mortality at a single site by incorporating detection probability into a binomial mixture model. EoAR (Evidence of Absence Regression) extends EoA to multiple sites or years, allowing for covariate estimation within the same model.

Currently, these models are applied to multi-site analyses using different parameter values per site (no pooling) or the same values for all sites (complete pooling). We explore partial pooling as an alternative to improve fatality estimates across sites.

Our goal was to estimate the variability of detection probability components across sites and assess how pooling methods affect bat mortality estimates. Using post-construction monitoring data from 20 wind farms in Iowa, USA, we compared mortality estimates from complete pooling (CP), no pooling (NP), and partial pooling (PP) for three bat species: Little brown bat (LBBA: *Myotis lucifugus*), Evening bat (EVBA: *Nycticeius humeralis*), and Hoary bat (HOBA: *Lasiurus cinereus*). We also evaluated simulated scenarios with varying site heterogeneity to assess pooling effects.

Results from real data show that model performance depends on species and project-specific conditions. The PP model performed well in some cases, especially for *M. lucifugus* and *N. humeralis*, sometimes outperforming CP and NP. However, its performance for *L. cinereus* was inconsistent. Despite this, PP's comparable results to NP suggest it is a viable option. Our findings highlight the need for refined model selection rather than relying solely on CP or NP, as no single model is best in all cases.

Simulated scenarios indicate that PP consistently provides the most accurate and precise fatality estimates across all evaluated conditions, outperforming CP and NP. These findings underscore the importance of selecting a model that remains robust across varying levels of heterogeneity while maintaining expected nominal accuracy, reducing overestimation, and improving parameter recovery and predictive performance.

Sample Size for Estimating Disease Prevalence in Free-Ranging Wildlife Populations: A Bayesian Modeling Approach

James G. Booth[1], *Brenda J. Hanley*[1], *Florian H. Hodel*[2], *Christopher S. Jennelle*[3], *Joseph Guinness*[1], *Cara E. Them*[4], *Corey I. Mitchell*[5], *Md Sohel Ahmed*[1], *and Krysten L. Schuler*[1]
1Cornell University, Ithaca, NY, USA
2Michigan State University, East Lansing, MI, USA
3Minnesota Department of Natural Resources, Saint Paul, MN, USA
4Cara Them Consulting, LLC, Corvallis, OR, USA
5Desert Centered Ecology, LLC, Tucson, AZ, USA

A two-parameter model and a Bayesian statistical framework are proposed for estimating prevalence and determining sample size requirements for detecting disease in free-ranging wildlife. Current approaches tend to rely on random (ideal) sampling conditions or on highly specialized computer simulations. The model-based approach presented here can accommodate a range of different sampling schemes and allows for complications that arise in the free-ranging wildlife setting including the natural clustering of individuals on the landscape and correlation in disease status from transmission among individuals. Correlation between individuals and the sampling scheme have important consequences for the sample size requirements. Specifically, high within cluster correlations in disease status can reduce sample size requirements by reducing the effective population size. However, disproportionate sampling of small subsets of subjects from the greater target population, combined with high correlation of disease status, tends to inflate sample size requirements, because it increases the likelihood of sampling multiple animals within the same highly correlated clusters, resulting in little additional information gleaned from those samples. Our results are consistent with those generated using both previously established approaches and extend their ability to adapt to additional biological, epidemiological, or societal sampling complications specific to wildlife health.

Reference: Journal of Biological, Agricultural and Environmental Statistics, 29(3):438-454 (2024)

Bayesian Hierarchical Modeling of Fusarium Head Blight Epidemics Using Zero-Inflated Data and Environmental Covariates

Wanderson B. Moraes[1], *Laurence V. Madden*[1], *Kelsey F. Andersen*[1], *Christina Cowger*[2], *Ruth Dill-Macky*[3], *and Pierce A. Paul*[1]

[1]Department of Plant Pathology, The Ohio State University, Wooster, OH 44691, USA
[2]Department of Plant Pathology, North Carolina State University, Raleigh, NC 27695, USA
[3]Department of Plant Pathology, University of Minnesota, Saint Paul, MN 55108, USA

Fusarium head blight (FHB) is a severe cereal disease impacting wheat yield and grain safety due to associated mycotoxins, particularly deoxynivalenol (DON). Traditionally, FHB index (“field severity”) is quantified as a continuous response variable between 0 and 1, frequently containing inflated zeros due to absence of disease symptoms. Standard modeling approaches often struggle to accurately interpret such zero-inflated data, compromising reliable inference on environmental and treatment effects.

We propose a Bayesian hierarchical modeling framework specifically designed to handle zero-inflated continuous responses and effectively assess plant disease epidemics across multiple locations and years. Our model adopts a zero–one inflated beta (ZOIB) approach, which extends a two-part (hurdle) framework to include discrete probabilities for no infection (index=0) and complete infected spikes (index=1), along with a beta-distributed component for partial infection (0 < index < 1). This structure overcomes the limitations of standard beta models by explicitly handling boundary values. Hierarchical modeling naturally incorporates multiple layers (cluster-within-plot, plot, experiment, state, and year) as random effects, improving model accuracy and inference robustness.

Using extensive multi-state, multi-year wheat plot data, we evaluate the effects of pre-anthesis simulated rainfall treatments, which strongly influence disease progression. Bayesian hierarchical methods enable effective incorporation of additional environmental covariates (temperature, relative humidity, and natural rainfall), reducing confounding and revealing nuanced relationships between rainfall patterns, FHB index, and subsequent mycotoxin contamination.

This methodology facilitates a deeper biological understanding of how environmental factors influence disease epidemics and grain contamination with DON. By explicitly handling zero-inflation and integrating hierarchical structures, our Bayesian approach enhances predictive accuracy and supports precise risk assessment, leading to improved plant disease management strategies.

Reporting and Analyzing Agricultural and Ecological Levels as Continuous Data: The Case of Grass-Legume Mixtures

***Nicolas Caram*[1]**, *Lynn E. Sollenberger*[2], *Edzard van Santen*[3]
[1]Facultad de Agronomía, Universidad de la República, Paysandú, Uruguay
[2]Agronomy Department, University of Florida, Gainesville, FL, USA
[3]Agronomy Department and IFAS Statistical Consulting Unit, University of Florida, Gainesville, FL, USA

The global pressure for sustainable development of agricultural systems and preservation and restoration of ecosystems requires optimizing resource-use efficiency, while mitigating environmental impacts. Achieving this goal demands more efficient designs, innovative research questions and statistical analyses that enhance inference and applicability at larger scales. Specifically, we propose shifting from discrete treatments or manipulations to continuous levels that enable more precise assessments of agroecological responses and inferences in experimental and observational studies. This transition requires a thorough literature review of previous research to refine designs and levels that improve statistical inference. In cases where agroecological questions have already been addressed using discrete treatments or manipulation levels across various sites and environmental conditions, future studies should focus on refining the understanding of responses and mechanisms along continuous gradient levels. This allows researchers to identify optimal or, preferably, minimum levels that maximize desired outcomes, improving resource use efficiency and guiding targeted interventions. Statistical tools addressing this aspect include linear and non-linear models within frequentists or Bayesian frameworks, and model averaging for more robust inferences. Reporting these optimal levels facilitates meta-analyses and synthesis studies that explore how environmental conditions and climate change influence agroecological mechanisms and responses at large scales. This approach has been widely used in nitrogen (N) fertilization rate studies but remains less explored in other agricultural and ecological contexts. Here, we review the case of legume proportion in grass-legume mixtures as an example elucidating the benefits of assessing treatment variables, such as legume proportion in the forage, as continuous rather than discrete. While there is extensive research across the globe confirming that incorporating legumes (analyzed as discrete) into only-grass systems can replace inorganic N fertilizers and improve forage nutritive value and animal efficiency, simply incorporating legumes does not guarantee optimal benefits. Reviewing limited data on tropical and subtropical grass-legume mixtures analyzing legume as a continuous gradient, we found that biomass production, animal performance and emission intensity are highly sensitive to shifts from an optimal legume proportion. Insufficient or excessive legume proportions can lead to substantial productivity losses and environmental impacts, associated with suboptimum biomass and N yield, animal productivity and emission intensity. Thus, we encourage future research to focus on identifying optimal levels of agroecological variables by shifting from discrete to continuous experimental and observational designs, enhancing resource-use efficiency, improving statistical inference, and strengthening management strategies.

Rice Monitoring Using Multispectral sUAS Imaging in Breeding Trials

***Kevin Carrillo*[1]*,** *Christian De Guzman[2], Danny McCarty[2], Austin Fruge[2], Helen Ellenburg[2], and Aurelie Poncet[1]*

[1]University of Arkansas, Fayetteville, Arkansas, USA

[2]University of Arkansas system Division of Agriculture, Rice Research and Extension Center, Stuttgart, Arkansas, USA

Traditional rice phenotyping methods are labor-intensive and time-consuming, creating challenges for the assessment of large genotypic numbers and variations among growth stages. Integrating remote sensing into breeding programs can enhance decision-making accuracy and efficiency in selecting superior genotypes while minimizing manual labor. Typically, multispectral images are captured at key phenological stages using small unmanned aerial systems (sUAS) to maximize spatial resolution and calculate vegetation indices (VIs). However, statistical analysis is often performed separately for each time point, leading to suboptimal model performance. This study aims to exploit the temporal relationships across remote sensing images to improve the accuracy and predictability of grain yield trait estimation in breeding trials.

A randomized complete block rice breeding trial with three replications was established in Stuttgart, Arkansas, to evaluate 200 advanced lines. Multispectral sUAS data were collected weekly over eight flights spanning from panicle initiation to flowering. Raw images in the green, red, red-edge, and near-infrared bands were stitched using Pix4Dmapper, where the flight altitude of 30 meters resulted in a ground sampling distance of 3 cm for the calculated orthomosaics. Agronomic traits such as days to heading, plant height at maturity, and grain yield were documented in the field.

The raw digital numbers were calibrated to the environmental conditions of the first flight using a uniform control, and the following VIs were calculated: Normalized Difference Vegetation Index (NDVI), Soil-Adjusted Vegetation Index (SAVI), Normalized Difference Red Edge Index (NDRE), and Green Chlorophyll Index (GCI). Values from each VI and flight were matched with field plot data. Statistical descriptors, including mean, quantiles, standard deviation, and coefficient of variation were then used to characterize the central tendency and variability of VI values within each plot. Functional principal component analyses (fPCA) were conducted to reduce data dimensionality and identify key spectral patterns associated with grain yield traits. K-means clustering was applied to the fPCA outputs to classify genotypes into distinct groups based on spectral and agronomic characteristics. The optimal number of clusters was determined using the gap statistics, ensuring robust and meaningful groupings. Correlation analyses between agronomic traits and the k-means groups allowed the development of a genotype yield performance scale.

The results validated the potential of using multispectral sUAS time-series data to establish a framework for identifying the best rice genotype candidates for the release of new varieties. Ongoing research will aim to further refine remote sensing-based high-throughput phenotyping approaches to improve rice genotype selection, with an emphasis on increasing resilience to abiotic and biotic stresses in varying environments.

Estimating Genetic Gain – Correctly, Poorly, or Badly?

Philip Dixon[1]

[1] Iowa State University, Ames, IA USA

Plant breeding is one of many reasons for the huge increase in corn yields from 1920 to now, but separating genetic gain from confounding factors is difficult. Don Duvick, of Pioneer Hy-Bred / Corteva / ISU, popularized the eRA study design to estimate the contribution due to plant breeding. Seeds from genotypes released in different years are planted in the same year and site and managed identically.

Duvick used a two-stage analysis: estimate the mean yield (or other trait) for each genotype after accounting for the experimental design, then regress the mean yield on year of release. The slope estimates the genetic gain per year. Recent analyses (numerous published papers) have replaced the genotype means with the genotype BLUPs from a random genotype model.

I show that using BLUPs in a two-stage analysis systematically underestimates genetic gain. Duvick's original proposal, a two-stage analysis using genotype means provides unbiased estimates of genetic gain and its standard error when the eRA study design is balanced. A one-stage analysis, including the year of release, a random effect for genotypes, and the experimental design, also provides unbiased estimates of genetic gain and its standard error but requires access to the observation-level data. I will discuss extensions of the two-stage analysis to unbalanced data and multi-environment eRA trials. The various analyses are illustrated using simulated data and a study of root traits.

Modeling Considerations and Challenges for Over-Dispersed Count Data in a Generalized Linear Mixed Model

Susan L. Durham[1]*, Nora M. Bello*[2]*, Daniel G. Palmer*[23] *and Walter Stroup*[4]
[1]Utah State University, Logan, UT, USA
[2]United States Department of Agriculture – Agricultural Research Service, USA
[3]University of North Dakota, Grand Forks, ND, USA
[4]University of Nebraska-Lincoln, Lincoln, NE, USA

Within the Generalized Linear Mixed Model (GLMM) ecosystem, count data pose their own set of unique challenges due to the very common phenomenon of overdispersion. Overdispersion may be apparent due to the misspecification of the linear predictor (e.g., missing covariates or nonlinear terms, excess zeros, outliers, incorrect link) or inherent due to distributional choices. Our focus here is on alternative plausible distributions that reflect the generation process of count data and that may be used to accommodate inherent overdispersion; we illustrate using a data example.

Consistent with other non-Gaussian GLMMs, we start by addressing the choice of inference space and inference target based on the research objective; this choice is necessitated by the non-equivalent natures of marginal and conditional models. We consider it a true marginal model when inference targets the mean response of the population. When inference instead targets the median response of the population, representing a “typical” population member, we examine conditional GLMMs with Poisson-normal, negative binomial (aka Poisson-gamma), and generalized Poisson distributions. In all cases, we address appropriate methods of estimation and subsequent implications for inference. We emphasize that selection of a “best-fitting” distribution is intelligently driven by consideration of data generating mechanisms, the overdispersion mechanism, and the experimental/sampling design.

We compare results obtained using off-the-shelf software on SAS and R platforms, as well as in-house custom scripts. When appropriate, we highlight misalignment of results among software platforms, even when coding was intended to implement the same methodological approach.

Conditional Simulation of Gaussian Random Fields

Somak Dutta[1]*, Debashis Mondal*[2]
[1]Iowa State University, Ames, IA, USA
[2]Washington University, St. Louis, MO, USA

In spatial analysis, conditional simulation of spatial variables at unobserved locations given the data at the observed location facilitates various statistical inferences but suffers from computational scalability when the sample size is large. In this paper, we develop a method for conditional simulation based on novel mathematical decompositions of the inverse-covariance matrix. The method applies to a broad class of spatial models, including the Gaussian Markov random fields, fractional Gaussian fields, and the Matérn models. Matrix-free computational techniques are also developed for scalability. I will describe a practical application to mapping groundwater arsenic exceedance regions.

Zero-Inflated Models: Best Practices for Diagnostics and Model Fitting Across Software Platforms

Conor Fair[1], Julia Piaskowski[2], Raúl Macchiavelli[3], Josefina Lacasa[4], Bruce A Craig[5], and Walter Stroup[6]
[1]University of Georgia, Griffin, GA, USA
[2]University of Idaho, Moscow, ID, USA
[3]University of Puerto Rico, Mayagüez, PR, USA
[4]Kansas State University, Manhattan, KS, USA
[5]Purdue University, West Lafayette, IN, USA
[6]University of Nebraska-Lincoln, Lincoln, NE, USA

Excessive zeros are a common problem with count data that include cases that may not experience the event being measured; excess zeros cannot be handled properly by standard distributions (e.g., Poisson, negative binomial, etc.), resulting in incorrect inference. Therefore, scenarios with excessive zeros call for the use of zero-inflated models to account for both data generating mechanisms, namely the binary presence-absence and the observed counts. From a diagnostic perspective, it is crucial to properly identify evidence for zero-inflation based on the predicted number of zeros, as determined by parameter estimates in the fitted model. A model-based approach is preferred over approaches that only consider the number of observed zeros in the raw data, as the latter tend to inflate false positive rates for testing for zero-inflation. We illustrate implementations for diagnosing zero-inflation in SAS and R. Specifically, we found the R packages **DHARMa** and **performance** to be well-documented and straight forward to use. Further, we describe best practices in fitting zero-inflated GLMMs across software platforms using proc **nlmixed** in SAS and the package **glmmTMB** in R. We also found it important to evaluate different sources of the zero inflation beyond the intercept when it has not been identified *a priori*. While R and SAS are both able to handle high model complexity, there nevertheless remains gaps in evaluating zero inflation in generalized linear mixed models.

Assessment of Crop Herbicide Injury with Image Segmentation, Probability Density Functions, and Machine Learning

Wesley France[1], *Aurelie Poncet*[1], *Thanh Bui*[2], *Mario Soto*[1] *and Cengiz Koparan*[3]
[1]University of Arkansas, Department of Crop Soil and Environmental Science, Fayetteville, AR, USA
[2]University of Arkansas, Department of Computer Science and Computer Engineering, Fayetteville, AR, USA
[3]University of Arkansas, Department of Agricultural Education, Communications, and Technology, Fayetteville, AR, USA

Scouting for stress provides vital information for crop management. Crop stress may manifest different symptoms at varying intensities. While imaging sensors can provide insights into crop development, teasing out the differences between stress symptomology is challenging. In this poster, we present a method for quantifying different types of herbicide injury symptoms in soybean.

A completely randomized small-plot soybean trial was established in Fayetteville, Arkansas to evaluate visible stress in response to herbicide application. Plots were treated with three herbicide chemistries expected to manifest unique symptomology. Two rates were applied to generate different injury intensity. Chlorosis (CHL), necrosis (NEC), biomass (BIO), and overall injury (OI) ratings were collected 21- and 28-days following treatment application. At the same time, a total of 30 overlapping red, green, and near-infrared (RGN) images of each plot were collected using a multispectral sensor mounted 1.5 m above ground on an aluminum frame. The red and near-infrared images were used to calculate the normalized difference vegetation index (NDVI).

First, generalized linear mixed-effect analysis was performed to characterize soybean symptomology in response to the herbicide treatments and rating dates. Separate models were computed for CHL, NEC, BIO, and OI. The results showed that the treatments created a gradient of herbicide injury comparable to that found in literature. Regression analysis was then performed to describe overall injury as a function of CHL, BIO, and NEC. The model descriptive performance was high, indicating strong dependencies between OI and other symptomology. Therefore, OI was excluded from further analysis.

Image segmentation and classification were conducted to remove soil and shadows. The empirical probability density function of the NDVI values of vegetation were calculated for each plot and rating date. For each treatment, one replicate was selected at random and set aside for validation. Relief-f feature selection was computed to identify the NDVI values that explain most of the variability in CHL, NEC, and BIO, separately. These NDVI values were used to reduce the dataset dimensionality and fit random forest models that estimate herbicide symptomology ratings from the segmented RGN images. The random forest residual root-mean-square-error (RMSE) values ranged from 4.7% to 8.7%, which was within the acceptable range of error as weed scientists are trained to rate injury within a 10% range. The fitted random forest models were then applied to the validation data. The predicted RMSE ranged from 9.1% to 20.1%. The predicted error did not fit within the desired 10% likely due to the paucity of the data. However, the goal of this research was proof-of-concept, and findings did demonstrate that the proposed methods could be used to quantify specific herbicide injury symptomology in soybean given a larger dataset size.

The Importance of Data Quality

***Julian Garcia-Abadillo*[1,2], *Diego Jarquin*[1]**
[1] Visitor Scholar – Agronomy Department, University of Florida, Gainesville, FL, USA
[2] PhD Candidate – Centre for Biotechnology and Plant Genomics, Universidad Politécnica de Madrid, Madrid, Spain

Data is often described as the fuel that powers the cutting-edge information technologies that are directly impacting the world. However, like raw fuel, data must undergo refinement to become truly valuable and reliable. In this context, standardized data collection protocols, preprocessing pipelines, and quality control are essential for maximizing its utility. However, these processes have not received the same level of attention as advancements in modeling and statistical learning techniques or performance benchmarking evaluations. In this presentation, we describe different stages in the data life cycle in which quality can be compromised for different reasons, ranging from technical and human-made errors to epistemological contradictions. The importance of data quality in practical agronomy implementations is exemplified through the prediction of plant phenotypes using genomic and environmental predictors. This plant breeding application, built under a supervised machine learning framework, will be used to show real-word examples of both poor and best practices in data analysis and result interpretation, especially in scenarios where operational decisions are typically data-driven.

Improving Wheat Genomic Prediction Through Random Regression and Optimized Sowing Dates

Guillermo Sniadower[1], *Rishap Dhakal*[2], *Paula Silva*[3], *Bettina Lado*[1], *Pablo Sandro*[2], *Inés Rebollo*[4], *Martin Quincke*[2], *Lucia Gutiérrez*[2], ***Pablo González Barrios***[1]

[1]Facultad de Agronomía, Universidad de la República, Av. Garzón, 780, Montevideo 12900, Uruguay

[2]Department of Plant and Agroecosystem Sciences, University of Wisconsin-Madison, 1575 Linden Drive, Madison, WI 53706, USA

[3]Instituto Nacional de Investigación Agropecuaria (INIA), La Estanzuela, Sistema Agrícola-Ganadero, 70006, Colonia, Uruguay

[4]Department of Agronomy and Plant Genetics, University of Minnesota, 1991 Upper Buford Cir Borlaug Hall, Saint Paul, MN 55108, USA

Wheat (*Triticum aestivum* L.) supplies 20% of global dietary protein and calories, making it a vital crop for food security. However, climate variability increasingly threatens sustained production. Genomic selection is a powerful, proven tool for predicting genetic merit across varied crops. It employs whole-genome data to enhance breeding efficiency in multi-environment trials. It has revolutionized plant breeding by enabling early selection of superior genotypes, offering a significant increase in genetic gain over traditional methods. Despite its success, a major challenge persists in explaining non-genetic full effects, particularly genotype-by-environment interactions (GEI), which shift genotype rankings across environments. This study integrates environmental covariates (ECs) into random regression models (RRM) to improve genomic prediction in multi-environment trials. We analyzed 4,291 genotypes from Uruguay's National Wheat Breeding Program (2010-2020) across 59 environments defined by location, sowing date, and year, deriving 42 ECs from vegetative, reproductive, and grain-filling phases. Using partial least squares regression, five ECs—cumulative precipitation and cool temperatures during vegetative and reproductive phases, plus frost days—were selected to model GEI. These phase-specific ECs, including reproductive precipitation, vegetative frost, minimum temperatures below 15°C, mean maximum temperature, and vegetative precipitation, captured distinct yield effects, outperforming models without ECs. Variance components analysis showed that year (27.1%) and site-by-year (19.3%) were the dominant sources of variation, while GEI (11.2%) outstripped genotypic variance (5.9%), emphasizing the environmental impact on yield. In cross-validation (CV1, CV2, CV0), RRM incorporating one or two ECs outperformed genomic best linear unbiased prediction (GBLUP) by 72-100%, improving predictions in 35 of 59 environments. Factor analytic models achieved the highest prediction ability (e.g., 0.42 mean in CV2) when full data were available. The findings highlight the adaptability of different EC modeling strategies: RRM for sparse covariates, and factor analytic models for richer datasets. Furthermore, May-June sowing optimized yield under cool vegetative conditions, with favorable genotype-specific responses. This phase-specific, GEI-focused approach, combined with optimized sowing strategies, offers a robust and practical framework for climate-adapted wheat breeding, offering a flexible tool to enhance selection efficiency in variable environments.

Launching and Sustaining Statistical and Data Science Consulting: Planning and Assessing for Success

Emily Griffith[1]***, Mara Rojeski Blake***[1]

[1]North Carolina State University, Raleigh, North Carolina, USA

The Data Science Consulting Service at North Carolina State University Libraries provides support for the NC State community for requests spanning the entirety of the data science lifecycle. Data-driven planning and assessment has helped this program be successful by allowing us to provide the best service, ensure our student workers have the knowledge needed to assist with the wide-ranging requests our service receives, and promote the program to higher-level administrators at NC State.

The collaborative consulting efforts of the Libraries and NC State's Data Science Academy leverage graduate student data science consultants, as they help us provide and scale data science service to our campus while providing an important learning opportunity for the students. We have found that it also opens graduate students to new options for future work they may otherwise not have encountered in their disciplinary programs.

We will share our experiences developing this partnership and staffing model and the successes of our program. We also will offer recommended practices for others to adapt to their local institutions and environments.

Some of this work was done in collaboration with Marianne Heubner (Michigan State University), Steven Pierce (Michigan State University), Micaela Parker (Academic Data Science Alliance), and Rachel Levy (NC State University) for an article in *Stat*.

Using Historical Control Data to Assess Study Quality for Medaka Extended One Generation Reproduction Test (MEOGRT)

***Yushan Gu*[1]**, *Julie Krzykwa[2], Natalie Burden[3], Valentin. Mingo[4], Davood Poursina[1], Constance A. Mitchell[2], Edward R. Salinas[5], and James R. Wheeler[6]*

[1]Corteva Agriscience, Indiana, USA
[2]HESI, Washington D.C., USA
[3]NC3Rs, London, UK. natalie.burden@nc3rs.org.uk
[4]Corteva Agriscience, München, Germany
[5]Bayer AG, Monheim, Germany
[6]Corteva Agriscience, Bergen op Zoom, The Netherlands

The Medaka Extended One Generation Test (MEOGRT; OECD TG 240/EPA OCSPP 890.2200) is an in vivo assay designed to assess adverse effects and endocrine-relevant endpoints through key stages of the fish life cycle. This assay requires a significant number of laboratory animals, making it essential for the test to be reliable and robust. The exposure begins with spawning fish (F0), includes multiple life stages of their progeny (F1), and continues until two weeks post-fertilization in the second generation (F2). Parameters such as survival, growth, sex ratio, and reproduction are evaluated for potential adverse effects.
This study analyzed control data from 25 control groups from 24 independent studies conducted following the MEOGRT test initially adopted in 2015, or providing similar data as in the MEOGRT, including 14 Medaka Multigeneration Tests (MMT) performed prior to the adoption of TG 240. Summary statistics of historical control data for all endpoints were provided. The attainability of current validity criteria, especially those related to reproduction endpoints, was analyzed using Logistic Regression or Gaussian models. Cross-laboratory and study variability were examined using fixed effect models, as random effect models and regular variance component analysis were not applicable. Additionally, simulations were used to estimate the test design's empirical type I error or false positive rate. The primary goal is to develop a knowledge base that can enhance test design, performance, and data interpretation.

We strongly encourage submissions that feature innovative, provocative subjects that reflect outside the box thinking. We want attendees to be exposed to diverse viewpoints that broaden understanding, challenge existing assumptions, and foster innovative thinking.

Generating High-Quality Simulated Data Using Process-Based Crop Models for AI Applications

***Rishabh Gupta*[1]**, *Satya K. Pothapragada[2], Joel B. Harley[2], Alina Zare[2], Lincoln Zotarelli[1]*

[1]Horticultural Sciences Department, University of Florida, Gainesville, FL, USA

[2]Department of Electrical & Computer Engineering, University of Florida, Gainesville, FL, USA

Despite the remarkable advancements in artificial intelligence (AI) and its powerful estimation capabilities, the complexities inherent of agricultural systems pose recurrent challenges, primarily due to limited field observations and lack of standardized organization in field trial data. The constraint of limited observations is unlikely to be resolved soon, as collecting accurate field data remains costly, labor-intensive, and time-consuming process. This led us to explore a novel approach: a cropping system model (CSM)-informed AI model. This approach involves training AI models using data generated from CSM. However, generating high-quality CSM data requires precise CSM calibration using field observations which brings us back to the fundamental issue of data organization and standardization. Hence, our collaborated efforts led us to develop a structured data template for organizing field experiment data to enhance the accessibility and usability of the data for modeling purposes. This development led us to effectively organize in-situ data from our old field trials conducted from 2011-2014 which subsequently was utilized to calibrate CSM for generating high-quality data to train AI models for agriculture system interpretation aiming to further improve crop nutrient recommendation. The data template eased the process of preparing input files for the DSSAT crop model (SUBSTOR-Potato) which was calibrated and evaluated using sparse data collected from field trials such as plant/tuber biomass and nitrogen (N) content, and soil N concentration. Afterward, approximately 4.5 million hypothetical potato management scenarios were created, such as varied planting dates, N fertilizer application rates and timings, irrigation strategies subjected to historical weather conditions from 2001-2024. The DSSAT generated (~700 million data rows) plant/tuber biomass, soil N, and N leaching. Subsequently, multiple AI models were trained using generated data to estimate the tuber growth, soil N concentration, and N leaching, to eventually be utilized for providing nutrient application rates and timings recommendation.

Random Regression Finlay-Wilkinson Models for Yield and Stability Prediction in Cereals

Pablo Sandro[1], Justin Blancon[2], and ***Lucia Gutierrez[1]***
[1]University of Wisconsin - Madison, Madison, WI, USA
[2]INRAe, Clermont Ferrand, France

Climate change poses a challenge for agriculture by increasing climatic variability and compromising crop yields, therefore, breeding for genotypes that can adapt to climate change is needed. One of the most important challenges for stability evaluations is that most stability indices require complete or balanced datasets and evaluations in a large number of environments, while most breeding programs typically generate sparse multi-environment datasets. Methods such as Finlay-Wilkinson random regression can deal with sparse datasets and incorporate genomic data to leverage phenotypic information from related genotypes to predict yield and yield stability. Our objective was to compare Finlay-Wilkinson random regression with genomic GBLUP models to predict yield and yield stability, and to evaluate the impact of the number of environments and the environmental variance on those predictions. We used three large datasets, one highly unbalanced dataset for oats, and two completely balanced datasets with a different number of environments in barley and wheat. We ran classic and random regression Finlay-Wilkinson models in various scenarios where we predicted stability and grain yield. The Finlay-Wilkinson random regression model had higher predictive ability for grain yield for un-phenotyped genotypes (i.e. new genotypes) compared to the genomic GBLUP model in both balanced and unbalanced datasets. Because the Finlay-Wilkinson random regression model borrows information from relatives, the number of environments needed for a high predictive ability was lower than using the classic Finlay-Wilkinson models. On the other hand, pairs of contrasting environments, that created large among environment variance, consistently yielded high predictive ability compared to classic Finlay-Wilkinson models, indicating that a smaller number of contrasting environments could be used as training populations when random regression models are used.

Envirotyping-Informed Mixed Models to Study the Climatic Drivers and Yield Seasonal Variation for Common Beans in Brazil

***Alexandre Bryan Heinemann*[1]*, David Henriques da Matta*[2] *and Luís Fernando Stone*[1]

[1]Embrapa Arroz e Feijão, Santo Antônio de Goiás, Goias, Brazil

[2]Universidade Federal de Goias, IME, Goiânia, Goias, Brazil

Common beans (Phaseolus vulgaris L.) are a staple food crop cultivated across various regions, seasons, and management systems in Brazil. To ensure production stability, it is essential to understand how climate factors affect cultivar development. This study aimed to determine the main edaphoclimatic drivers influencing the seasonal variation of common bean yield and their impact on genotype ranking across Brazil. Utilizing extensive databases, such as historical field trial records, allows for deeper insights into the impacts of environmental features on phenotypic variation, guiding plant breeders in addressing genotype-by-environment interactions that limit cultivar targeting and genetic progress. We applied an envirotyping-informed (EI) linear mixed-effects model (LMM) to assess climatic drivers and their effects on yield variation across diverse years, elite genotypes, and regions. Our approach was implemented in three steps. Step 1 entails collecting agronomic variables from the EMBRAPA's dry bean breeding program trials and linked with climate and soil features to create the data set, hereafter named only as "Environmental Features (EF)". Step 2 entails the selection of environmental characteristics using the stepwise method based on the Bayesian Information Criterion (BIC). At the end of the stepwise method, non-significant covariates (with p-value > 0.05) were removed from the adjusted envirotyping-informed linear mixed-effects model" (EI-LMM). Finally, step 3 entails the estimation of the percentage contribution of the environmental features (EF), the predicted genotypic potential (RPIPY, %), and the genotype relative importance (GRI). Our findings identified distinct seasonal environmental types within each region. Air temperature emerged as a key factor, explaining 40% to 80% of the phenotypic variation in grain yield. The Midwest region, where the main breeding nursery is located, is primarily limited by temperature, while other regions, such as the Southeast, exhibit different factors affecting yield variations. The inclusion of EI-LMM enabled cultivar ranking based on genetic mean incremental predict value and the calculation of genotype relative importance using analysis of variance (ANOVA). These outcomes connect data from advanced breeding trials and inform decisions about cultivar development, considering regional environmental specificities and within-season variations. Future studies should incorporate genotype-by-environment-by-management interactions to better understand climate adaptation in common beans, bridging the gap between breeding efforts and farmer needs.

Keywords: Environmental profiling; Linear mixed-effects model; Adaptation; Precision Breeding; Climate; Enviromics; Phaseolus vulgaris L.

Allowing Negative Variance Component Estimates in REML: Inferential Consequences for Fixed Effects and Type I Error Control

Bipin Poudel[1]*,* ***Réka Howard***[1]*, Nora M. Bello*[2]*, Walt Stroup*[1]
[1] University of Nebraska – Lincoln, NE
[2] US Department of Agriculture, Agricultural Research Service, Northeast Area - Beltsville, MD

This study examines the inferential implications of allowing for negative variance component estimates in mixed model analyses using Restricted Maximum Likelihood (REML). Specifically, we make comparisons with the common practice of constraining variance estimates to be non-negative, thus effectively dropping random effects from the linear predictor when variance estimates are set to zero when estimating equations produce a negative solution. We show this common practice to be ultimately undesirable as it yields a model specification that misrepresents the data generation process. We further evaluate the impact of these constraints on the variance estimates on practical inference in the context of common multi-level experimental designs. Specifically, we explore inferential implications on significance levels for fixed effects and on estimation of standard errors and confidence intervals. Consistent with the limited prior research on this topic, our findings indicate that constraining variance component estimates to be non-negative inflates the Type I error rate when testing the significance of fixed effects. Similar inferential implications are apparent from standard errors and confidence intervals. This study underscores the need for proper model specification that adequately reflects the data generation process, and its corresponding implementation in statistical software packages. Ultimately, we advocate for best practices in the implementation of mixed model analyses to ensure accurate and reliable inference.

ANOVA P-Values and Confidence Sets via Monte Carlo

***Benjamin L. Jacobs*[1]**, *Peng Liu*[1]

[1] Department of Statistics, Iowa State University, Ames, IA, USA

In the traditional exposition of fixed-effects linear model theory, confidence intervals for linear combinations of the regression coefficients are derived by transforming the confidence distribution of the coefficients, which is a multivariate T-distribution. In contrast, the ANOVA test is derived from the sampling distribution of the ratio of the mean squares, which under the null has an F-distribution. Confidence sets for the parameters of interest can be derived by pivoting the ANOVA test statistic but are conceptually de-emphasized in this approach. This switch from the confidence distribution to the sampling distribution creates a conceptual rift in the theory of the two methods. This can obscure how the ANOVA test logically relates to the confidence intervals.

I will provide an alternative explanation of the ANOVA p-value by directly transforming the confidence distribution of the regression coefficients. In this approach, the ANOVA p-value can be interpreted as 1 minus the size of the smallest origin-containing central confidence region for the difference in fitted means of two nested models. Such an explanation unifies the conceptualization of what ANOVA is doing with the conceptualization of what the confidence intervals are doing. A Monte Carlo algorithm for computing ANOVA p-values via direct simulation from the confidence distribution of the coefficients is provided. This algorithm makes the interpretation more concrete. It also allows computation of the p-value without directly reckoning degrees of freedom or having to consider the F-distribution. The F-distribution itself fades into the background, as all calculations are done in terms of the multivariate T-distribution and the ANOVA test statistic itself is interpreted in terms of the radius of the confidence region. Connections with objective Bayesian inference will be drawn, and an example with a real data set will be provided.

A Guide for Implementing Linear Mixed Models in R for Common Agricultural Experimental Designs

***Harpreet Kaur*[1]** *and Julia Piaskowski*[1]
[1]Statistical Programs, University of Idaho, Moscow, ID, USA

Linear mixed models have long been the established standard for analyzing experimental data. While agriculture scientists increasingly rely on the R programming language for statistical analysis, the implementation of mixed models for complex designs using R is challenging. This is due to the broad landscape of modelling packages and the lack of centralized educational effort. To address this gap, we have created an online tutorial, the "Field Guide to the R Mixed Model Wilderness", that provides guidance on implementation and of LMMs for common experimental designs using R. The intended audience is researchers who have some knowledge of frequentist statistics, LMMs and R programming, but limited experience in implementing LMMs in R. This guide covers the application of LMMs using "lme4" and "nlme", two widely used packages for mixed modeling in R. The experimental designs addressed in this tutorial are randomized complete block (including factorial designs), split-plot, split-split plot, strip-plot, incomplete-block and experimental designs with repeated measures. The application of lme4 and nlme packages is demonstrated through examples from previously published agriculture experiments. The tutorial also includes details on data integrity checks to conduct before and after analysis to confirm adherence to LMM model assumptions and a troubleshooting guide for how to handle common problems like unequal variance, departures from normality, and non-convergence. With this guide, we hope to bolster the quality of statistical analysis among agricultural scientists and anyone else that may benefit from understanding how to properly implement LMMs in R.

Aggregation of Yield Data from Thyrow Long-Term Experiments as a Basis for a Meta-Analysis Using Winter Rye (Secale Cereale L.)

***Bärbel Kroschewski*[1]**, *Gemma Dini*[1] *and Konrad Neugebauer*[1]
[1]Humboldt-Universität zu Berlin, Berlin, Germany

Long-term experiments (LTE) are designed to gain insights into slowly changing soil processes and thus contribute to a better understanding of the long-term effects of agricultural measures.
The joint analysis of data from several LTEs in the framework of a meta-analysis can lead to even more valuable results. The results of similar studies can gain in statistical power and new questions can be answered that were not part of the original research projects.

When aggregating LTE data from different sites and/or years, treatment effects may be confounded with site and year effects. In Thyrow, an experimental station of the Humboldt University near Berlin, several long-term experiments are conducted. If selected treatments of these LTEs are aggregated, the site and weather conditions in the same year are largely identical, and possible differences in yield should then be mainly due to differences in treatment. For the aggregation, winter rye was selected as the test crop because it is grown in six of the Thyrow long-term trials in every fourth year. The grain and straw yield for six years in the period 2002–2022 were examined. The main question is whether similar treatments from different LTEs lead to comparable grain and straw yields.

The aggregation was carried out in the following steps:

- Selection of treatments,
- Definition of a criterion for grouping the treatments,
- Assignment of treatments to the variant groups,
- Evaluation of the assignment (visualization, analysis of variance components),
- If necessary, modification of the assignment by adjusting the assignment criteria.

The main criterion was the gradient of mineral nitrogen fertilization in combination with organic fertilization, followed by preceding crop, crop rotation, and others. Using suitable graphics, it was examined whether the same treatments in the same experimental year lead to largely the same yields and whether the gradient behind the variant groups is reflected in increasing yields.

A subsequent variance component decomposition using a random effects model (year, LTE, variant group and all two- and three-way interactions) showed that 75% of the variation in grain and straw yield was determined by variant group, followed by year (6%) and LTE (2-3%). The amount of interaction between the three factors was very low, so that the grouping of treatments can be considered largely independent of year and LTE.

Finally, a joint analysis of all variant groups is carried out in the framework of a network meta-analysis. If the treatments occur together in all LTEs, their differences can be assessed by direct comparisons. If this is not the case, indirect comparisons must be carried out using a common reference treatment. For the Thyrow site, the treatment that corresponds to the best practice at the site was defined as the reference treatment.

Handling Within-Block Heterogeneity Using Smoothing Splines

***Josefina Lacasa[1]**, Susan L. Durham[2], Nora M. Bello[3] and Walter Stroup[4]*
[1]Kansas State University, Manhattan, KS, USA
[2] Utah State University, Logan, UT, USA
[3]United States Department of Agriculture – Agricultural Research Service, USA
[4]University of Nebraska-Lincoln, Lincoln, NE, USA

Spatial variability is a common feature of agricultural field experiments and may produce misleading results if not addressed. Agricultural field experiments often aim to address spatial variability using blocking strategies assumed to produce approximately homogeneous areas in space. However, when blocks are large in size, the assumption of within-block homogeneity often suffers. In this case, a researcher may adapt their analytic approach by modeling spatial correlations explicitly. Spatial correlations may be specified parametrically (e.g., with exponential or Gaussian correlation functions) or non-parametrically (e.g., with smoothing splines on the residuals). Here, we focus on non-parametric smoothing splines for scenarios in which the spatial process cannot be modeled with simpler, parametric correlation functions. We implement and compare three approaches to modeling spatial variability, namely (i) by specifying block effects only, (ii) by specifying smoothing splines on the residual surface, or (iii) by specifying smoothing splines on both residuals and expected values. We illustrate differences in the implementation of smoothing splines across software platforms using PROC GLIMMIX in SAS, and the mgcv and glmmTMB packages in R.

How are Large Language Models Changing the Analytics Industry?

James David Long[1]
[1]Palomar Specialty Insurance, La Jolla, CA, USA

Large language models (LLMs) are fundamentally transforming the analytics industry in waves of adoption that are reshaping how organizations derive insights from data. This presentation explores this transformation, highlighting current impacts, future trajectories, and critical pitfalls.

The first wave of LLM adoption has centered primarily on code generation and text writing assistance. This focus stems from LLMs' inherent advantages in this domain: they've been trained on billions of lines of code from diverse repositories, can recognize structural patterns across programming languages, and effectively translate natural language requirements into executable solutions. For analytics professionals, this has created unprecedented productivity gains. GitHub's studies indicate developers using AI assistance complete tasks approximately 55% faster than peers without such tools. McKinsey research shows coding productivity improvements of 20-45%, with analytics workflows particularly benefiting from this efficiency boost.

Simultaneously, LLMs have gained traction in text-heavy domains requiring sophisticated summarization and semantic search capabilities. Legal research, medical literature analysis, and market intelligence have pioneered these applications. The analytics sector has quickly adopted similar approaches for processing unstructured data sources that were previously difficult to incorporate into quantitative frameworks.

The emerging second wave of adoption promises even more profound changes. We're witnessing a shift toward domain-specific LLMs fine-tuned for specific industries and use cases. These specialized models are being integrated directly into analytics workflows. This democratization effect is allowing non-technical stakeholders to perform sophisticated analyses through conversational interactions. Organizations implementing these capabilities report significant reductions in time-to-insight, with some documenting 40-60% decreases in reporting cycles.

Despite their capabilities, relying on LLMs as "stores of knowledge" is fraught with significant problems. These models frequently hallucinate information, produce confident but incorrect responses, and struggle with certain types of logical and spatial reasoning tasks. The contrast between simple foundational models and Retrieval-Augmented Generation (RAG) approaches is critical. Used poorly, these tools hallucinate and provide misleading feedback; used properly as summarization tools, writing assistants, and knowledge-fetching agents, they can accelerate specific classes of analysis and human learning.

As a CTO of a specialty insurance company that includes crop insurance, I will focus this presentation on trends I see from my industry experience, highlight implementation challenges, and propose a framework for researchers to effectively consider LLMs in their analytics ecosystems.

Analysis of Repeated Measures Revisited: Are Recommended Best Practices Applied Consistently Across and Within Software Platforms?

***Raúl E. Macchiavelli*[1]**, *Nora M. Bello*[2] *and Walter Stroup*[3]

[1]University of Puerto Rico, Mayagüez, PR, USA

[2]United States Department of Agriculture, Agricultural Research Service, USA

[3]University of Nebraska-Lincoln, Lincoln, NE, USA

While the main methodological debates in the analysis of repeated measures for Gaussian data have been settled for more than 20 years, reliable inference depends heavily on bias corrections and good approximations to the distribution of the test statistics. Another critical issue to sound inference is the selection of a covariance structure, which in turn depends on the penalized likelihood criterion used to compare model fit to data. As illustrated in the data examples presented in this poster, we found that bias correction methods (e.g., Kenward-Rogers) and penalized likelihood criteria (e.g. CAIC, BIC) were not consistently defined across SAS and R software platforms, neither between R packages, nor between SAS procedures. Furthermore, the availability of a computationally stable suite of candidate covariance structures also differed between software platforms. Taken together, these inconsistencies caused substantial discrepancies in the model selected for final inference, and thus in the conclusions obtained. The inferential implications of software differences in mixed models' implementation were further exacerbated by the small sample size of the dataset analyzed relative to the number of repeated observations.

Measuring Spatial Consistency in Simulated Crop Yield with Tensor-Product Splines and Interclass Correlation

***Payton Miloser*[1], *Philip Dixon*[1]**
[1]Iowa State University, Ames, IA

Crop yield is spatially consistent when the high-yield areas of a field (or low-yield areas) occur in the same places from year to year. Using precision ag data from previous years to design a management plan for the current year assumes there is spatial consistency, but there are few statistical methods to assess this consistency, especially for data from more than two years. Generating simulated field data from a Gaussian process, we show the use of generalized additive models with tensor-product splines on a range of consistent to inconsistent spatial trends across multiple years of data. This avoids pairwise comparisons, expanding on current methodology used for precision ag data. Measuring this consistency is as important as identifying it, thus using interclass correlation coefficients we have the ability to detect small amounts of inconsistencies across the field and locate those for improvements in farming effectiveness.

Developing *paar*: An R Package for Yield Data Filtering and Management Zone Delineation in Precision Agriculture

Pablo Paccioretti[1,2]*, Mariano Cordoba*[2]*, Monica Balzarini*[2]*, Laila Puntel*[1,3] *and Guillermo Balboa*[1]
[1]University of Nebraska–Lincoln, Lincoln, Nebraska, USA
[2]National Scientific and Technical Research Council, Argentina
[3]Syngenta group, Digital, Basel, Switzerland

Precision agriculture aims to enhance crop management by leveraging data-driven insights to optimize yield and resource use efficiency. The technologies used in this approach collect hundreds of thousands of georeferenced data points within a single field, generating highly detailed yet complex datasets. The spatial correlation inherent in this data affects statistical analysis, making data processing and interpretation more challenging for decision-making. We present *paar*, an R package designed to streamline the analysis of yield data and automate and validate the delineation of homogeneous management zones in precision agriculture. The package integrates robust statistical methods and geospatial analysis tools to improve decision-making at the field level.

paar offers functions to automate error removal from yield maps and facilitate zone delineation using multivariate clustering techniques. It includes functions for pre-processing raw yield data, identifying and removing outliers, and generating cleaned yield maps that can be used for spatial interpolation. During depuration, observations close to field boundaries, global outliers, and spatial outliers can be removed. The zone delineation function applies clustering techniques that account for spatial correlation. Multivariate spatial analysis using principal components that explain a high percentage of total variability is used as input for clustering. A spatial principal component analysis (PCA) is performed, followed by a k-means method. Several statistical indices are provided to help determine the optimal number of zones. Furthermore, delineated zones can be validated by statistically comparing the mean value of each zone.

The automated workflow reduces the time required for data preprocessing, making advanced precision agriculture data analysis techniques accessible to a broader audience. The effectiveness of *paar* is demonstrated through case studies in diverse agricultural settings, showing improvements in yield data filtering and homogeneous zone delineation. The package is open-source and available on CRAN, encouraging contributions and adaptations to suit different agricultural contexts. By simplifying complex Precision Agriculture data Analysis tasks in R, *paar* empowers researchers and practitioners to make more informed decisions, ultimately enhancing productivity and sustainability in agriculture.

Mixed Models-Based Precision and Power Analyses for Design Comparisons Using PROC GLIMMIX and glmmTMB

Daniel G. Palmer[1,2], *Nora M. Bello*[2], *Josefina Lacasa*[3], *and Walter W. Stroup*[4]
[1]University of North Dakota, Grand Forks, ND, USA
[2]United States Department of Agriculture – Agricultural Research Service, USA
[3]Kansas State University, Manhattan, KS, USA
[4]University of Nebraska – Lincoln, Lincoln, NE, USA

In recognizing experimental design and statistical modeling as two sides of the same coin, we use a mixed models-based closed-form approach to illustrate precision and power calculations for designed experiments with applications in SAS and R. First, we identified candidate experimental designs for set resources, namely balanced incomplete blocks, control versus treatment, disconnected nested factorial (i.e. split plot), and “forced” randomized complete blocks. For each candidate design, we conducted precision analyses to assess standard errors for contrasts of interest and corresponding power calculations for a set sample size. Rather than a standard simulation-based approach, we used the strategy of retrofitting mixed models to leverage expected treatment effects and variance components held constant. We illustrate side-by-side implementations of mixed model-based precision and power analyses using PROC GLIMMIX in SAS and glmmTMB/emmeans in R. Both software platforms yielded the same results, so either may be used to inform practical steps in research planning. We discuss differences in software implementation and briefly mention syntax discrepancies between versions of glmmTMB. We finish by briefly discussing extensions to generalize linear mixed models.

Generalized Foldover Designs: A Class of Designs for Irregular Fractional Factorial Experiments with Bias Protection Properties

***Vinny Paris**[1], Max D. Morris[1]*
[1] Iowa State University, Ames, IA, USA

Fractional factorial experiments are often used when the total number of runs in the full factorial experiment is too large. This introduces bias if higher order interaction effects are unaccounted for in the model. The original foldover method proposed by Box and Wilson allows for complete dealiasing of main effects from unaccounted for second-order interactions in fractional factorial experiments. A new class of designs for irregular fractional experiments based on a generalization of the foldover method will be presented. By restricting the design space, this new class of designs can dealias main effects from a subset of unaccounted for second-order interactions. Simulations will be presented as well as a brief comparison of the currently tabulated D-optimal generalized foldover designs against D-optimal designs from an unrestricted design space. This will be a continuation and update of the work presented at the Utah State 2023 incarnation of this conference.

Using Randomized Quantile Residuals to Assess Model Fit of Generalized Linear Mixed Models

Julia Piaskowski[1]

[1]Statistical Programs, University of Idaho, Moscow, ID, USA

Assessing if a generalized linear mixed model (GLMM) is appropriately specified for a given data set and data generating process is challenging since often GLMMs have no defined distribution for model residuals. Detecting deviations from model expectations (e.g. dispersion, zero-inflation) is difficult given that some non-normal distributions have mathematically defined mean-variance relationships, and they may have discrete outcomes that are less amenable to direct interpretation via plotting. Scaled residuals such as Pearson or studentized residuals can alleviate these challenges to an extent, but they also can be difficult to correctly interpret for over- and under- dispersion. Randomized quantile residuals are another option, simulating data for each data point using the fitted GLMM for the distribution and its parameters and obtaining the quantile for each observation from the cumulative distribution function of the simulated data. Thus, quantile residuals are a measure of how well the data conform to the specified model and distribution. Quantile residuals are expected to follow a uniform distribution [0,1] and can be interpreted similarly to common diagnostic residuals plots such as the q-q plot and the residual-versus-fitted values plot. Many hypothesis tests can be conducted using quantile residuals including tests for over- or under-dispersion, uniformity, zero inflation, and categorical dependence among the independent variables. This method is implemented with the R package “DHARMa”. Although implementations in other languages are not known, the method itself is relatively simple and can be written in other statistical programming languages. The purpose of this talk is to describe the process of generating and evaluating randomized quantile residuals for different aspects of fitted GLMMs. Randomized quantile residuals are a useful model diagnostic tool, but it should be noted that at this time, there is a lack of information on their performance and reliability across a broad range of conditions.

Assessing the Efficiency and Heritability of Blocked Tree Breeding Trials

Hans-Peter Piepho[1], *Emlyn R. Williams*[2], *and Maryna Prus*[1]
[1]University of Hohenheim, Stuttgart, Germany
[2]Australian National University, Canberra, Australia

Progeny trials in tree breeding are often laid out using blocked experimental designs, in which families are randomly assigned to plots and several trees are planted per plot. Such designs are optimized for the assessment of family effects. However, tree breeders are primarily interested in assessing breeding values of individual trees. This paper considers the assessment of heritability at both the family and tree levels. We assess heritability based on pairwise comparisons among individual trees. The approach shows that there is considerable heterogeneity in pairwise heritability, primarily due to the differences in both genetic as well as error variances among within- and between-family comparisons. Our results further show that efficient blocking positively affects all types of comparison except those among trees within the same plot.

Soybean Classification Using Time-Series Satellite Remote Sensing

Aurelie M. Poncet[1], *Ikram Morso*[1], *Jason A. Tullis*[1], *Ujjwal Sigdel*[2], *O. Wesley France*[1]
[1]University of Arkansas, Fayetteville, AR, USA
[2]University of Georgia, Athens, GA, USA

Effective and timely crop monitoring is essential to minimizing yield loss caused by stress. However, traditional scouting methods are limited in coverage and frequency. Satellite remote sensing and vegetation indices (VIs), such as the Normalized Difference Vegetation Index (NDVI), offer scalable solutions for monitoring crop health and development. While typical data analysis methods tend to focus on single time points, understanding temporal patterns is required to gain deeper insights into the processes that contribute to yield variability in production fields. The goal of this project was to characterize and classify in-field changes in crop development by analyzing spatial and temporal variations in crop NDVI trajectories.

A randomized complete block strip seeding rate trial with four replications was established in a production soybean field in Arkansas. The seeding rate treatments were: 185, 247, 309, 371, and 432 thousand seeds ha^{-1}. Soil samples and stand counts were collected every 0.5 ha to characterize in-field changes in soil fertility, soil texture, and crop emergence. Yield monitor data were collected at harvest. PlanetScope satellite imagery was downloaded for 72 of 156 days between planting and harvest and used to calculate NDVI. Time after planting was quantified using cumulative growing degree days (cGDD) calculated from public hourly air temperature data.

The field was divided into a fishnet grid with 861 square cells. Kriging, polynomial regression, or spatial intersection were used to quantify soil fertility, texture, yield, and cGDD-specific NDVI in each grid cell. The cGDD-specific NDVI values were interpolated every 25 cGDD from the original time-series. Functional principal component analysis was used to identify cGDD values that described the most important site-specific crop development dynamics occurring in the field. K-means clustering was then applied to the top three functional principal components to classify the fishnet grids at the selected cGDD values. The gap statistic was used to determine the optimum number of clusters that represented areas with comparable crop development characteristics across the field. Results were correlated with ground-reference measurements of soil properties, emergence, and yield.

From the six expanded trifoliate (V6) growth stage to flowering (R1), crop vegetative growth was classified into two zones with different soil pH and percent sand. At flowering, suboptimal hydrology due to erosion resulted in the development of a third zone. By the end of seed filling (R5) five zones were identified across the field. Three of the five zones delineated field areas with different potentials due to varying soil properties. The yield achieved in the other two zones was below soil potential due to other stressors. Soybean yield was established by the end of the seed filling growth stage. The results validated the use of satellite time-series analysis for optimized soybean management.

Quantifying Irrigation Influence on Crop Likelihood in the Central and Eastern US

Lokendra Rathore[1] *and Emily Burchfield*[1]
[1]Emory University, Atlanta, GA, USA

Rising demand for food, feed, fuel, and fiber, alongside climate change and biodiversity loss, will drive transformative shifts in agricultural systems. Understanding how these interconnected challenges will impact cultivation geographies is critical to developing resilient cropping systems and optimizing agricultural resource use efficiency. While the majority of existing literature emphasizes the role of biophysical factors in shaping agricultural land use, human interventions such as management practices, inputs, governmental policies, and socioeconomic factors strongly shape regional cropping systems. We utilize Random Forests to predict crop occurrence for four major row crops in the Central and Eastern United States - corn, soybeans, wheat, and cotton. We leverage SHapley Additive exPlanations (SHAP) values to evaluate the relative contribution of climate, soil, topography, inputs, irrigation, management, economics, infrastructure, and demographics in predicting historical cultivation geographies. We focus on the critical role of irrigation, finding that irrigation changes have diverse effects on crop likelihood, varying by crop and farm resource region. Moreover, the complete collapse of irrigation significantly reduces the likelihood across all major crop-producing areas.

Multinomial Logistic GLMMs: a Patchy Landscape of Software Implementations

Quentin D. Read[1]*, Bruce A. Craig*[2]*, Philip M. Dixon*[3] *and Walter W. Stroup*[4]

[1]USDA Agricultural Research Service, Raleigh, NC USA
[2]Purdue University, West Lafayette, IN USA
[3]Iowa State University, Ames, IA USA
[4]University of Nebraska, Lincoln, NE USA

A cumulative logistic mixed model (CLMM) is appropriate for ordered categorical response data with a multilevel structure. A key issue when fitting CLMMs is determining whether the assumption of proportional odds holds. The proportional odds assumption means that the effect of an explanatory variable on odds ratios is consistent across ordered categories. It can be evaluated graphically or by fitting a model that relaxes the assumption followed by a likelihood ratio test. When the assumption of proportional odds does not hold, software implementation of CLMMs is unfortunately incomplete in both R and SAS platforms. Misleading inference may result from inappropriately choosing a proportional odds model. In this talk, we present available Frequentist implementations, or lack thereof, of CLMMs in R (packages ordinal, glmmTMB, and others) and SAS (proc glimmix and proc nlmixed) both under proportional odds and otherwise. We describe cases for which CLMMs may be fit using off-the-shelf software code, as well as cases for which the user must write the likelihood function explicitly. We further introduce Bayesian implementations of CLMMs using Stan and its R wrapper brms. We demonstrate that even for a very small example dataset, the data overwhelm weakly informative priors, yielding nearly identical results as the maximum likelihood fit. We end with recommendations for future software development work. Collaborators from the stats community are welcome to help!

The Modifiable Areal Unit Problem and the Use of Frame Independent Analytical Methods in Studies of Crop Root System Distribution

Simon S. Riley[1], *Edzard van Santen[2]*
[1]IFAS Statistical Consulting Unit, University of Florida, Gainesville, FL, USA

It is common for studies examining the spatial distribution of crop root systems to characterize that distribution in terms of the total root length or root length density in each of several soil strata or depth classes. Moreover, the choice of how large these strata are defined to be is often arbitrary: chosen for convenience rather than being derived from some feature of the soil or crop. This introduces the issue, known as the modifiable areal unit problem (MAUP), wherein the estimates and test results are conditional on the chosen depth class size, and even the study's overall conclusions may be sensitive to changes in the scale of aggregation prior to analysis. This study performed a sensitivity analysis to ascertain the extent to which the results of two previously published studies were affected by the MAUP. The study also re-analyzed those data sets using a frame independent analysis – one in which the response variable is not defined in terms of spatial units – to illustrate how the MAUP can be avoided in future studies of root system distribution. The results indicate that at least some studies of crop root system distribution may be quite sensitive to the MAUP, although not all estimates and tests were equally affected. It also finds that the results from frame independent analyses are highly consistent, regardless of the scale at which data are collected or aggregated prior to analysis.

Combining Genomic Prediction and Machine Learning Approaches to Improve Genotype Recommendations

Vitor Seiti Sagae[1,2]*, Moysés Nascimento*[2]*, Ana Carolina Campana Nascimento*[2] *and Diego Jarquin*[1]
[1]Jarquin's Lab, Agronomy Department, University of Florida, Gainesville, FL, USA
[2]Laboratory of Computational Intelligence and Statistical Learning (LICAE), Federal University of Viçosa, Viçosa, MG, BRAZIL

Genomic prediction has been used in plant breeding programs to predict genomic estimated breeding values of phenotyped and unphenotyped individuals in observed and unobserved multi-environment trials, aiming to select and recommend promising genotypes. However, for complex highly quantitative traits, genotype-by-environment (G×E) interactions impose challenges for breeders by alternating rankings of individuals across environments. Machine learning approaches enable flexibility in modeling G×E interactions by leveraging high-dimensional variables such as weather, soil features and other sources of information without requiring any assumption and potentially capturing non-linear patterns. In this study, we aimed to combine the efficiency of genomic prediction to obtain genomic estimated breeding values and the potential of random forest (RF) method to characterize and predict new environments using environmental variables information. 32 soybeans check genotypes from the SoyNAM dataset were used to evaluate the proposed approach. The genotypes were evaluated for grain yield (kg ha-1) in 9 locations spread out through different states (Iowa, Illinois, Indiana, Kansas, Michigan, Missouri, Nebraska and Ohio) during the 2012 season. Initially, a GBLUP accounting for G×E was fitted to obtain within-environment genomic estimated breeding values for each genotype. Then, the RF approach was employed individually for each genotype by using the within-environment genomic estimated breeding value as response variable and 38 environmental variables collected daily during the crop season. Lastly, cultivar recommendation was made based on predictions in a buffer area around the evaluated environments. The combination of GBLUP model and RF enabled an increase in resolution to position the genotypes across the cultivated area.

Using Segmented Regression to Estimate Decay of Curve in Seismic Data

Rebekah Scott[1] *and Elizabeth Sunday*[2]
[1]Department of Statistics, Iowa State University, Ames, IA, USA
[2]Department of the Earth, Atmosphere, and Climate, Iowa State University, Ames, IA, USA

In a recent statistical consultation, we assisted a geology graduate student in interpreting simulated seismic data. The problem was to estimate the decay of the curve, the relationship between the log frequency and amplitude after an abrupt change called corner frequency, in seismic waves at nine observations points along a simulated slip fault. We did this using segmented regression. Some of the issues we had to deal with were deciding between a separate or combined analysis for the observation points, determining the source of larger variability in the central observation point, and helping the client interpret the coefficients of this non-linear model. Estimating the decay of the curve in the simulated data allowed our client to validate the simulation and compare the results to current seismic theory.

Determinants of Smallholder Rice Farmers' Willingness-To-Pay for Private Extension Services in Liberia: The Case of Gibi District

***Togba V. Sumo**, Cecilia Ritho, Patrick Irungu*
Department of Agricultural Economics, University of Nairobi, Nairobi, Kenya

Globally, many policymakers and extension professionals have advocated for the privatization of extension services to reduce the burden of funding faced by the state as well as to adequately respond to the low productivity problem of farmers as they endeavor to tackle productivity problems. This study assessed smallholder rice farmers willingness-to-pay (WTP) for private extension services and identified the determinants of their WTP using Gibi District of Liberia as a case study. A multistage sampling technique was used in selecting 296 smallholder rice farmers in the district while the double-bounded dichotomous choice contingent valuation method was used to elicit maximum WTP value for farmers. Descriptive statistics were computed and the double-bounded logit model used to analyze the data. The findings revealed that 78.7% of the rice farmers were willing to pay for privatized extension services and on average, a farmer was willing to pay US$11.21 per farm visit, almost twice the average daily wage rate of a skilled worker in Liberia. The results from the model showed that WTP was significantly positively influenced by the household head's age, years of schooling, household size, annual income, and distance to extension service provider. The study recommends that the Liberian government and its development partners should encourage the private sector to invest more in extension services to take advantage of the relatively high farmers' WTP and effective demand. In addition, the government should design and implement programs that reduce transaction costs in addition to increasing farmers' income to enhance their capacity to pay for privatized extension services.

Canonical Discriminant Analysis – A Multivariate Technique You Didn't Think You Needed

***Edzard van Santen**[1], Simon Riley[1], Hans-Peter Piepho[2]*
[1]IFAS Statistical Consulting Unit, University of Florida, Gainesville, FL, USA
[2]Biostatistics Unit, Institute of Crop Science, Hohenheim University, Stuttgart, Germany

Agronomic experiments *sensu lato* are characterized by having numerous measured response variables, maybe even a plethora. This then often leads to a universal phenomenon, which in English is described as "Can't see the forest for the trees", meaning that the researcher has difficulty grasping the overall picture of an experiment because he/she is too focused on individual response variables. This is where canonical discriminant analysis (CDA), a class-directed dimension-reducing multivariate technique, which considers pre-existing information such as treatment structure, can be extremely helpful. It is not concerned with predicting group membership but only with differences among pre-existing groups; one could call it damage assessment. The purpose of an initial CDA within the analytical workflow is to (a) detect differences among groups when all response variables are jointly considered, and (b) determine which response variables are the main drivers of group differences. This then can inform the necessary univariate analyses – for an agronomist, individual response variables are meaningful and important – and guide subsequent presentation and discussion of results. CDA is dimension-reducing by creating orthogonal canonical axes (canonical variates) from linear combinations of original variables. It does so by maximizing the differences among groups while minimizing the differences within groups. The underlying assumption for CDA is that the response variables are correlated (but not perfectly) and that the number of observations is greater than the number of variables. The first CDA axis (canonical variate) accounts for most of the original multi-variance; successive axes account for a small and smaller proportion. The maximum number of axes that can be extracted is one less than the total number of groups; in practice, the process is stopped when 75% of the original multi-variance have been accounted for. Our presentation will begin with a motivating example involving life history traits in annual bluegrass that enabled a student 25 years ago to make sense of his data. This will be followed by a simple example demonstrating the statistical tests involved, such as Hotelling's T^2 and related tests, canonical correlations, Mahala Nobis D^2 distance, etc. We will then use examples of forage quality in alfalfa, land-management effects on soil properties, and comparing soils impacted by phosphorous application to further demonstrate the utility of this multivariate technique when trying to make sense of data from experiments involving numerous measured variables.

Identifying Key Regulatory Genes in Sorghum Tillering Through Fused Graphical Lasso

Shili Wu[1], *Rajan Kapoor* [1], *Aniruddha Datta*[1] *and Scott Finlayson*[2]
[1]Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA
[2]Soil and Crop Sciences, Texas A&M University, College Station, TX, USA

Shoot branching is an important process in plant growth that shapes a plant's overall structure and has a direct impact on crop performance and farm economics. In grasses, the formation of side shoots—known as tillering—plays a critical role in determining the crop's shape, yield, and efficiency during processing. For example, forage sorghum benefits from producing many tillers, which boosts biomass production, while grain sorghum performs best with fewer tillers to maximize grain yield. In contrast, sorghum grown for sugar extraction is most effective when it produces no tillers, as this simplifies processing and improves sugar recovery.

On a biological level, shoot branching is controlled by both physical signals and plant hormones. Two hormones, Abscisic Acid (ABA) and Jasmonic Acid (JA), are especially important as they work together to regulate the plant's growth and branching patterns. In our study, we aim to find the controlling genes by using the Gaussian graphical lasso—a statistical tool that allows us to determine which genes have the greatest influence on the regulation of shoot branching.

In our analysis, the gene C5YX70 appears to be a strong candidate for controlling ABA pathways, while the gene C5YEI3 seems to play an important role in managing JA levels. The combined activity of these genes appears to be crucial for maintaining the hormonal balance needed for proper shoot branching. Our research is supported by studies of orthologous genes in Arabidopsis thaliana, where similar gene functions have been observed. Furthermore, our approach has revealed additional genetic interactions that may also affect this balance, suggesting potential avenues for further research in crop improvement.

Our findings have direct applications in agricultural production, where tailoring the tillering characteristics of crops can lead to significant improvements in efficiency and yield. Overall, the innovative use of the Gaussian graphical lasso in our study offers a framework for identifying key genetic regulators, thereby contributing to improved crop performance and supporting economic sustainability in modern agriculture.

Artificial Intelligence for Nature-Related Financial Disclosures

Xiao Xu[1]

[1]University of New South Wales, Sydney, NSW, Australia

With the increasing complexity of nature-related financial risks, AI-based solutions are playing a critical role in enhancing risk assessment capabilities. This study introduces AI-driven approaches to address the physical, transition, and systemic risks covered by the Taskforce on Nature-related Financial Disclosures (TNFD) framework.

Three distinct AI models are proposed: flood risk assessment using satellite imagery and Random Forest classification, biodiversity risk monitoring through neural networks trained on images from camera traps, and systemic risk analysis via public sentiment following natural disasters using Natural Language Processing (NLP) techniques like topic modeling.

These AI methodologies provide innovative tools to improve the identification, monitoring, and management of nature-related risks, which are increasingly significant for businesses in the context of evolving regulatory requirements. By leveraging these technologies, the study bridges the gap between theoretical frameworks and practical applications, contributing to more accurate and proactive risk management in nature-related financial disclosures.

This presentation will showcase how artificial intelligence can revolutionize nature-related financial risk assessment by integrating cutting-edge techniques such as satellite imagery analysis, neural networks for biodiversity monitoring, and natural language processing for systemic risk evaluation. By moving beyond traditional financial disclosures and extending the framework beyond TCFD to TNFD, this study challenges existing risk assessment methodologies and demonstrates how AI can provide deeper insights into nature-related financial risks. The presentation will encourage attendees to explore new perspectives on AI-driven risk management, regulatory adaptation, and the future of sustainable finance.

Using Generative Artificial Intelligence to Convert Legacy Statistical Code

Linda J. Young *and Alex Tarter*
USDA National Agricultural Statistics Service, Washington DC, USA

The United States Department of Agriculture's National Agricultural Statistics Service (NASS) has hundreds of programs, many of which were written in code that is no longer supported, not recommended for use, or too expensive. NASS plans to convert all legacy code into freeware languages, such as R or Python, within the next three years. Although the agency does not have the resources to pay an individual to manually convert the code to a modern language, generative AI may be a feasible solution. In a pilot study to assess the viability of this approach, SAS AF programs associated with the Genesis cycle are being converted into Python using GitHub Copilot. During the Genesis cycle, run in March of each year, most samples are drawn for the more than 100 surveys conducted during the upcoming year. As full support of SAS AF ends in 2025, conversion of this code is required to avoid potential risks to the success of the Genesis cycle beginning in 2026. The effects of CoPilot's main attractive translation features, such as the built-in large language model assistant, in-line integrated code commenting, and automatic support of production of scripts with proper syntax, will be discussed. Assessment of the financial and time savings during the program translation while assisted by CoPilot will be detailed. Potential drawbacks or challenges of using CoPilot, including learning curves in using the built-in assistant and manual fact-checking of the generated output lines of text, will also be described. Best practices for using the software, particularly the breaking down of programs into more readily translated contextual "chunks" of script, will be highlighted. Finally, the path forward will be discussed.