

Sunday - April 19th

7:30-8:30 am *Registration and Check-in for Workshop Participants, Pre-Convene*

Workshop, 8:30 am - 5:00 pm

Regency Ballroom B

Analysis of Messy Data Revisions

Dr. Dallas Johnson & Dr. George Milliken, Kansas State University

Dallas Johnson and George Milliken earned international reputations with the publication of their seminal book “The Analysis of Messy Data” in 1984. They have subsequently taught the methods described in the text to a whole generation of statisticians and subject matter scientists.

They have recently revised and updated the book, incorporating the developments in theory and statistical software that have been made in the period since the first edition was published. The second edition of the text was published in 2008.

Their workshop will describe many of these advances and will enable attendees to learn the latest techniques in dealing with nonstandard data situations.

Dallas and George are both emeritus professors at Kansas State. Each earned the doctorate under the direction of Dr. Franklin Graybill at Colorado State, and subsequently joined the faculty Department of Statistics at Kansas State University; George in 1969 and Dallas in 1975.

Tentative outline for the course is the following:

I. Analysis of experiments with heterogeneous errors using mixed models

II. Simultaneous inference procedures and new error rate definitions

III. Multilevel designs: split-plots, strip-plots

Break: 10:00 – 10:15 am ***Conf. I, II, III***

IV. Random effect models- parameter estimates, hypothesis tests, and various types of confidence intervals for variance components

Lunch: Noon – 1:00 pm ***Regency Ballroom B***

V. Mixed model theory- analyzing the random effects part of the model, analyzing the fixed effects part of the model, sample size and power considerations

Break: 2:30 – 2:45 pm ***Conf I, II, III***

VI. Split-plot-in-time analyses of repeated measures experiments

VII. Analyzing repeated measure experiments using MANOVA methods

VIII. p-value adjustment method

IX. Analyzing repeated measures experiments using MIXED model software

X. Crossover experiments with carryover

XI. Lots and lots of examples

Please note: break times are approximate

8:00 – 11:00 pm: ***Social Mixer (Snacks & Cash Bar)*** ***Regency Ballroom A***

Monday - April 20th

8:00-10:00 am Registration for Conference Participants, Pre-Convene

8:30-8:45 am Welcome

Regency Ballroom B

Session #1A, 8:45–9:15 am

Regency Ballroom B

Associating Single Nucleotide Polymorphisms (SNPs) with Binary Traits

Alexander E. Lipka, George P. McCabe, R.W. Doerge, Purdue University

Association mapping involves the implementation of statistical analyses to test for associations between Single Nucleotide Polymorphisms (SNPs) and traits. SNPs are usually located inside or near genomic regions of interest. A statistical association between a SNP and a trait provides evidence for a biological association between the nearby genomic regions and this trait. SNPs, which are predominantly diallelic and abundant in the DNA of most species, are found by comparing the nucleotide sequences of a subset of individuals in a dataset. The statistical analyses conducted in association mapping are divided into three subcategories: single SNP tests, multiple SNP tests, and haplotype-based tests.

This research focuses on the implementation of logistic regression to assess the additive, dominance, and epistatic associations between at least one SNP and a binary trait. In particular, the impact of the phenomenon of *quasi-separation of points* on logistic regression parameter estimates is investigated. Quasi-separation of points occur in a binary trait association mapping dataset when, for at least SNP genotype (or *SNP type*), almost all observed trait values are either “0” or “1”. As a result, the maximum likelihood estimates (MLEs) of logistic regression parameters have large standard errors, and statistical significance between the SNP and trait is not attained. Simulation studies are conducted and a dataset is analyzed to examine the statistical power of these MLEs, and to compare them to the power of MLEs obtained by maximizing *Firth’s penalized likelihood function*. *Firth’s MLEs* have been shown to provide adequate statistical power when quasi-separation of points is present in a dataset.

Session #1B, 9:15–9:45 am

Regency Ballroom B

Bayesian Nonparametric Bioassay Estimation of an Unknown Dose

Bahman Shafii, William J. Price, University of Idaho

Estimation of unknown pesticide levels in experimental samples is an important aspect of many agricultural and environmental studies. Such measurements are often made utilizing a standard dose-response curve. This methodology compares the biological response of a target organism at known dosages to the response of the same organism exposed to an unknown sample. These bioassays are typically more efficient in time and resources than direct chemical assessment of the unknown sample. The form and choice of the standard curve, however, is subjective and can influence the estimation of the unknown dose. Problems may arise when the effects of these problems is to use a more generalized nonparametric estimation approach. This work will outline an alternative bioassay method based on a Bayesian nonparametric standard curve estimation technique. Applications will be demonstrated using an organic weed control method on carrots.

Relative Potency Estimation in Direct Bioassay with Measurement Errors*Weixing Song, Kansas State University*

Measurement errors in the response of the direct bioassays often neglected in the statistical inferences. This paper proposes several estimation procedures for the relative potencies in the direct bioassays by taking the measurement errors into account. Numerical simulations are included to show the superiorities of the new estimators over the current estimators. An R package is build to facilitate the proposed estimation procedures.

10:15 am**Break & Poster Session****Conf I, II, III****Session #2, KEYNOTE ADDRESS, 10:45 am–12:00 pm****Regency Ballroom B****The Process of Designing Experiments***George Milliken, Kansas State University*

The traditional ways design of experiments are taught and/or discussed in text books are not the ways design of experiments are or should be used for real world applications. Designs of experiments are generally taught as single entities such as a completely randomized design, randomized complete block, nested design, Latin square, etc. where examples of each design are discussed. A study in the real world generally consists of a series of steps that might require the use of a series of designs. The steps of the process can impose restrictions that must be incorporated in the appropriate model, but the restrictions are often ignored. The new paradigm for teaching design of experiments must include real world case studies that involve a series of steps imposing restrictions on the collected data. Examples are presented that seem straight forward, but a more complex model is required after recognizing the restrictions imposed by the study process.

12:00 pm**Lunch****Regency Ballroom A****Session #3A, 1:30–2:00 pm****Regency Ballroom B****Predicting the Time of 50% Seed Viability in Maize***Allan Trapp II, Philip Dixon, Iowa State University**Mark Widrechner, USDA-ARS North Central Regional Plant Introduction Station*

The following paper is about predicting seed longevity. Its purpose is to provide USDA-ARS genebank curators a statistical procedure to schedule accession specific seed viability tests. Using data on 2,666 maize accessions collected from the North Central Regional Plant Introduction Station, we fit a random coefficients quadratic linear regression of germination rate against seed age. Seed age predictions at 50%

germination, noted as t_{50} , are derived algebraically. Next, a nonparametric bootstrap is incorporated to derive empirical distributions of the t_{50} 's. The lower 5% quantiles of the bootstrapped t_{50} distributions are recommended as the testing times for each accession. An ROC curve and subsequent viability tests are used to verify the appropriateness of the 5% cut-off quantile.

Session #3B, 2:00–2:30 pm**Regency Ballroom B****Comparisons of Three Parametric Methods for Testing Equality of Binomially Distributed Populations**

Jixiang Wu, Mississippi State University

Johnie N. Jenkins, Jack C. McCarty, USDA-ARS

Many variables like cotton boll retention, worm survival, and disease resistance can be expressed as a binomial distribution. Many parametric methods such as likelihood ratio (LR), Wald, and chi-square tests can be used for detecting the equality among several binomially distributed populations; however, little information is known regarding the properties of these methods when they are used. In this study, using the Monte Carlo simulation technique, three parametric methods, LR, Wald, and chi-square tests, will be compared regarding their Type I error and power under different conditions. The simulated results will be presented. An actual cotton plant mapping data set for boll retention will be analyzed and demonstrated using these three methods. Such an investigation should add new information when binary data are analyzed.

Session #3C, 2:30–3:00 pm**Regency Ballroom B****Bayesian Statistical Methods to Model Heterogeneity in Cow- and – Herd level Relationships between Production and Reproduction in Dairy Cows**

Nora M. Bell, Robert J. Tempelman, Michigan State University

Two of the most important broad classifications of phenotypes for successful dairy production are milk yield and fertility. The nature of the relationship between milk production and reproductive performance of dairy cows is uncertain due to conflicting results reported in many studies. A common deficiency in many such studies is an under appreciation of the dual dimension of the production-reproduction relationship, as defined by within-herd (cow level) and between-herd (herd level) sources of (co) variation. Our overall hypothesis is that the within- and between-herd relationships between milk production and reproduction in dairy cows are heterogeneous and depend upon various herd-related and management factors, such as herd size and parity. Our objective is to develop hierarchical Bayesian extensions to modeling multilevel heterogeneity of covariance matrices. Using such extensions, the relationship between milk production and reproductive performance in dairy cows can be modeled as a function of herd-related factors and management strategies, taking into joint consideration the within-herd and between-herd levels of data (co) variability.

We use a simulation intensive Markov Chain Monte Carlo algorithm to jointly model two continuous traits in our multivariate Bayesian model. A square-root free Cholesky decomposition is applied to the variance-covariance matrices of the residuals (within-herd level) and random effects (between-herd level). As a result, the within- and between- herd generalized autoregressive parameters (GARP), respectively. These GARP, which specify the relationship between the traits, are then modeled as functions of relevant fixed and random effects, thereby providing a mixed model extension of Pourahmadi's method. We validate our

method using a simulation study and apply it to data on 305-day milk yield and calving interval of Michigan dairy cows.

3:00 pm

Break & Poster Session

Conf I, II, III

Session #4, KEYNOTE ADDRESS, 3:30–4:45 pm

Regency Ballroom B

Some Messy Experimental Designs

Dallas Johnson, Kansas State University

This talk considers the analysis of several experiments that scientists had conducted prior to consulting a statistician. After completing the experiment, they came to a statistician for advice on how to analyze the data collected. The designs used are not standard designs, and the talk will consider determining the appropriate error terms for comparing different treatment effects to one another.

6:30–8:00 pm

Student Pizza Party

9:00 pm–12:00 am

Kansas Country Dance

Tuesday - April 21st

Session #5A, 8:30–9:00 am

Regency Ballroom B

Statistical Issues in Next-Generation Sequencing Data

Paul Livermore Auer, R.W. Doerge, Purdue University

High throughput deep-sequencing has emerged as an exciting new tool in a variety of applications such as variant discovery, profiling of histone modifications, identification of transcription factor binding sites, resequencing, and transcriptome characterization. The availability of this technology has generated unprecedented amounts of data in the scientific community (NCBI short read archive) even though few studies have looked carefully at its inherent variability. Considering the time and resources dedicated to studying reproducibility in microarray applications, a similar investment should be made for high throughput deep-sequencing.

Recent studies on mRNA expression levels found little appreciable technical variation in the Illumina sequencing platform (formerly Solexa sequencing). Although these results are encouraging, they are limited to a specific platform and application, and a full investigation of technical variation is still lacking. Furthermore, data from the Illumina platform tend to be quite large and, depending on the application, may require complex analysis techniques. Next-generation sequencing data are truly a messy data problem of

the 21st century. This paper and presentation discuss the key statistical issues involved in defining, identifying, quantifying, and interpreting technical variation in the Illumina sequencing platform.

Session #5B, 9:00–9:30 am

Regency Ballroom B

Comparative Study of Time Series and Multiple Regression for Modeling Dependence of Cattle Body Temperature on Environment Variables During Heat Stress

Manoj Pathak, Dr. A.M. Parkhurst, Dr. Rodrigo Arias, Dr. T.L. Mader, University of Nebraska-Lincoln

During the summer, a challenging thermal environment is known to cause a significant reduction in food intake, growth, milk production, reproduction and even death in cattle. In this study, we attempt to characterize the relationship of cattle body temperature with several environmental variables, such as air temperature, soil surface temperature, relative humidity, solar radiation, wind speed, incoming and outgoing short and long wave radiation. For these variables, the measurements taken over time are correlated. This places severe restrictions on the applicability of many conventional statistical methods that depend on the assumption of independent and identically distributed errors. In addition to these assumptions, there is serious collinearity among several weather variables and the variables are stochastic. Commonly used multiple regression models can be misleading when predictor variables are stochastic and issues of collinearity and stationary are ignored.

In this paper, time series analysis is used as a tool to investigate the adequacy of classical regression models. Various aspects of dynamics of cattle body temperature and its relationship to environmental variables are discussed using the time domain analysis. Finally, we present a detailed approach for fitting cattle body temperature using a transfer function model with multiple environmental variables as input.

Session #5C, 9:30–10:00 am

Regency Ballroom B

Application of the DYA Method to Compare Wheat Variety Yields

Arlin Feyerherm, Kansas State University

At the 1998 Conference, we proposed use of the DYA (differential yielding ability) method for comparing cultivars for yields. Since then we have applied the method to wheat performance variety trials over a nine-state region. Results were updated annually and shared with breeders and specialists. In this paper we review the DYA method: its formulas, its assumptions and models, its generality in application to messy data, and its ability to predict trends in adoption of new varieties by growers. We will also suggest some areas for further study to refine the results. A data set of over 100,000 observations is available for such inquiries.

10:00-11:00 am

***Break & Poster Session
Special Feature for Posters***

Conf I, II, III

A General P-value and Applications to Multiple Testing*Joshua Habiger, Edsel A. Pena, University of South Carolina*

Most multiple testing procedures for False Discovery Rate control are defined in terms of the p-values of the individual tests, and further assume that each p-value is uniformly distributed under the null hypothesis that the gene is not differentially expressed across treatment groups. However, when the sample size for each test is small (which is often the case in microarray analysis), a T-test has Uniformly distributed p-value under the null hypotheses only if the data from each group are independent and normally distributed, and nonparametric tests requiring fewer modeling assumptions typically yield a discrete p-value which is again not uniformly distributed under the null hypotheses. This paper provides a general definition of the p-value, allowing for randomization if necessary, and gives necessary and sufficient conditions in terms of the size of the test for it to be uniformly distributed under the null hypothesis. Using this result, a randomized Wilcoxon ranked sum p-value whose null distribution is uniformly distributed under fewer modeling assumptions is constructed. Simulation studies suggest that the new p-value yield less biased and less variable estimators of the false discovery rate in empirical Bayes procedures than the typical p-value from a t-test. The new p-value is also shown analytically and through simulation to perform favorably in the classical B-H procedure over both the t-test and Wilcoxon ranked sum test p-values. It is also applied to a real microarray data set and performs well.

Statistical Methods for Affymetrix Tiling Array Data*Gayla R. Olbricht, Bruce A. Craig, R.W. Doerge, Purdue University*

Tiling arrays are a microarray technology currently being used for a variety of genomic and epigenomic applications, such as mapping of transcription, DNA methylation, and histone modifications.

Tiling arrays are designed to provide high-density coverage of a genomic region through the systematic placement of probes from one end of the region to the other without regard to genome annotation. This unbiased genomic coverage provides flexibility in the application of tiling arrays since some biological mechanisms, such as epigenomic modifications, may occur anywhere in the genome. In this talk, methods will be proposed to address statistical and bioinformatic issues for data generated from Affymetrix tiling arrays for the model organism *Arabidopsis thaliana*.

12:00 pm**Lunch****Regency Ballroom A**

Correcting Bias in Angler Effort Estimation Using Aerial Counts

Byran Smucker, Penn State University

Robert Lorantas, Pennsylvania Fish and Boat Commission

James Rosenberger, Penn State University

A critical part of an angler survey is angler effort estimation. There have been various methods proposed and implemented to calculate this quantity, and one of the most effective utilizes aerial counts of anglers. In this talk, we describe the effort estimation methodology used in Juniata/Susquehanna River Creel Survey conducted in 2007 by the Pennsylvania Fish and Boat Commission. Daytime angler effort estimates are calculated using an augmented aerial survey, which includes both aerial counts and interview data collected by creel agents. The interview data, obtained via a modified roving ground survey, is used to produce estimates of the daily effort distributions, which are then used to expand instantaneous counts to daily effort estimates.

We present two ratios which ameliorate biases introduced by the aerial survey. An angler-to-people ratio is calculated from the interview data and accounts for shore persons who are not anglers. A ground-truthing ratio is also computed, and corrects for people and boats that were missed during the aerial counts. Variance estimation and angler effort results are discussed as well.

The Morrow Plots: A National Landmark with continued Scientific Relevance

Susanne Aref, Aref Consulting Group LLC

Michelle M. Wander, University of Illinois at Urbana-Champaign

The Morrow Plots experiment is the longest running agricultural field experiment in the US. The history of the Plots' development shows the willingness of the stewards of the Plots to emulate current farming practices with restraint. Although the Plots suffer from a non-random basic design and the usual confounds that pertain to long-term experiments, there are still lessons to be learned and trends to acknowledge. The focus of the experiment has been the productivity and sustainability of continually cultivated land. The Plot's yield records in general continue to increase, while the soil organic carbon continues to decline. Both yield levels and carbon content levels depend on the history of the individual plots. The relationship between soil organic matter and yield is parallel in shape for the different rotations and fertilization schemes except for the main underlying manure treatment, where the shapes of the relationships differ according to the presence or absence of manure. Reduction in tillage depth and intensity have not altered these relationships. Other recent studies concern effects of weather on yield and the effects of fertility treatments and rotation on carbon and nitrogen differences between the introduction of commercial fertilizer and the present.

Mixed Model Quantitative Trait Loci Analysis

Cherie A. Ochsenfeld, Kristofer Jennings, R.W. Doerge, Purdue University

Quantitative Trait Loci (QTL) Analysis has been an effective tool for locating regions of the genome associated with a trait. Due to the unknown nature of the error terms and the complexity of the model, asymptotic thresholds have been difficult to derive. Permutation testing has successfully provided significance thresholds; although, due to the need for exchangeability QTL analysis, using permutation thresholds have been limited to simple linear models. This limitation does not allow researchers to include important covariates into the analysis.

In order to address this limitation, a mixed model that incorporates the ability to include both fixed and random effects into a QTL analysis is proposed. A bootstrapping algorithm is employed to establish significance thresholds that are appropriate for a mixed model analysis. Simulation studies demonstrate an improved detection and estimation of additive effects in QTL studies when influential covariates are incorporated into the analysis model.

3:00 pm

Break

Conf I, II, III

A Tree-specific Stem Profile Model for White Spruce: A Calibration approach by Nonlinear Mixed-effects Modeling

Yuqing Yang, Shongming Huang, Shawn X. Meng, Forest Management Branch, Alberta Department of Sustainable Resource

A tree-specific stem profile (taper) model was developed for white spruce (*Picea glauca* (Moench) Voss) in Alberta, Canada using a nonlinear mixed model approach. Sectioned data from felled trees in a forest inventory program were used for model development. Four candidate variable-exponent taper functions were evaluated as base models. After comparing fit statistics by nonlinear least squares, one model was chosen for further modeling by a mixed model approach. Besides breast height diameter and total tree height, crown ratio, the ratio of crown length to total tree height, was also found to be a significant predictor and incorporated into the model. Between-tree variations in stem form were modeled by incorporating three random parameters into the model. Autocorrelation, a result of multiple measurements taken on each tree, was directly modeled by a banded Toeplitz covariance structure with four bands (TOEP(4)). The developed tree-specific taper model was evaluated for the accuracy of calibrations on stem diameter and tree volume using an independent data set. To make tree-specific calibrations, prior measurement on one or more stem diameters should be available for each tree. For this study, one, two, three, and four prior diameter measurements at various stem locations were evaluated. For each option, random parameters were predicted by an approximate Bayesian estimator based on prior diameter measurement(s). Tree-specific calibrations were subsequently derived from the taper model. Population level predictions of stem diameter and tree volume were also obtained for comparison. Calibration results showed that the developed taper model provided accurate predictions of stem diameter and tree volume at both the population and tree-specific levels.

Statistical Challenges in Genomic Mapping of Complex Traits*Jianming Yu, Kansas State University*

Genomic mapping of complex agronomic and physiological traits presents many challenges to statisticians and geneticists as we embrace the ultrahigh throughput sequencing and genotyping technologies. Extension of mixed model to genomic mapping has been in *in silico* mapping, association mapping, and genome-wide selection. In this report, we showcase both the great potential of the integration of statistics and genetics and further questions that need to be addressed in model selection, R square, and nonmetric multidimensional scaling.

A Sequential Bayesian Calculation for the Classification of Discrete, Ordered Data*Michael Anderson, Kansas State University*

DNA barcodes are short strands of nucleotide bases taken from the cytochrome c oxidase subunit 1 (COI) of the mitochondrial DNA (mtDNA). It has been proposed that these barcodes may be used as a method of differentiating between biological species (Hebert, Ratnasingham, and deWaard 2003). Current methods of species classification and clustering utilize distance measures that are heavily dependent on both evolutionary model assumptions as well as a clearly defined "gap" between intra- and interspecies variation (Meyer and Paulay 2005). We point out the limitations of such distance measures and propose a character-based method of species classification which utilizes an application of Baye's rule to overcome these deficiencies. The proposed method is shown to provide accurate species-level classification and provide answers to important questions not addressable with current methods.

Statistical and Numerical Dependence in Gene Expression Summaries*John R. Stevens, Utah State University**Gabriel Nicholas, University of Wisconsin at Madison*

The Dynamic Bayesian Network (DBN) is a valuable approach for inferring gene networks from temporal microarray expression data. It is a directed graphical model of stochastic processes that can incorporate hidden variables as driving factors (e.g., genes not included on a microarray or transcription factors). Our current work uses an iterative empirical hierarchical Bayesian estimation procedure in tandem with Kalman estimators to estimate the posterior distributions of network parameters. Significant network edges are chosen based on a Bayesian tail probability calculated from the posterior distribution of the interaction matrix and a multiple testing correction. In the early stages of this work we have demonstrated that our method has the potential to identify gene networks on par with existing methods, and within a reasonable amount of computational time.

Non-Randomness in Genic Order and Spatial Structure of a Genome*Douglas Baumann, Purdue University*

Mechanisms for gene movement, including inversions and transposable elements, most likely drive this structure toward linearity, increasing transcription efficiency and organism fitness. Characterizing these processes may lead to increase fitness and a greater understanding of disease resistance, yield, and other socioeconomic traits. Descriptions of these processes, as well as their influence on gene order and spatial structure, are discussed with an eye toward statistical modeling. Non-randomness is assessed through comparisons of spatial data from both completely random simulations and simulations generated using known biological mechanisms. These exploratory analyses and simulations are presented graphically in order to introduce this area of research and motivate discussion.

Using Time Series to Study Dynamics of Sweat Rates of Holstein Cows Exposed to Initial and Prolonged Solar Heat Stress*Bixia Liang, A.M. Parkhurst, University of Nebraska-Lincoln**C.N. Lee, University of Hawaii at Hawaii**K.G. Gebremedhin, P.E. Hillman, Cornell University, Ithaca**R.J. Collier, University of Arizona*

Sweating is a very important way for cows to cope with heat stress. We are interested in the ability of Holstein cows to sustain high sweat or evaporation rates when exposed to solar radiation. In this study, there were two solar heat stress treatments: onset and prolonged. The onset data provides an opportunity to examine the impact of suddenly turning on an additional solar thermal load. The prolonged data allows us to examine the impact of exposure to solar heat stress for an extended period (3 hr). Two questions of interest are: Do cows sweat at a constant or cyclic rate? Is there a difference in the dynamics of the two treatments: onset and prolonged solar heat stress? We examine the data for stationary. We fit ARIMA models and estimate the parameters in the time domain. In the frequency domain, we use both parametric and nonparametric spectral estimation to identify cyclic patterns in the sweat rates. We present the results and discuss the usefulness of each technique for analyzing the dynamics of sweat rates.

Comparing Experimental Designs for a Bi-Logistical Model used to Estimate Heat Stress when Moving Feedlot Cattle*Xiapeng Li, Anne M. Parkhurst, Terry L. Mader, University of Nebraska-Lincoln*

Processing and handling cattle requires expenditure of energy causing an elevation of body temperature, depending on the ambient conditions. During summertime heat waves, caution should be exercised in moving cattle. More knowledge of the dynamics of body temperature, T_b , could lead to specific recommendations of how far cattle can be moved before becoming thermally challenged. Data comes from feedlot trials conducted over four days. A bi-logistic mixed model of T_b is used to describe the handling process.

This model provides estimates for several important biological parameters describing the thermal challenge and recovery: the maximum Tb challenge, rate of heat challenge (rate of increase in Tb), the Tb at the challenge inflection point, baseline for recovery, recovery rate (rate of decrease in Tb) and Tb at the recovery inflection point. Fitting a nonlinear mixed model with six parameters under extremely variable animal and environmental conditions is difficult especially when the treatment factor (distance) is introduced into the model. Additional difficulties in fitting the model arise as the experimental design increases in complexity from a CRD to a repeated Latin Square. The objectives of this study are: to examine the bi-logistic model with distance as a treatment factor and estimate the relative efficiencies as the experimental design is simplified.

Poster

Conf I, II & III

Using Time Series to Study Effect of Air Temperature and Humidity on Body Temperature of Cows in Puerto Rico

*Yan Zeng, A.M. Parkhurst, University of Nebraska-Lincoln
J. Pantoja, University of Puerto Rico*

Body temperature is an important measure for monitoring the health status of cows. In this study we attempt to determine if a cow's body temperature is related to ambient temperature, relative humidity and/or the temperature humidity index (THI) and look for signs of heat stress. The data was collected at five minute intervals during the winter months in Puerto Rico. A succession of time series analyses are conducted in both the frequency and time domains. The stationarity of the variables is examined. ARIMA models are applied to detect the dynamic properties of cows body temperature. Nonparametric spectral estimation is also performed in the frequency domain. A search for indications of heat stress is performed by characterizing the relationship between body temperature and environmental factors. We present a detailed comparison of multiple regressions with auto correlated errors and a transfer function model in time domain.

Poster

Conf I, II & III

Development of Precipitation-Runoff Relationship in Nekarood watershed in Mazandaran Province, Iran

Karim Soleimani, Mahmood Habibneshad, College of Resources, University of Mazandaran.Sari

Application of precipitation-runoff relationships or models developed based on easily accessible variables is essential, since the number of hydrometry stations providing necessary information on watershed runoff is very limited. In this research an attempt has been made to compare different models and developed the best model of precipitation-runoff. The study area located in the north of Iran and comprises 1992 km². In the present study because of the large area we should divide that in 3 part, Sefidchah station in up. Gelevard station in middle and abloo station in down. Then analyze 3 station separate. The modeling process was followed by software called Spss and different bivariate regression. The relationships were investigated using bivariate in different forms linear, logarithmic, power, cubic, quadratic, inverse growth curve. Furthermore, consideration of results show that power model had highest residual sum of square (R.S.S.). After that this model has select for the basin.

Using Compartment Models to Evaluate the Fate of Sulfamethazine in Surface Water

*Garritt Page, Phillip Dixon, Joel Coats, Iowa State University
Thomas Moorman, USDA-ARS National Soil Tilth*

Sulfamethazine (SMZ) is an antibiotic widely used in livestock production; this drug can be transported to surface water bodies from manured fields. As the presence of SMZ may have implications on organisms present in surface water, it is beneficial to understand the adsorption and degradation processes of SMZ in surface water. This was investigated using small pond water microcosms. Four types of microcosms were used: pond water only. For each treatment the microcosms were replicated 4 times at 5 different time periods. We fit compartment models to characterize the kinetics of adsorption and degradation and suggest possible new pathways. Preliminary results suggest that major mechanisms of dissipation included adsorption to sediment and photodegradation.

A Simulation Study to Evaluate the Analysis of Non-normal Data with Correlated Errors

Emily Hagel, Walter Stroup, University of Nebraska

Experiments with correlated errors are common in many disciplines, including agriculture. Examples include repeated measures or spatial variability. While, analysis of correlated error models with normal data is fairly well understood, analysis of correlated error models with non-normal data, such as binary or count response data is not as well understood. With the availability of software such as SAS PROC GLIMMIX, it has become dramatically easier to fit generalized linear models with correlated errors to non-normal data. However, there are contending approaches on how to accomplish this. The traditional approach uses working correlation matrices. Alternative approaches use so called G-side modeling or radial smoothing. The purpose of this paper is to present the results of simulation studies using these contending approaches. Two scenarios will be considered. One will be repeated measures data using a design structure similar to the one Guerin and Stroup presented at the 2000 Applied Statistics in Agriculture Conference. The other is based on a spatial example first discussed by Gotway and Stroup (JABES, 1997).

The Statistical Analysis of Dynamics of Organic Substance and Microbiotic Complex in Root-inhabited Media and their Relationship with Plants Productivity in Conditions of Primary Soil Formation

V.K. Mukhomorov, Agrophysical Institute, St. Petersburg Russia

The statistical analysis of results of intensive multivegetative experiment (during the course of 23 continuous vegetations) by growing of a tomato and spring wheat plants on an originally abiogenic mineral substrate (crushed granite and zeolite) in conditions of primary soil formation is executed. Statistically safely is established the interrelation of the organic substance and microbiotic complex with plants productivity and the contents of nitrates in reproductive organs. It is shown that the operation of acid base regeneration of RM results in statistically reliable gives raise to plants productivity in conditions of a

primary soil formation. Using methods of an information theory the statistical information parameter describing information interchanging between multicomponent systems of organic substance and microbiotic complex is introduced. It is established the information streams between these dynamic systems statistically safely correlate with dynamics of plants productivity.

Poster

Conf I, II & III

Integrated Analysis of Genomic and Quantitative Trait Data

Tilman Achberger, Ivan Baxter, James Fleet, David Salt, R.W. Doerge, Purdue University

A fundamental goal in genetics research is to locate genomic regions associated with the measurable traits of an organism. Quantitative Trait Loci (QTL) mapping, expression-QTL (eQTL) mapping and Expression Level Polymorphism (ELP) mapping are three methods for finding genomic regions associated with phenotypic traits, gene expression traits, and changes in gene expression under different environmental conditions, respectively. While each of these methods can be used separately to investigate relevant research hypothesis, there is no current method to combine all three of these methods in any meaningful way.

This poster presents the work conducted at Purdue University to integrate these methods into on statistically powerful and informative analysis. Our approach makes use of the dependencies between traits, which will allow researchers to investigate relationships between gene expression levels and their downstream results as phenotypic traits. Preliminary results will be presented using *Arabidopsis thaliana* data, that combine gene expression data (Dr. Dina St. Clair, UC Davis), ionomic trait data (Dr. David Salt, Purdue University), and traditional phenotypic trait data (Dr. Olivier Loudet, INRA France).

Poster

Conf I, II & III

Inferring Genetic Regulatory Interactions with Dynamic Bayesian Networks and Kalman Estimators

Andrea Rau, Purdue University & INRA

Florence Jaffr zic, Jean-Louis Foulley, INRA

Rebecca Doerge, Purdue University

Gene regulatory networks are collections of genes that interact with each other and with other substances in the cell. Such interactions regulate the rate and degree to which genes are transcribed into mRNA and proteins. Even though these systems are typically difficult to elucidate due to the large number of genes and the limited number of biological replicates, by measuring gene expression over time, it is possible to estimate the gene network involved in a particular cell process.

The Dynamic Bayesian Network (DBN) is a valuable approach for inferring gene networks from temporal microarray expression data. It is directed graphical model of stochastic processes that can incorporate hidden variables as driving factors (e.g., genes not included on a microarray or transcription factors). Our current work uses an iterative empirical hierarchical Bayesian estimation procedure in tandem with Kalman estimators to estimate the posterior distributions of network parameters. Significant network edges are chosen based on a Bayesian tail probability calculated from the posterior distribution of the interaction matrix and a multiple testing correction. In the early stages of this work we have demonstrated

that our method has the potential to identify gene networks on par with existing methods, and within a reasonable amount of computational time.

Poster

Conf I, II & III

What caused the Death? A Random Act or Exceeding an Individual Tolerance, and how to tell.

Man-Yu Yum, Philip M. Dixon, Iowa State University

One explanation of the dose-response relationship in toxicology is the individual tolerance or individual effective dose (IED) model proposed by Gaddum. It is postulated that each individual has a unique innate IED, beyond which it dies. An alternative explanation is that survival or death of an individual is stochastic. These two models have drastically different implications for repeated exposures to a toxicant. Despite the wide acceptance of the IED assumption, it has seldom been verified.

One way to distinguish the contribution of the two mechanisms is to kill an organism twice. We propose using the correlation between the two times to death as a measure of relative contribution of the IED and stochastic model. The closer the correlation is to 1, the larger the contribution from IED. One can't kill something twice, but Zhao and Newman (2007) collected a special type of bivariate censored time-to-death data from freshwater amphipods that were twice exposed to the same dose of toxicant. We develop a maximum likelihood estimator of the correlation from these data.

For the toxicant CuSO₄, the estimated correlations for the two replicated experiments are 0.80 and 0.97, with 95% profile CI (0.14, 0.95) and (0.66, 0.99). For another toxicant NaPCP, the estimated correlations are lower, 0.67 and 0.66, with 95% profile CI (0.29, 0.85) and (0.34, 0.83). These suggest that CuSO₄ may be more dominated by IED than stochasticity compared to NaPCP.

Poster

Conf I, II & III

Associating (Sorghum) GUS expression with Cinfu and Huck Promoters

Brian Denton, Purdue University

Balasubramaniam Muthukuma, Jeff Bennetzen, University of Georgia

R. W. Doerge, Purdue University

Retrotransposons are components of DNA that have the ability to amplify themselves in a genome; they are particularly numerous in plant genomes. While it is known that 70% of the Maize genome is comprised of retrotransposons and that Maize is highly related to Sorghum, less is known about Sorghum. In Maize there are two retrotransposons known as Cinfu and Huck that occupy 9% and 10% of the Maize genome, respectively. Work from the Bennetzen laboratory (University of Georgia) in collaboration with Purdue University, has employed Cinfu and Huck primers (or DNA sequence) to amplify a family of long terminal repeats (LTR; retrotransposons) in Sorghum. The specific question addressed by this investigation is whether there are statistically significant associations between nucleotide position in the promoter sequences, for either Cinfu and Huck, and the GUS reporter system in Sorghum. The GUS reporter system is typically employed to evaluate the activity of promoters by assessing the expression of a gene under a specific promoter (e.g., Cinfu or Huck).

A repeated measures analysis of variance (ANOVA) model was used to estimate the correlation with GUS expression. The statistical significance of nucleotide effects was assessed at the 0.05 level of significance with a Bonferroni adjustment for multiple comparisons. The evaluation of Cinfu1 revealed 10 significant base pair positions out of a total of 643, with one region having two consecutive significant base pair positions. Huck sequences contained 46 significant base pair positions out of a total of 1,479, with six regions that consisted of two consecutive significant base pair positions, and one region of three consecutive base pair positions. These results demonstrate a statistically significant association between nucleotide position in the promoter sequences Cinfu1 and Huck and the expression of GUS.

Poster

Conf I, II & III

Proc Report in Style

Wendy Boberg, AFMC

I would like to demonstrate how you can create colorful PDF files using ODA (Output Delivery System) with different predefined style templates for the layout and color scheme of both tables and graphs. There are a few different ways you can customize your reports in the Report Procedure. Traffic lighting is a very popular technique and can be achieved by changing the font and/or the background colors. I prefer to customize my reports by adding background color to highlight a row, column, and/or cell in a table.

I will share my favorite styles and show you how to look at your tables in the different styles available. I will briefly explain how you can create your own style by changing the color scheme of an existing style so that it can be used for all of your reports. I will explain how I used the pieces of code to get the table with a row, a column and individual cells highlighted with several different colors.

I will focus on creating PDF files using ODS, but these methods can be used with other output destinations.

This paper is for beginner to intermediate level SAS programmers with experience creating tables with Report procedure.

Poster

Conf I, II & III

Seven Years of Statcom at Purdue University: Managing a Growing Number of Student Volunteers

Douglas Baumann, Andrea M. Rau, Purdue University

Statistics in the Community (STATCOM) at Purdue is a volunteer community outreach organization directed and staffed by graduate students which provides free professional statistical consulting services to governmental and nonprofit groups. Since its establishment in 2001, the STATCOM student volunteer base at Purdue has grown to over 55 students across several departments. In addition to growing its volunteer base, STATCOM at Purdue has developed new ways to serve the community, such as P-12 Outreach and STATCOM Network components. This growth in student participation and available services has required STATCOM at Purdue to develop a broader organizational infrastructure, adding the roles of Project Manager, Network Outreach Coordinator, and P-12 Officer.

