NISS at JSM 2023 Toronto

Sunday, Aug 6: 2:00 PM - 3:50 PM

Use of Color in Statistical Charts - **Heike Hofmann**, Chair (Iowa State University)
Panelists: **Lyn Bartram**, Simon Fraser University; **Danielle Szafir**, UNC-Chapel Hill; **Ian Lyttle**, Schneider Electric; **Gail Mulligan**, National Center for Education Statistics
Organizer: **Haley Jeppson**, Iowa State University and NISS
1551 Invited Panel Session
Convention Centre Room: CC-810
Main Sponsor: Section on Statistical Graphics

Panel: ***Use of Color in Statistical Charts***

Data visualization communicates the underlying data by mapping the data to the visual properties of the graphic, such as position, size, or color. For these encodings to be effective, the mapping between the visual properties and values must preserve important differences in the data and be perceivable by the viewer. Effective color encodings in data visualization is a complex challenge, and its impact on the accuracy of visual communication is often overlooked. While it is widely accepted that color scales can deceptively manipulate a viewer's perception of the underlying data, guidelines for constructing color scales are loosely defined, difficult to express mathematically, and often difficult to apply. At this panel, researchers at the forefront of investigating the use of color in data visualizations will share their current research and tools for selecting, constructing, and evaluating color scales. We will discuss the effective matching of a color scale to the data it is encoding, perceptual uniformity, accessibility concerns, and more.

Sunday, Aug 6: 4:00 PM - 5:50 PM

Innovations in Estimation Approaches for Biomedical Data - **Luca Sartore, Chair (NISS/NASS)**
0019 Contributed Papers
Convention Centre Room: CC-809
Main Sponsor: Biometrics Section

Sunday, Aug 6: 4:00 PM - 5:50 PM

Machine Learning & Imputation Techniques for Survey Design & Missing Data - **Elizabeth Petraglia**, Chair (Westat)
0021 Contributed Papers
Convention Centre Room: CC-206B
Main Sponsor: ASA Government Statistics Section

Presentation: ***Early Season Corn Acreage Estimates in the Presence of Extreme Weather***

The United States Department of Agriculture National Agricultural Statistics Service (NASS) provides timely and accurate statistics in service to U.S. agriculture. An example includes planted acreage estimates for corn. NASS conducts surveys in March and June to provide early season estimates of corn acreage. Since planting typically occurs in May and June, farmers are generally reporting planting intentions in March. The information collected through the June survey is typically a close representation of what is planted since corn planting generally is complete by the end of June. It is possible, however, that planting can be prevented by extreme weather. If this is the case, the June survey may still include planting intentions, which can bias the results when intentions are changed due to weather conditions. More information is necessary to mitigate this potential source of bias. The objective of this study is to use machine learning to combine the June survey estimate with precipitation, temperature, economic and other data to forecast corn planted acreage. The accuracies of the model estimates are measured based on the relative error with respect to official acreages.

Co-Authors: **Luca Sartore, NISS/NASS**; **Noemi Guindin**, **Kevin Hunt**, & **Claire Boryan**, USDA/NASS
First Author & Presenting Author: **Jonathon Abernethy**, USDA/NASS

Presentation: ***Optimizing Imputation for an Area Survey***

Every year the U.S. Department of Agriculture's National Agricultural Statistics Service (NASS) conducts the June Area Survey (JAS) based on an area frame, which has complete coverage of all land in the contiguous U.S. The data collected from the JAS are used to supply direct estimates of acreage and measures of sampling coverage for NASS's list frame, which consists of all known farms in the U.S. Response rates have been declining in many federal surveys, including the JAS, leading to heavier reliance on imputation. NASS has begun exploration of utilizing automatic imputation for the JAS using various machine learning models. Previous research has found that NASS's Predictive Cropland Data Layer (PCDL) has good predictive power at certain entropy levels for major U.S. crop commodities. This paper explores the interaction between the entropy levels, PCDL values, and other data from administrative and survey sources to determine which entropy levels are appropriate for the purposes of imputing the JAS.

Co-Authors: **Luca Sartore, NISS/NASS**; **Arthur Rosales** & **Tara Murphy**, USDA/NASS
First Author & Presenting Author: **Sean Rhodes**, USDA/NASS

Sunday, Aug 6: 2:00 PM - 3:50 PM

Time Series Analysis for Complex Business and Economic Data - **David Matteson**, Chair (Cornell University)
Organizer: David Matteson, Cornell University
1787 Topic-Contributed Paper Session
Convention Centre Room: CC-206D
Main Sponsor: Business and Economic Statistics Section
Co Sponsors: IMS & Section on Nonparametric Statistics

Sunday, Aug 6: 2:00 PM - 3:50 PM

Jerry Sacks Memorial Session - **James Rosenberger**, Chair (Penn State University and NISS)
James Rosenberger & **Lingzhou Xue**, Organizers (Pennsylvania State University and NISS)
2009 Invited Paper Session
Convention Centre Room: CC-718A
Main Sponsor: Memorial

Presentation: ***Jerry Sacks and NISS***

Even before NISS existed, Jerry Sacks was a moving force in its creation, by understanding and articulating the need for cross-disciplinary research involving the statistical sciences. Based on my eight years as Associate Director under Jerry and 14 years as Director, I will highlight the ups and downs of the early years of NISS. During that time, NISS evolved from four offices in rented space to a building of its own, which was ultimately shared with SAMSI. I will also focus on Jerry's and NISS' scientific heritage, both in terms of specifics and in making cross-disciplinary research commonplace in statistics. The presentation will include many photographs, not only of Jerry, but also others who helped put NISS on the path that has enabled it to thrive for more than 30 years.

Speaker: Alan Karr, AFK Analytics, LLC
Speaker: William Welch, University of British Columbia
Speaker: Piyushimita Thakuriah, Rutgers University, New Brunswick, New Jersey
Speaker: James Berger, Duke University

Sunday, Aug 6: 4:00 PM - 5:50 PM

Statistics Informing Decisions for Our Communities on Gun and Crime Policy - **David Banks**, Chair (Duke University)
Aleksandra Slavkovic, Organizer (Pennsylvania State University)
1436 Invited Paper Session
Convention Centre Room: CC-203C, CC-203D
Main Sponsor: Statistics and Public Policy

Presentation: ***Statisticians Engage in Gun Violence Research***

Government reports document more than 14,000 homicides and more than 195,000 aggravated assaults with firearms in 2017. There were 346 mass shootings, with 4 or more victims, including over 2000 people shot. And these statistics do not include suicides (two-thirds of gun deaths) or accidents (5% of gun deaths). This article describes statistical issues discussed at a national forum to stimulate collaboration between statisticians and criminologists. Topics include: (i) available data sources and their shortcomings and efforts to improve the quality, and alternative new data registers of shootings; (ii) gun violence patterns and trends, with statistical models and clustering effects in urban areas; (iii) research for understanding effective strategies for gun violence prevention and the role of the police in solving gun homicides; (iv) the role of reliable forensic science in solving cases involving shootings; and (v) the topic of police shootings, where they are more prevalent and the characteristics of the officers involved. The final section calls the statistical community to engage in collaborations with social scientists to provide the most effective methodological tools.

Speaker: James Rosenberger, Penn State University and NISS

Monday, Aug 7: 8:30 AM - 10:20 AM

Statistical Learning with Practical Constraints - **Amita Manatunga**, Chair (Emory University) *by COPSS Leadership Academy Award Recipients*
Organizer: Amita Manatunga, Emory University
1343 Invited Paper Session
Convention Centre Room: CC-718B
Main Sponsor: Committee of Presidents of Statistical Societies

Presentation: ***Theoretical Guarantees for Sparse PCA based on the Elastic Net***

Sparse principal component analysis (SPCA) has been widely used for dimensionality reduction and feature extraction in high-dimensional data analysis. Despite there being many methodological and theoretical developments of SPCA in the past two decades, the theoretical guarantees of the

popular SPCA algorithm proposed by Zou, Hastie, and Tibshirani (2006) based on the elastic net are still unknown. To close this important theoretical gap, we first revisit the SPCA algorithm of Zou et al. (2006) and present our implementation. Also, we study a computationally more efficient variant of the SPCA algorithm in Zou et al. (2006) that can be considered as the limiting case. We provide the guarantees of convergence to a stationary point for both versions of SPCA. We prove that, under a sparse spiked covariance model, both algorithms can recover the principal subspace consistently under mild regularity conditions. We show that their estimation error bounds match the best available bounds of existing works or the minimax rates up to some logarithmic factors. Moreover, we demonstrate the numerical performance of both algorithms in simulation studies.

Co-Authors: Teng Zhang & **Haoyi Yang**, Penn State University
Speaker: Lingzhou Xue, Pennsylvania State University and NISS

Monday, Aug 7: 10:30 AM - 12:20 PM

Hierarchical Models for Survey Data - **Akhil Vaish**, Chair (RTI International)
0046 Contributed Papers
Convention Centre Room: CC-713B
Main Sponsor: Survey Research Methods Section

Presentation: ***Mixture Model and Its Application***

Federal Statistical Agencies are required to produce estimates of subpopulations with few samples. Traditional design-based estimates based on only sample data from subpopulations are unreliable. For over forty years, the Fay-Herriot model has been widely used to produce reliable small area statistics. This model develops prediction of small area of interest based on a linear regression on auxiliary variables. The Fay-Herriot model are treated as independent and normally distributed zero-mean random variables with an unknown variance. It is sensitive to outliers because the outliers may result in overestimation of the model variance. In this talk, we propose a new robust estimation approach to estimate small area populations. The robustness property is achieved by replacing the standard normality assumption of the model errors by a mixture of two normal distributions with different variances, making this mixture model less sensitive to outliers. Finally, we compare the estimates from the proposed mixture model to alternative existing methods using study data set from the Cash Rents Survey conducted by the National Agricultural Statistics Service (NASS).

Co-Author: Lu Chen, NISS/NASS; Gauri Datta, University of Georgia
First Author & Presenting Author: Yang Cheng, USDA/NASS

Monday, Aug 7: 2:00 PM - 3:50 PM

Contributed Poster Presentations - **Jacob Bien**, Chair (University of Southern California)
Convention Centre Room: CC-Hall E
Main Sponsor: Business and Economic Statistics Section

Contributed Poster Presentation: ***04 Drift vs Shift: Decoupling Trends and Changepoint Analysis***

We introduce a new approach for decoupling trends (drift) and changepoints (shifts) in time series. Our locally adaptive model-based approach for robustly decoupling combines Bayesian trend filtering and machine learning based regularization. An over-parameterized Bayesian dynamic linear model (DLM) is first applied to characterize drift. Then a weighted penalized likelihood estimator is paired with the estimated DLM posterior distribution to identify shifts. We show how Bayesian DLMs specified with so-called shrinkage priors can provide smooth estimates of underlying trends in the presence of complex noise components. However, their inability to shrink exactly to zero inhibits direct changepoint detection. In contrast, penalized likelihood methods are highly effective in locating changepoints. However, they require data with simple patterns in both signal and noise. The proposed decoupling approach combines the strengths of both, i.e. the flexibility of Bayesian DLMs with the hard thresholding property of penalized likelihood estimators, to provide changepoint analysis in complex, modern settings. Our framework is outlier robust & can identify a variety of complex changes.

First & Presenting Author: David Matteson, Cornell University

Tuesday, Aug 8: 8:30 AM - 10:20 AM

Cell Suppression Methods for Economic Census and Establishment Surveys - **Yang Cheng**, Chair (USDA/NASS)
Discussant: Saki Kinney, RTI International
Organizer: Yang Cheng , USDA/NASS
1835 Topic-Contributed Paper Session
Convention Centre Room: CC-715B
Main Sponsor: Survey Research Methods Section
Co Sponsors: Government Statistics Section; Social Statistics Section

Presentation: ***Overview of Cell Suppression Methods***

Statistical agencies widely use cell suppression methods for economic censuses and establishment surveys to protect sensitive tabular data from disclosure to the public. The goal is to reduce the risk of disclosure through first identifying sensitive cells as primary suppressions and then finding additional cells as the complementary (secondary) suppressions to protect the primary cells against an attacker. In general, cell suppression

problems (CSP) can be described as a linear programming problem. In this presentation, cell suppression models are reviewed, with a focus on network flow models with heuristic solutions for two-dimensional tables as well as exact optimal solutions. Applications of cell suppression methods from statistical agencies are highlighted. The extension of the solutions to high-dimensional, hierarchical, and linked tables is also discussed.

Co-Author(s): Lu Chen, NISS/NASS; **Yang Cheng** & **Michael Jacobsen**, USDA/NASS
Speaker: Ruiyi Zhang, NISS/NASS

Tuesday, Aug 8: 8:30 AM - 10:20 AM

Statistical Challenges in Public Health and Medicine - **James Rosenberger**, Chair (Penn State University and NISS)
Discussant: Lingzhou Xue, Pennsylvania State University and NISS
Organizers: Lingzhou Xue & **James Rosenberger**, Penn State University and NISS
1278 Invited Paper Session
Convention Centre Room: CC-718B
Main Sponsor: National Institute of Statistical Sciences (NISS)
Co Sponsors: Biometrics Section & ENAR

Presentation: *Challenges of Modeling Longitudinal Intensive Care Unit Data*

Prediction of health outcomes is an important component for determining how to make recommendations and treat individuals. Regarding treatment, the intensive care unit (ICU) is a place where many such decisions are made. A primary goal for ICU patients is treating them to achieve positive outcomes (e.g., hospital discharge alive, improvement from in-hospital ailments, extended survival). A major analytical issue is the preponderance of information available at ICU entry (e.g., age, sex, co-morbidities, prescriptions, vital signs), and especially longitudinally (e.g., vital sign changes, dynamic renal function, in-ICU treatment). I will present some interesting analytic challenges utilizing longitudinal data for predictive modeling that my collaborators and I have encountered from a large ICU database, and discuss a few remedies that we have investigated.

Speaker: Joel Dubin, University of Waterloo

Presentation: *Ensemble Methods for Testing a Global Null with Application to Whole Genome Sequencing Association Studies*

Testing a global null is a canonical problem in statistics and has a wide range of applications. In view of the fact of no uniformly most powerful test, prior and/or domain knowledge are commonly used to focus on a certain class of alternatives to improve the testing power, e.g., the class of alternatives in the scenario of the same effect sign or signal sparsity. However, it is generally challenging to develop tests that are particularly powerful against a certain class of alternatives. In this paper, motivated by the success of ensemble learning methods for prediction or classification, we propose an ensemble framework for testing that mimics the spirit of random forests to deal with the challenges. Our ensemble testing framework aggregates a collection of weak base tests to form a final ensemble test that maintains strong and robust power. The key component of the framework is to introduce a certain random procedure in the construction of base tests. We then apply the framework to four problems about global testing in different classes of alternatives arising from Whole Genome Sequencing (WGS) association studies. Specific ensemble tests are proposed for each of these problems,…

Speaker: Xihong Lin, Harvard T.H. Chan School of Public Health

Presentation: *Mediation Analysis with External Summary Data on Total Effect*

As modern assaying technologies continue to improve, environmental health studies are increasingly measuring endogenous omics data to study intermediary biological pathways of outcome-exposure associations. Mediation analysis is often carried out when there is a well-established literature showing statistical and practical significance of the association between an exogenous exposure and a health outcome of interest, or the total effect. For example, there are a plethora of studies associating maternal phthalate exposure with preterm delivery, and researchers are now trying to characterize the mechanisms by which phthalate exposure impacts final gestational age. Existing methodology for performing mediation analyses does not leverage the rich external information available on the total effect. We show that incorporating external summary-level information on the total effect improves estimation efficiency of the natural direct and indirect is a function of the partial R-squared comparing the outcome model with and without the mediators. Additionally, we discuss how to handle incongruous external information which can arise from transportability violations or fundamental…

Speaker: Bhramar Mukherjee, University of Michigan

Tuesday, Aug 8: 10:30 AM - 12:20 PM

Contributed Poster Presentation - **Jacob Bien**, Chair (University of Southern California)
Convention Centre Room: CC-Hall E
Main Sponsor: Section on Statistical Learning and Data Science

Contributed Poster Presentation: *33 Inference for Nonstationary Economic Time Series*

Many data in economics are observed over time, admitting temporal correlation and also exhibiting persistent upward and downward movements. A relaxation of standard assumptions is nonstationarity modeled through locally stationary processes with a smoothly varying trend.

This talk will present novel estimators for high-dimensional autocovariance and precision matrices that uses the locally stationarity property. The estimators are used to derive consistent predictors in nonstationary time series. Besides some theoretical verifications, we illustrate the finite sample properties of the new methodology by a simulation study and an application to economics data.

Co-Author: **David Matteson**, Cornell University
First & Presenting Author: **Marie-Christine Duker**, Cornell University

Wednesday, Aug 9: 2:00 PM - 3:50 PM

Weighting Adjustments - **Evrim Oral**, Chair (LSUHSC School of Public Health)
0141 Contributed Papers
Convention Centre Room: CC-717B
Main Sponsor: ASA Survey Research Methods Section

Presentation: ***Consistency of Survey Estimates through Adjusted Integer Weights***

Calibrating survey weights can improve estimates by enforcing consistency constraints derived from external known benchmarks, i.e., the weighted sums of certain variables should be equal to their known population totals. Popular calibration methods, even though robust, often fail when multiple benchmarks are to be satisfied simultaneously. In this paper, new extensions of the integer calibration method adopted by NASS for its 2017 US Census of Agriculture are discussed. New constraint relaxation techniques and new "distances" between calibrated and design weights are introduced for big-data applications, where millions of records are processed by benchmarking simultaneously several thousand variables. The same algorithm can also be adopted to produce fractional adjusted weights with simple arithmetic operations. The consistency of the estimator and the accuracy of the results are investigated through a simulation study using the R-package inca. An application using NASS Farm Labor Survey data illustrates how extreme weights are distributed among other records while maintaining benchmarking relationships.

Co-Author: **Lu Chen, NISS/NASS**
First Author & Presenting Author: **Luca Sartore, NISS/NASS**

Thursday, Aug 10: 8:30 AM - 10:20 AM

Contributions to Inference from Survey Samples - **Guofen Yan**, Chair (University of Virginia)
Organizer & Discussant: Balgobin Nandram, Worcester Polytechnic Institute
1187 Invited Paper Session
Convention Centre Room: CC-715A
Main Sponsor: Survey Research Methods Section
Co Sponsors: Government Statistics Section; Indian Statistical Institute

Presentation: ***Hierarchical Bayesian Model for State-Level Cash Rental Rates***

The United States Department of Agriculture's National Agricultural Statistics Service provides state estimates of the cash rent paid for various land-use categories based on the Cash Rents Survey (CRS). Some of the realized sample sizes are too small to support reliable direct estimates, and there are outliers. In addition, quantities of interest for geographically contiguous small areas in CRS display a spatial pattern. Statistical agencies are increasingly considering the use of small area models in the estimation process. These models can provide indirect but reliable estimates for small areas. Therefore, we propose a hierarchical Bayesian area-level two-component mixture model with spatial random effects to account for outliers and spatial correlation. The model incorporates two years of data and a discounting factor for the first year provides a prior for the hyperparameters that is not too tight. We assess the effectiveness of the spatial model based on a simulation study and a case study from 2022 CRS. The results show superior performance of the proposed model over the direct estimates and the original Fay-Herriot model.

Co-Author: **Balgobin Nandram**, Worcester Polytechnic Institute
Speaker: **Lu Chen, NISS/NASS**

Thursday, Aug 10: 8:30 AM - 10:20 AM

Innovation in Time Series Modeling, Estimation and Testing - **Tae-Hwy Lee**, Chair (University of California-Riverside)
0143 Contributed Papers
Convention Centre Room: CC-803B
Main Sponsor: Business and Economic Statistics Section

Presentation: ***Likelihood Inference for Possibly Non-Stationary Processes via Adaptive Overdifferencing***

We make a simple observation that facilitates valid likelihood-based inference for the parameters of the popular ARFIMA or FARIMA model without requiring stationarity by allowing the upper bound for the memory parameter to exceed 0.5. We observe that estimating the parameters of a single non-stationary ARFIMA model is equivalent to estimating the parameters of a sequence of stationary ARFIMA models. This enables improved inference because many standard methods perform poorly when estimates are close to the boundary of the parameter space. It also allows us to leverage the wealth of likelihood approximations that have been introduced for estimating the parameters of a stationary process. We explore how estimation of the memory parameter depends on the upper bound and introduce adaptive procedures for choosing. Via simulations, we examine the performance of our adaptive procedures for estimating the memory parameter when the true value is as large as 2.5. Our adaptive

procedures estimate the memory parameter well, can be used to obtain confidence intervals for the memory parameter that achieve nominal coverage rates, and perform favorably relative to existing alternative

Co-Author(s): David Matteson & **Gennady Samorodnitsky**, Cornell University
First & Presenting Author: Maryclare Griffin

Thursday, Aug 10: 10:30 AM - 12:20 PM

Advances in Bayesian Modeling for Non-Gaussian, Time Series and Complex Structured Data - **Changwoo Lee**, Chair (Texas A&M University)
0155 Contributed Papers
Convention Centre Room: CC-205C
Main Sponsor: Section on Bayesian Statistical Science

Presentation: ***Trend Filtering with Adaptive Bayesian Changepoint Analysis for Count Time Series***

Model development for sequential count-valued data characterized by small counts and non-stationarities is essential for broader applicability and appropriate inference in the scientific community. Specifically, we introduce global-local shrinkage priors into a Bayesian dynamic generalized linear model to adaptively estimate both changepoints and a smooth trend for count time series. We utilize a parsimonious state-space approach to identify a dynamic signal with local parameters to track smoothness of the local mean at each time-step. This setup provides a flexible framework to detect unspecified changepoints in complex series, such as those with large interruptions in local trends. We detail the extension of our approach to time-varying parameter estimation within dynamic Negative Binomial regression analysis to identify structural breaks. Finally, we illustrate our algorithm with empirical examples in social sciences.

Co-Author: David Matteson, Cornell University
First & Presenting Author: Toryn Schafer, Texas A&M University