Conference Program

The 2022 Meeting of

The International Society for Data Science and Analytics

May 31 and June 1, 2022

Notre Dame, IN

& Zoom

Organizing Committee

- Hawjeng Chiou, Distinguished Professor National Taiwan Normal University
- Karl Ho, Associate Professor, University of Texas at Dallas
- Wen Qu, Junior Associate Professor, Fudan University
- Jiashan Tang, Professor, Nanjing University of Posts and Telecommunications
- Ke-Hai Yuan, Professor, University of Notre Dame
- Zhiyong Zhang, Professor, University of Notre Dame

Sponsored by

Lucy Family Institute for Data & Society

UNIVERSITY of NOTRE DAME
Department of Psychology

師大管理學院
College of Management
National Taiwan Normal University

南京郵電大學
Nanjing University of Posts and Telecommunications

INTERNATIONAL SOCIETY FOR DATA SCIENCE AND ANALYTICS

Please contact the organizing committee at meeting@isdsa.org for any feedback.

Schedule

In-Person Location: E102 & E108 Corbett Family Hall, University of Notre Dame
Zoom:
https://notredame.zoom.us/j/94987508545?pwd=QVRHRTEvRnR4bVoyTDV4ODRqalN2Zz09
Meeting ID: 949 8750 8545 Passcode: 548636
Please include your full name (preferably with your institution information) when joining in on Zoom.

May 31, 2022

Time	Speaker and Title
Morning	
8:00-8:45	Registration and badge pickup & light breakfast
8:45-9:00	Opening remarks by the organizing committee
9:00-9:30 zoom	**An Application of the Predator-Prey Model in Psychology: The Dynamic Relationship Between Food Cravings and Unhealthy Snacking** Yueqin Hu*; Beijing Normal University; CN Xiaohui Luo; Beijing Normal University; CN
9:30-10:00 IN PERSON	**Transformations of Continuous Time Dynamic Models into Alternative Discrete Time Models** Sy-Miin Chow*; Pennsylvania State University; US Diane Losardo; Pennsylvania State University; US Jonathan Park; Pennsylvania State University; US Peter Molenaar; Pennsylvania State University; US
10:00-10:30 IN PERSON	**Using Later Retrieval to Handle Missing Data in Ecological Momentary Assessments** Manshu Yang*; University of Rhode Island; US
10:30-10:45	Coffee break
10:45-11:45	*Speed talks* **1. A Comparison of Tree Based Supervised Machine Learning Methods for Prognostication of Patients with Traumatic Brain Injury** zoom Vineet Kumar Kamal, National Institute of Epidemiology, Indian Council of Medical Research, IN R. M. Pandey, Department of Biostatistics, All India Institute of Medical Sciences, IN Deepak Agrawal, Department of Neurosurgery, Neurosciences & Gamma Knife Centre, All India Institute of Medical Sciences, IN

	2. A Growth of Hierarchical Autoregression Model to Capture Interindividual Differences in Intraindividual Changes in Psychological Processes IN PERSON Yanling Li*; The Pennsylvania State University; US Chelsea Muth; US Sy-Miin Chow; The Pennsylvania State University; US Zita Oravecz; The Pennsylvania State University; US **3. Calculating the Economic Impacts of Food Gentrification on Communities of Color in Portland** IN PERSON Karishma Shah*; University of Chicago; US **4. Modeling US-China Trade Relations: A Time Series Machine Learning approach using MNC stock data** zoom Min Shi*; School of Economic, Political and Policy Sciences, The University of Texas at Dallas; US Karl Ho; School of Economic, Political and Policy Sciences, The University of Texas at Dallas; US **5. Text Analytics Models of US News Media: The Case Study of US-China Trade Relations** zoom Wen Si*; The University of Texas at Dallas; US
11:45-1:00	Lunch (provided)
12:15-12:45 IN PERSON	*Lunch talk* **Introduction to Choice Modeling** Meghan Cain*; StataCorp LLC; US
Afternoon	
1:00-1:30 IN PERSON	**Meta-Analysis of Correlation Coefficients: A Cautionary Tale on Treating Measurement Error** Qian Zhang*; Florida State University; US
1:30-2:00 IN PERSON	**Handling Non-Normality in SEM** Han Du*; University of California, Los Angeles; US
2:00-2:30 IN PERSON	**Recent Advancements of Moderation and Mediation Analyses**

	Ke-Hai Yuan*; University of Notre Dame, USA; US Hongyun Liu; Beijing Normal University, China; CN
2:30-2:45	Coffee break
2:45-3:15 zoom	**Practicalities of the Power/sample Size Analysis of Tabulated Data** Oleksandr Ocheredko*; Vinnytsya National Medical University; UA
3:15-3:45 zoom	**The Big-Fish-Little-Pond Effect in Mathematics Classes Across Nations and Over Years** Ze Wang*; University of Missouri; US
3:45-4:15 zoom	**Customising Data Distribution Types for Statistical Power Analyses in R** Antoine Bagnaro*; Department of Marine Science, University of Otago; NZ
4:15-5:15	**Campus tour**
5:15-8:00	**Dinner**
June 1, 2022	
Time	
Morning	
8:00-9:00	Registration and badge pickup & light breakfast
9:00-9:30 zoom	**Bayesian Evaluation of Predictors' Relative Importance** Xin Gu*; East China Normal University; CN
9:30-10:00 zoom	**A New Hybrid Statistical Method for Demographic Analysis** Gloria Gheno*; Ronin Institute; US
10:00-10:30 zoom	**Comparison of Anomaly Detection Methods for Bot Detection in Online Likert-type Questionnaires** Max Turgeon*; University of Manitoba; CA; US
10:30-10:45	Coffee break
10:45-11:45	*Speed talks* **1. Identification of Heterogeneity of Growth Trajectory Using Mixture Modeling with Bayesian Estimation: An Example of Analysis of Wage Trajectory with the Age Period and Cohort Effects** IN PERSON Hawjeng Chiou*; National Taiwan Normal University Department of Business Administration, Department of Educational Psychology and Counseling; TW **2. Sending and Receiving Effect of Personality on Friendship: A Social Network Analysis Approach** IN PERSON

	Haiyan Liu*; University of California, Merced; US Ren Liu; University of California, Merced; US **3. Stochastic Approximation Expectation-Maximization SAEM Algorithm for Fitting Differential Equation Models with Random Effects in dynr** IN PERSON Xiaoyue Xiong*; College of Health and Human Development, The Pennsylvania State University; US Hui-Ju Hung; The Pennsylvania State University; US Sy-Miin Chow; College of Health and Human Development, The Pennsylvania State University; US **4. Comparison of Methods for Imputing Social Network Data** IN PERSON Ziqian Xu*; University of Notre Dame; US Zhiyong Zhang; University of Notre Dame; US Jiarui Hai; Tsinghua University; CN Yutong Yang; Renmin University of China; CN **5. Impact of different school attendance modes on secondary school students' intention to pursue higher studies** zoom Bernice Wong; Kolej Tuanku Ja'afar; MY Joanne Yim*; Tunku Abdul Rahman University College; MY **6. Social Network Analysis in the Framework of Structural Equation Modeling** IN PERSON Zhiyong Zhang*; University of Notre Dame Notre Dame, IN 46556 USA; US
11:45-1:00	Lunch (provided)
12:15-12:45	*Lunch talk* **Causal Mediation Analysis with the Latent Growth Curve Mediation Model** IN PERSON Xiao Liu*; Department of Psychology University of Notre Dame; US Lijuan Wang; Department of Psychology University of Notre Dame; US
Afternoon	
1:00-1:30	**Adaptive Respondent Driven Sampling of Social Networks: A Simulation based Study using Machine Learning** IN PERSON Josey VanOrsdale*; University of Nebraska-Lincoln; US

1:30-2:00	**Isotonic regression and Machine Learning** zoom Karl Ho*; University of Texas at Dallas; US Chuan-Fa Tang; University of Texas at Dallas ; US
2:00-2:30	**The Role of Personality in Trust in Public Policy Automation** zoom Philip Waggoner*; YouGov America & Columbia University; US Ryan Kennedy; University of Houston; US
2:30-2:45	Coffee break
2:45-3:15	**Sleep-Wake Classification of Actigraphy Data: A Machine Learning Approach** IN PERSON Linying Ji*; The Pennsylvania State University; US Meng Liu; The Pennsylvania State University; US Lindsay Master; The Pennsylvania State University; US Orfeu Buxton; The Pennsylvania State University; US Soundar Kumara; The Pennsylvania State University; US Sy-Miin Chow; The Pennsylvania State University; US
3:15-3:45	**A Data-Driven Method for Capturing Comorbidity Structure in Mental Disorders** zoom Hojjatollah Farahani*; Department of Psychology, Tarbiat Modares University; IR Parviz Azadfallah; Department of Psychology, Tarbiat Modares University; IR Peter Watson; Cognition and Brain Sciences Unit, University of Cambridge, UK; GB Marija Blagojević; University of Kragujevac, Faculty of Technical Sciences, Čačak; CS
3:45-4:45	*Speed talks* **1. Can Eves’ Daughters Survive: Massively Unheeded Psychological Intimate Partner Violence Looks Up To Cognitive Dynamics** zoom Samina Ashraf*; UUM University of Utara Malaysia; PK **2. Did Restricting President Trump's Election 2020 Tweets Reduce Their Impact** zoom Zhuofang Li*; California Institute of Technology; US

	3. How did Advertisement Attributes Impact Recruitment of Hard-to-Reach Study Population using Social Media zoom Abraham Liu*; Brown University; US Tyler Wray; Brown University School of Public Health; US Tao Liu; Brown Data Science Initiative, Brown University; US **4. Household Poverty Levels in Namibia and Their Associated Sociodemographic Factors: A Statistical Analysis of the Namibia Household Income and Expenditure Survey** zoom Opeoluwa Oyedele*; University of Namibia; NA **5. A Note on Extreme Value Analysis of Mortality at the Oldest Ages** zoom Yuancheng Si*; Anhui Agricultural University and Bank of Huzhou; CN Zezhi Tang; University of Sheffield; UK **6. Robust Bayesian Analysis of Longitudinal Data using Conditional Quantiles** IN PERSON Xin Tong*; University of Virginia; US
	End of the conference

Directions and Lodging

Our meeting is held in Room **E108 Corbett Family Hall**, University of Notre Dame, Notre Dame, IN 46556.

Getting Here

<u>By Air</u>

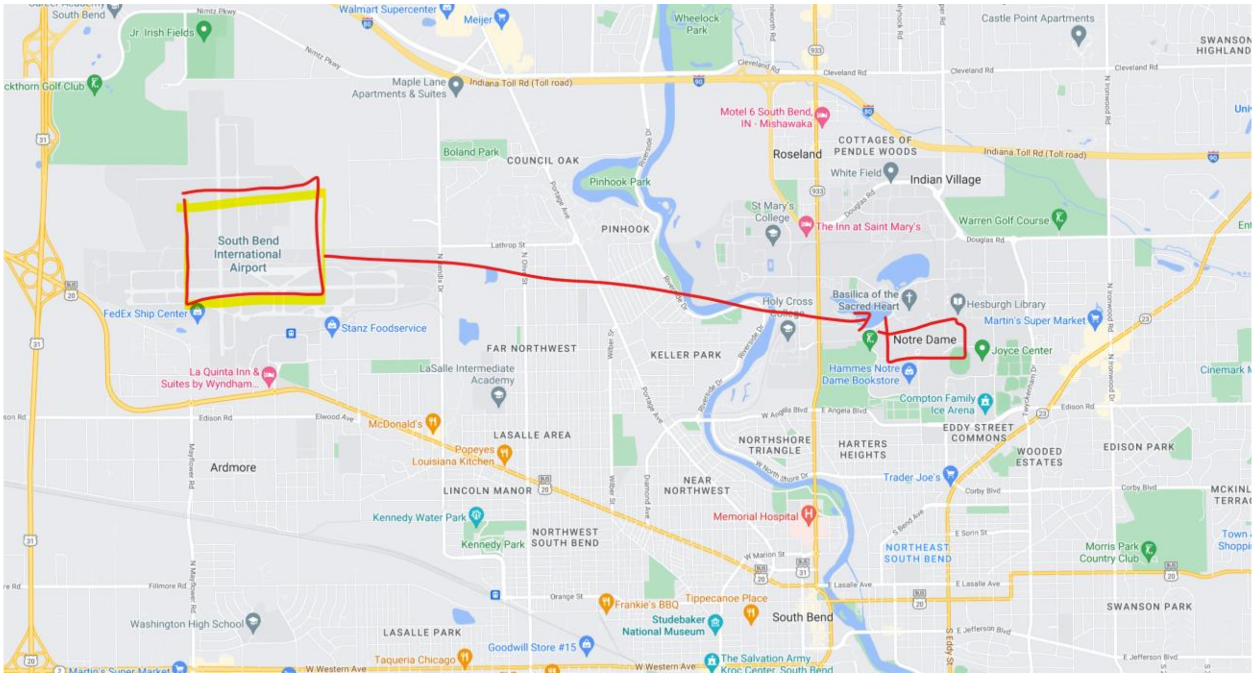
South Bend International Airport is about 15 minutes by car from the Notre Dame campus (flights should be booked to South Bend, Indiana -- airport code SBN). Various transportation methods are available (e.g., taxi, rental car, limo).

Visitors also can fly into Chicago then drive or take a bus to Notre Dame. The University is about two hours by car from Chicago’s O’Hare International Airport and about 90 minutes from Midway International Airport.

<u>By Train</u>

The South Shore Line trains run directly from the Chicago Loop (at the corner of Michigan and Randolph) to South Bend International Airport (about a three-hour trip). From the airport, the Notre Dame campus is approximately a 15-minute ride by car. Various transportation methods are available (e.g., taxi, rental car, limo).

More information can be found here: https://www.nd.edu/visit/

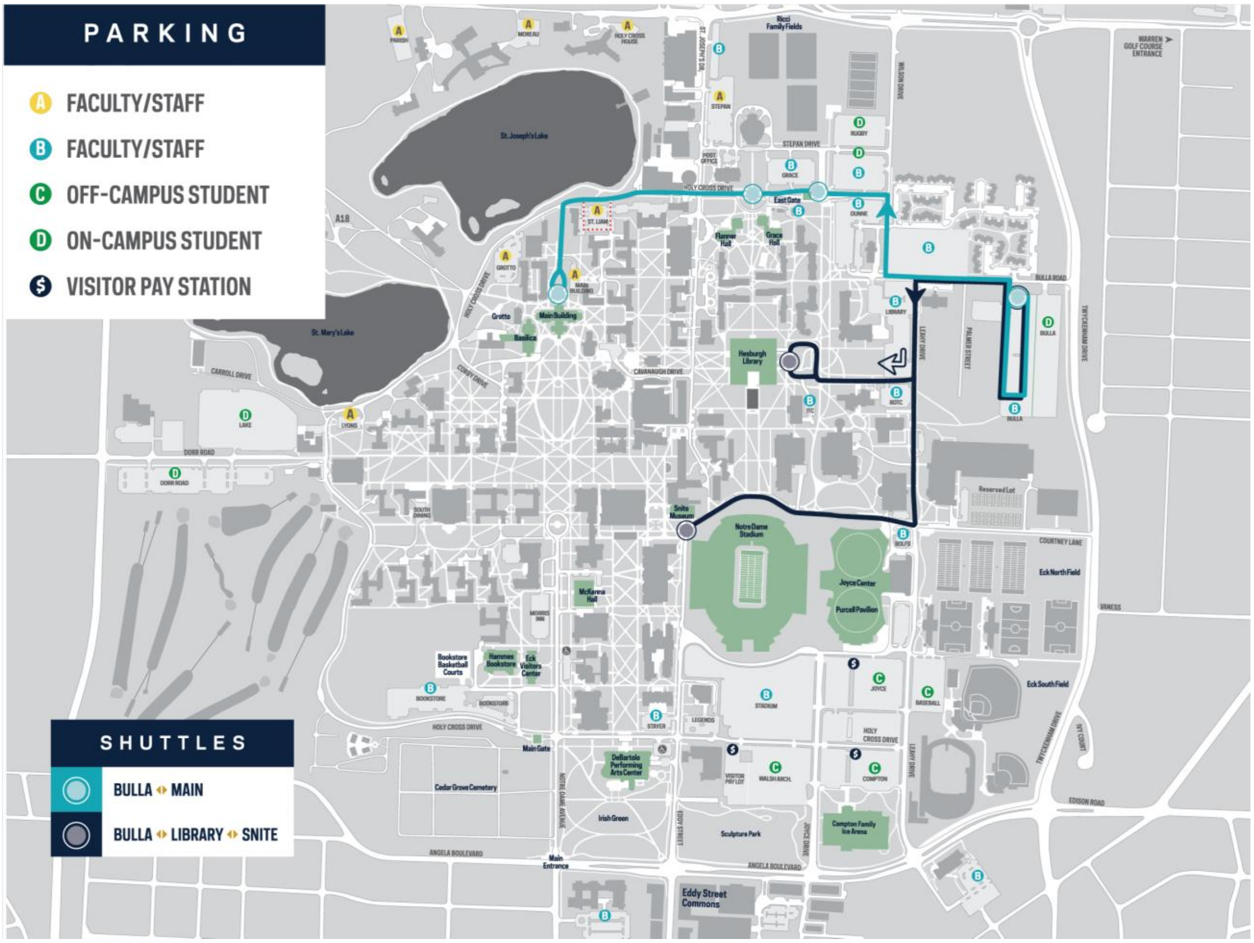


Parking

You can park in the Joyce Lot (just south of the entrance to Purcell Pavilion; see map below). During regular business hours (Monday-Friday, 6 a.m. - 4 p.m.), visitors must purchase a permit at a pay station (credit cards only). The permit must be displayed face up on the driver's side of the vehicle's dashboard, fully visible to parking enforcement staff.

The current rate schedule is:

- 1hr - Free (must obtain permit and display on dash)
- 2hrs - $1
- 3hrs - $2
- 4hrs - $3
- 4+hrs - $8



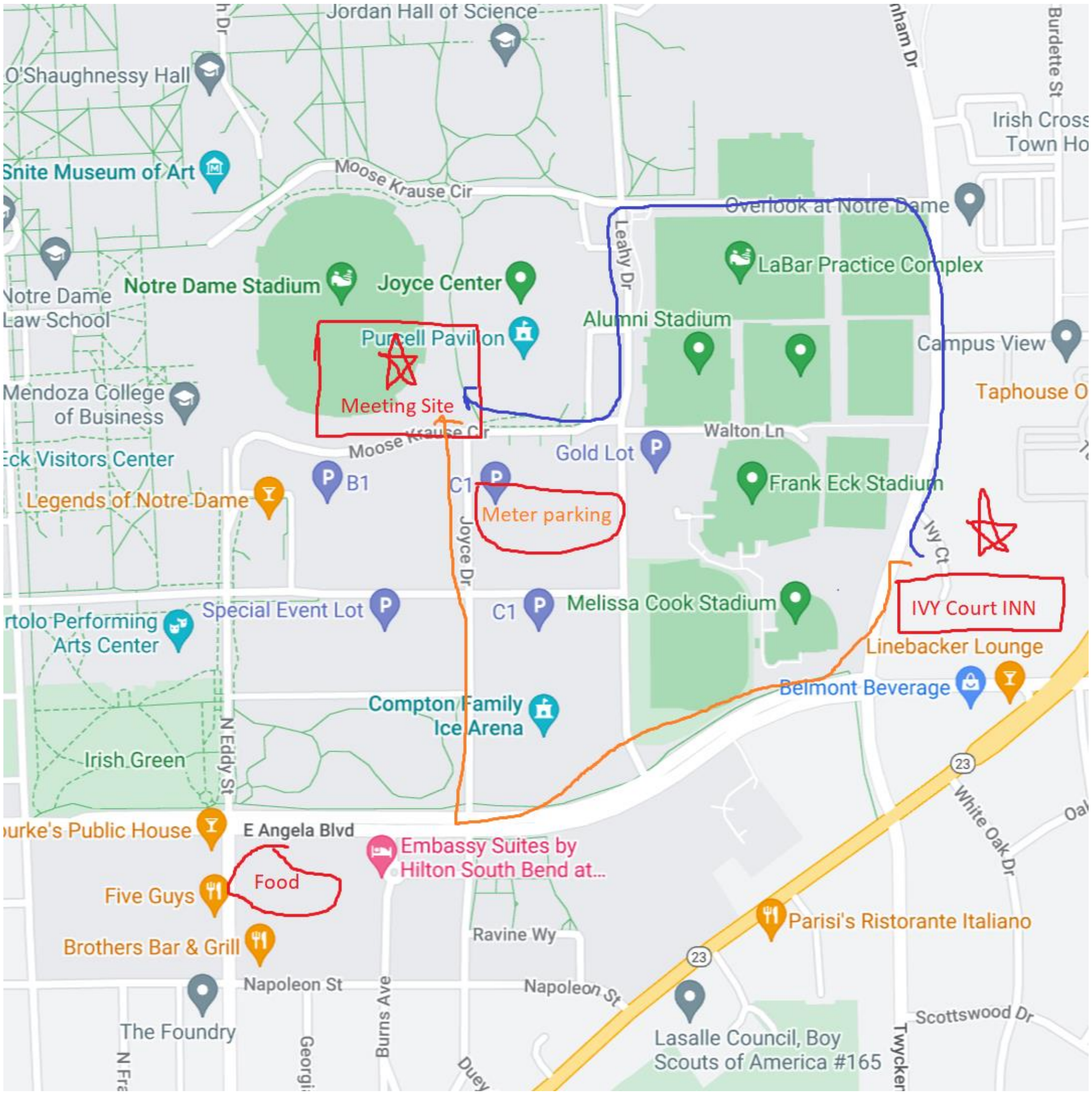
Hotels

We have blocked some rooms at Ivy Court Inn & Suites with the best rate. You can book your room directly on its website. When checking out, input the code "22-157" as the "Group Attendance" code. Please book your room by May 16. Rate will increase after it.

IVY COURT INN & SUITES

Price: $110+tax for standard double and king rooms; $134+tax for King suites.
Address: 1404 Ivy Ct, South Bend, IN 46637
Phone: (574) 277-6500
Website: https://www.ivycourt.com/

The hotel is within walking distance to our meeting site.



Food

We will provide light breakfast and boxed lunch at the meeting site. There will be a dinner on May 31, 2022.

Below is a list of restaurants around the meeting site.

- Rohr’s, American Food, on campus, 1399 N Notre Dame Ave, South Bend, IN 46617
- Modern Market, on campus, Duncan Student Center, South Bend, IN 46617
- O'Rourke's Public House, 1044 E Angela Blvd #103, South Bend, IN 46617
- Blaze Pizza, N Eddy St #1234, South Bend, IN 46617
- Five Guys, 1233 N Eddy St Unit 10, South Bend, IN 46617
- J.W. Chen's, 1835 S Bend Ave, South Bend, IN 46637
- Cre-Asian, 1639 N Ironwood Dr, South Bend, IN 46635

Abstracts

Can Eves' Daughters Survive: Massively Unheeded Psychological Intimate Partner Violence Looks Up To Cognitive Dynamics

Samina Ashraf*; UUM University of Utara Malaysia; PK

Intimate partner violence (IPV) is a soaring matter in social science works. Most of the researches have focused on its aftermath. Nonetheless, there is scarcity of investigations that determines how to reduce IPV with cognitive Dynamics. Consequently, the current research intends to measure the impact of humor on PIPV under the moderation of self-efficacy. In order to meet the objectives of the study, the data were composed from 504 battered working women in Pakistani women colleges through self-administered surveys. The data were analyzed by employing SPSS approach using SPSS v23. The outcomes of the investigation elucidated a negative link between humor and PIPV. The research also utilized full moderation of self-efficacy amid the relationship of humor with PIPV. The study also explains the research implications, limitations and future directions in the last section.

Customising data distribution types for statistical power analyses in R

Antoine Bagnaro*; Department of Marine Science, Te Tari Pūtaiao Taimoana,
University of Otago, Te Whare Wānanga o Otāgo; NZ

The performance of statistical tests is usually assessed with two different metrics: their power and their significance. Statistical significance has long been integrated in ecological studies. Statistical power, however, remains a marginal metric. It nonetheless gives confidence in deciding whether the results of a test may be extrapolated at the population level – or at the regional level – which is crucial for the design of appropriate ecological monitoring studies.

Ecological data often rely on the census of particular species of interest in a number of selected locations. The density of these species (in number of individuals per unit of area) or other ecologically relevant measures (e.g. weight, size…) rarely follow normal, unimodal distribution types. Adapting statistical power tests to accommodate distribution types that are relevant to every scenario is key to ensure their accuracy for ecological monitoring.

The aim of this short case-study will be to provide a workflow for the design of user-defined distribution types for statistical power analyses. Particular attention will be given to the determination of the necessary sampling intensities for cases of non-normal, zero-inflated, and multi-modal data distributions, with graphical output examples. All analyses will be performed with the help of the R software.

Introduction to Choice Modeling

Meghan Cain*; StataCorp LLC; US

Discrete choice models are used across disciplines to analyze choice behavior, for example: voters choose their candidate or party, commuters choose a mode of transportation, college freshman choose majors, and employers choose job candidates. In all of these cases, we observe decision making entities that are faced with a set of alternatives to choose from. Discrete choice models with alternative-specific variables allows for including variables that can vary both over decision makers as well as choice sets. In other words, we can incorporate attributes of the decision maker as well as attributes of the alternatives into our analysis.

In recent years, it's become apparent that psychological factors, such as emotional states, attitudes, and personality characteristics, are important attributes of the decision maker that need to be considered when modeling many types of choices. This talk will provide a quick introduction to choice modeling, and a demonstration of fitting and interpreting choice models in Stata.

Identification of heterogeneity of growth trajectory using mixture modeling with Bayesian estimation: An example of analysis of wage trajectory with the age period and cohort effects

Hawjeng Chiou*; National Taiwan Normal University Department of Business Administration, Department of Educational Psychology and Counseling; TW

The major advantage of mixture modeling applying to longitudinal data analysis is to identify the unobserved heterogeneity in the development of an outcome over time. It provides the technical solution for longitudinal study to deal with the dependency between time-(in)variant variables with heterogeneity. For example, as a traditional practical issue in human resource management field, the growth pattern of wage for different populations could be separated and affected by the age, period, and cohort in certain ways. However, the specification of latent classes of trajectory with covariates increase of complexity of model and the amount of parameters being estimated dramatically. The missing values also cause the difficulty of estimation. According to literatures, Bayesian estimation is one of the alternative to maximum likelihood approach for dealing with these issues. In order to examine the efficiency of Bayesian estimates in mixture growth modeling for real data, this study selected 5851 Taiwanese adults with age from 25 to 65, acquired 32816 observations from the Panel Study of Family Dynamics dataset at the 1999 to 2018 period. To figure out the mechanism of wage profile and dispersion affected by APC variables, this study fit a series of latent growth model to the data first, following by the identification of heterogeneity using Bayesian estimates and comparing with ML approach using Mplus8. Results identified three latent classes of wage trajectory. The major group (94%) shown a curvilinear trajectory with positive quadratic term. The other two groups have a negative quadratic term with different wage level at the intercept term. Although the covariates reveal different effects for the three trajectories, in general, the three covariates age, cohort, and period along with gender and educational level have significant effects on wage growth. The cohort effect with wage change revealed a relative stable curve relationship while the model only included cohort variable alone. After taken age and period into account, cohort effect of wage was not statistically significant. Gender as well as educational year can significantly predict wage trajectory, as the same findings for work hour per week. Bayesian estimation were superiors over ML estimates for fitting the mixture models with missing data, but also shown the

efficiency for examining the covariates and APC effects. Beyond the methodological implementation, the present study has contribution to utilize the existing secondary dataset and to provide constructive findings about wage change overtime for applications of human resource management in practice.

Transformations of Continuous Time Dynamic Models into Alternative Discrete Time Models

Sy-Miin Chow*; Pennsylvania State University; US
Diane Losardo
Jonathan Park
Peter Molenaar

Irregularly spaced longitudinal data often arise in experience sampling studies that use partially random sampling intervals to capture the participants' status *in the moment*. Many structural equation modeling (SEM) approaches for fitting longitudinal or dynamic models to intensive longitudinal data treat the time intervals between successive occasions as equally spaced (e.g., in computing lagged covariance or correlation matrices for model fitting purposes) and are not well suited for use with irregularly spaced data. Several authors have introduced continuous-time models in the form of linear stochastic differential equation (SDE) models as a way to accommodate such irregularly spaced time intervals, and discussed their parallels with the SEM framework. Unfortunately, the relations between SDEs and their discrete time counterparts such as vector autoregression (VAR) and structural VAR (SVAR) models are not well understood, and these models are sometimes regarded as completely distinct modeling options. In this talk, we present and discuss the relations and transformation functions for mapping linear SDE models to VAR and SVAR. Code and demonstrations for fitting these models to irregularly spaced data using an R package, *dynr*, are provided, followed by discussions of some of the caveats, challenges, and possible extensions to leverage these transformations to fit continuous-time dynamic network models.

Handling Non-Normality in SEM

Han Du*; University of California, Los Angeles; US

The mainstream estimators for structural equation modeling are based on normal theory, but real data are unlikely to be exactly normally distributed. To improve estimation and inference with non-normal data, I will introduce two studies that I work on in recent 3 years. In Study 1, my colleagues and I propose a distributionally weighted least squares (DLS) estimator for parameter estimation (Du & Bentler, in press; Du et al., 2022). We find that DLS works well with both normal and non-normal data in both factor analysis and growth curve models. It generally can provide more accurate and efficient estimates than the classical estimators. In Study 2, to help evaluate model fit, we propose to use an unbiased distribution free weight matrix estimator in robust test statistics and extend it to models with mean structures (Du & Bentler, under review; Du, under review). We find that the Satorra–Bentler statistic and Hayakawa’s T_MVA2 coupled with unbiased distribution free weight matrix could control Type I error rates better than other statistics.

A Data-Driven Method for Capturing Comorbidity Structure in Mental Disorders

Hojjatollah Farahani*; Department of Psychology, Tarbiat Modares University; IR
Parviz Azadfallah; Department of Psychology, Tarbiat Modares University; IR
Peter Watson; Cognition and Brain Sciences Unit, University of Cambridge, UK; GB
Marija Blagojević; University of Kragujevac, Faculty of Technical Sciences, Čačak; CS

The concurrent presence of a mental disorder with another mental disorder is common in the clinical practice of comorbidity structure research. In this study we look at the structure of the comorbidity, assessing the degree of overlap among the measured signs and symptoms of two mental disorders. In this paper, the newly advanced and graphical statistical method of network analysis is introduced and described. This data driven method helps mind researchers to be able to capture the most important relationships among variables in a complex and complicated system. The stages for running the network analysis using R software are explained. Accuracy testing and stability centrality measures are investigated using bootstrapping. As a practical example, this method was used on the data obtained from 254 Multiple sclerosis (MS) patients to capture the comorbidity structure between depressive and anxiety symptoms. The results are presented and discussed. Network analysis as a data-driven based model can be of interest to all mind researchers especially the researchers working in clinical, cognitive and social psychology.

A new hybrid statistical method for demographic analysis

Gloria Gheno*; Ronin Institute; US

The analysis of the evolution of the population is fundamental for the public administration, which must necessarily have projections, not only on the total global population, but also the division into single age groups, as each of them represents the number of potential future users of the various types of services which it provides. For example, it is effective to predict the number of school-age children in order to size schools and ancillary services in good time. For this reason, the scholars propose new methods to study demographic. To estimate the development of the future population, three main demographic factors are considered: mortality, fertility and migration. Hunsinger (2010) proposed an ARIMA model to analyse these three factors, adding the possibility of assuming certain values, suggested by experts in the field, to the parameters of the classic AR(1) model. Vollset et al. (2020) propose a mathematically well-structured model to estimate the future population. To estimate the 3 demographic factors mentioned above, Vollset et al. (2020) additional variables, considering so other causes of population evolution. This paper had a lot of notoriety but had also critics because of the presence or absence of some demographic assumptions (O'Sullivan, 2021). My work links these two models, considering not only the observations of the experts, but adding also other determined factors present in the literature. This new model is applied to some European Countries, with demographic differentiations, to see its adaptation to various situations, so as to help improve the future developments of Countries.

Hunsinger, E. (2010). An expert-based stochastic population forecast for Alaska, using autoregressive models with random coefficients. SSRN Electronic Journal, 9.

O' Sullivan, J.N. (2021). Trends in population health and demography. The Lancet, 398(10300)

Vollset, S. E., Goren, E., Yuan, C. W., Cao, J., Smith, A. E., Hsiao, T., ... & Murray, C. J. (2020). Fertility, mortality, migration, and population scenarios for 195 countries and territories from 2017 to 2100: a forecasting analysis for the Global Burden of Disease Study. The Lancet, 396(10258), 1285-1306.

Bayesian evaluation of predictors' relative importance

Xin Gu*; East China Normal University; CN

This study develops a Bayesian evaluation approach for the relative importance of predictors. Several importance measures are introduced, among which two are highlighted: dominance analysis measures and relative weights. Using these two measures, the contribution to the variation of the outcome variable is attributed to each predictor. The importance indices are often computed through the R-squared change by adding a predictor into a model, which can be derived from the correlation matrix of the data. Researchers' theories about relative importance are represented by order constrained hypotheses. Support for or against the hypothesis is quantified by the Bayes factor. The proposed method is extended to assessing the relative importance of latent predictors, where the model implied correlation matrix of the latent variables can be used to calculate the R-squared change, the importance measure and the Bayes factor. Simulation studies are conducted to investigate the performance and properties of the proposed method. A real data example is used to illustrate how the relative importance can be evaluated.

Isotonic regression and Machine Learning

Karl Ho*; University of Texas at Dallas; US
Chuan-Fa Tang; University of Texas at Dallas ; US

In this study, we review the rationale behind the family of isotonic (monotone) regression methods (Barlow and Brunk 1972; Dougherty 2015; Leeuw, Hornik and Mair 2009, Tibshirani, Hoefling and Tibsibrani 2011) and compare them with traditional supervised and unsupervised machine learning methods. We further illustrate how to apply the methods in social and political data sets and evaluate them from a machine learning perspective.

An Application of the Predator-Prey Model in Psychology: The Dynamic Relationship Between Food Cravings and Unhealthy Snacking

Yueqin Hu*; Beijing Normal University; CN
Xiaohui Luo; Beijing Normal University; CN

The predator-prey model is widely used in various disciplines, but not yet in psychology, where the competitive interactions featured in this model are actually not uncommon. This study attempts to introduce this model into psychology, using the relationship between food cravings and unhealthy snacking as an example. Self-efficacy, a determinant of health behavior, was used

as a moderator of this relationship. Fifty-nine female undergraduates completed a self-efficacy scale and a seven-day ecological momentary assessment in which they reported craving intensity and snack consumption five times per day. Results showed that unhealthy snacking acted as the predator that "preyed" on food cravings, that is, growing desire for food stimulated snacking, and snacking subsequently reduced food cravings. For individuals with low self-efficacy, food cravings led to more subsequent consumption of unhealthy snacks. The predator-prey model reflects a facilitation-inhibition-bidirectional relationship or negative feedback mechanism, which should have broader applications in behavioral science

Sleep-Wake Classification of Actigraphy Data: A Machine Learning Approach

Linying Ji*; The Pennsylvania State University; US
Meng Liu; The Pennsylvania State University; US
Lindsay Master; The Pennsylvania State University; US
Orfeu Buxton; The Pennsylvania State University; US
Soundar Kumara; The Pennsylvania State University; US
Sy-Miin Chow; The Pennsylvania State University; US

Background: Actigraphy has been widely used as an unobtrusive and relatively low-cost way to collect objective sleep data. However, compared with polysomnography (PSG), the current gold standard for measuring sleep, researchers found the typical algorithms normally used to identify the sleep/wake status of actigraphy data was less sensitive in identifying wake, and tends to over-classify sleep. As a result, it is necessary to develop better models to enable more accurate classification of sleep-wake status using actigraphy.

Purpose: Leveraging the previous classification results offered by extreme-gradient boosting(XGB), a machine learning approach, we aim to 1) explore whether and in what ways adding time series and dynamical systems-inspired data features to the XGB classifier can help improve classification results; and 2) evaluate ways to improve the robustness of the classifier through addition of cross-validation and dimensionality reduction procedures.

Participants and methods: Actigraphy and PSG data were collected from 54 participants in four different experimental studies. The number of sleeping periods ranged from 3 to 11 per participant. Actigraphy data were collected with 30-seconds epochs. We included actigraphy data features such as summary statistics (e.g. mean, percentiles, standard deviations, etc.), dynamic features (e.g. entropy, max level/variance shift, Hurst, etc.), and lagged activity count records. Different dimension reduction methods, including t-SNE, a nonlinear embedding technique, and principal component analysis were compared and evaluated. A three-fold rolling window cross-validation was used to improve the robustness of the classifier during optimization with training data. Area Under the Curve (AUC) was used as the metric to drive parameters tuning handling the imbalanced sleep-wake status.

Results: Variable importance analysis shows that dynamic features, such as max level shift, max Kulback-Leibler shift, and standard deviation of the first derivative of the time series applied to relatively large window sizes (i.e. 121, 30-s epochs, ~ 1h) ranked high in their importance in predicting sleep-wake status. The machine learning model we developed using XGB and cross-validation greatly improved the classification performance from the device manufacturer (Actiware versions 5.57 and 5.59, Philips-Respironics). Specifically, AUC of our approach is .85 (manufacturer: 0.69), balanced accuracy is 0.76 (manufacturer: 0.69). In addition, our approach performed much better in identifying wake during sleep periods with specificity of 0.74 (manufacturer: 0.43). See Supplemental Table 1.

Conclusion: Machine learning classifiers of sleep-wake states greatly improved the manufacturer's classification performance with PSG-derived sleep-wake states as ground truth. Including dynamic features derived from longer time windows was particularly helpful in optimizing model performance and predicting sleep health attributes, such as total sleep time and wake-after-sleep onset.

A comparison of Tree based supervised machine learning methods for prognostication of patients with traumatic brain injury

Vineet Kumar Kamal, National Institute of Epidemiology, Indian Council of Medical Research, India
R. M. Pandey, Department of Biostatistics, All India Institute of Medical Sciences, New Delhi -110029, India
Deepak Agrawal, Department of Neurosurgery, Neurosciences & Gamma Knife Centre, All India Institute of Medical Sciences, New Delhi -110029, India

Introduction: Traumatic brain injury (TBI) is a significant public health problem in all regions of the globe, especially in developing nations. Statistical modelling is essential for prognostication, hypothesis generation and stratification of patients in research studies in case of TBI. Machine learning (ML) has been successfully applied to give support to clinical diagnosis and prognosis prediction. Tree based ML techniques allow us to build accurate predictive models with even uncontrolled data (i.e. missing values, lots of variables, nonlinear relationships, outliers etc.).

Objective: To compare tree-based supervised machine learning methods: Classification and Regression Tree (CART), Stochastic Gradient Boosting (GB), and Random Forest (RF) in order to deliver enhanced predictions of outcome in future patients with traumatic brain injury using routinely collected demographic, clinical, CT variables, and laboratory variables.

Methods: We retrospectively used trauma database of India's largest level I trauma centre. For development of models, this study included all patients (n=1466) with moderate and severe TBI admitted to emergency department and later on, shifted to ICU under the Department of Neurosurgery during May 19, 2010 to July 31, 2012 using logistic regression (LR), CART, Gradient Boosting, and Random Forest to predict In-hospital mortality and Unfavourable outcome at 6-months. Further, a data set of all eligible patients (n=316) from the same hospital were used for external validation. Comparisons of models were done in terms of discrimination (Area under ROC) and calibration ability (graphical representation of observed vs predicted) in both development and external validation data set.

Results: For LR, CART, GB, and RF in external validation data set; area under the ROC curves (95% CI) were 0.85 (0.80, 0.90), 0.80 (0.75, 0.86), 0.88(0.83,0.92), and 0.78 (0.74, 0.83), respectively to predict In-hospital mortality; and 0.90 (0.86, 0.94), 0.83 (0.79, 0.88), 0.89 (0.85, 0.93), and 0.90 (0.86, 0.94), respectively to predict Unfavourable outcome. the Gradient boosting had the best agreement between observed and predicted outcome followed by LR, Random Forest and CART.

Conclusion: Gradient boosting ML techniques seems to be the best method for prognostication of patients with TBI. All these methods outperformed the stand-alone CART to predict long term outcome. We can rely on Tree based modern machine learning algorithms - GB and RF to predict the future outcome and to address the disadvantages of logistic regression in case of TBI in Indian and similar settings.

A Growth of Hierarchical Autoregression Model to Capture Interindividual Differences in Intraindividual Changes in Psychological Processes

Yanling Li*; The Pennsylvania State University; US
Chelsea Muth; US
Sy-Miin Chow; The Pennsylvania State University; US
Zita Oravecz; The Pennsylvania State University; US

Quantitative researchers have successfully employed dynamical systems models to measure a set of three dynamic characteristics, which collectively summarize the unique dynamic profiles of psychological phenomena such as cognitive processing and affective experience. As demonstrated in Figure 1 in the supplementary material, these dynamic characteristics can be defined as: (1) baseline: representing a person's average levels of experience; (2) intraindividual variability (IIV): representing how dramatically or imperceptibly a person's level of experience fluctuates in momentary experience; and (3) regulation: representing how quickly or slowly a person returns to their baseline levels.

Importantly, these dynamic features may fundamentally change over time on a longer timescale – that is developmental changes might occur over the course of weeks, months, or years. It is also reasonable to expect developmental changes of these dynamic features over the course of an intervention, as illustrated by the pre-to-post intervention change in Figure 1. Although recent methodological developments such as dynamical models with time-varying parameters allow for changes on multiple time scales, few of them simultaneously account for developmental changes of multifaceted characteristics of human processes, such as the three dynamic characteristics mentioned above.

In this study, we propose a growth of hierarchical autoregression (GoHiAR) model which combines autoregressive (AR) and growth curve models (GCM) to simultaneously evaluate developmental changes in dynamic characteristics (i.e., baseline, regulation, and IIV) and individual differences therein. Specifically, GCM models are fitted to dynamic AR model parameters, including the intercept, AR parameter, and process noise variance, to respectively account for changes of baseline, regulation, and IIV over the long-term timescale and individual differences. The Bayesian estimation framework allows for all model parameters to be estimated at once to avoid accumulation of estimation errors caused by two-step estimation approaches. Different from standard dynamical models with time-varying parameters, the GoHiAR model allows for both person-specific and time-varying parameters to simultaneously account for changes of dynamic parameters and between-person differences. In addition, the GoHiAR model extends the autoregressive latent trajectory (ALT) model by allowing for not only time-varying intercepts but also time-varying AR parameters and process noise variances.

We note that the proposed model also offers a more comprehensive modeling framework for exploring the efficacy of interventions. In contrast to traditional intervention evaluations that merely focus on baseline changes over time, the proposed model allows for evaluation of between-group differences (i.e., control vs. intervention group) in a set of dynamic parameters (i.e., intercept, AR parameter, and process noise variance) that flesh out dynamic characteristics (i.e., baseline, regulation, and IIV) of intervention outcomes. Further, the multilevel structure of the GoHiAR model makes it straightforward to include person-level predictors and moderators on changes of dynamic characteristics to explore individual differences related to heterogenous intervention outcomes. In our empirical illustration, the GoHiAR model was fitted to data from an 8-week mHealth intervention study (n=160) to investigate changes of dynamic characteristics of "Meaning of Life", defined as the extent to which one felt belonging to something larger than him/herself (e.g., participating in community activities). Findings suggested that participants generally experienced decreased fluctuation and increased regulation of Meaning of Life over the course of the intervention, and the mHealth interventions did not have added benefit over digital monitoring (control group) at the group level.

Did Restricting President Trump's Election 2020 Tweets Reduce Their Impact

Zhuofang Li*; California Institute of Technology; US

We examine whether Twitter's actions to label or restrict the circulation of former President Trump's tweets during the final stages of the 2020 election affected subsequent conversations about election and voting fraud online. We evaluate the impact of Trump's tweets, with and without an action taken by Twitter, using a dataset of over 15 million tweets. Using this unique dataset, we use sophisticated time series methods to estimate whether Twitter's restrictions on Trump's tweets between November 3 and November 11, 2021 influenced the subsequent amount of discussion on Twitter about election and voting fraud, the overall sentiment of this discussion, and the topics being discussed. We also look at the partisan differences of these effects. We find that controlling for the time series nature of the data, there is little evidence that suggests Twitter's action had a significant impact on Republican's discussion of election and voter fraud online. However, Twitter's action works in reducing Democrats' election fraud discussion and alleviating negative sentiment and toxicity for Twitter users of both parties. We conclude our paper discussing the methodological and substantive implications of our research.

How did Advertisement Attributes Impact Recruitment of Hard-to-Reach Study Population using Social Media

Abraham Liu*; Brown University; US
Tyler Wray; Behavioral and Social Sciences, Center for Alcohol and Addiction Studies, Brown

University School of Public Health; US
Tao Liu; Department of Biostatisticsm Center for Statistical Sciences, Brown Data Science Initiative, Center for AIDS Research, Advance-Clinical and Translational Research, Brown University School of Public Health; US

We visit the important problem of classification through advertisements. To be clear, we wish to understand which aspects and features of ads best classify those who were screened for a nationwide study registry and those who enrolled into said study among the people who were eligible to. Furthermore, we want to dive deeper into this subject and do a subgroup analysis on said eligible people conditioning on African American and Hispanic participants. Several different methods of analysis were used, which include univariate logistic regression, multivariate logistic regression, LASSO logistic regression, and random forest classification algorithms. Based on the analysis, it was found that certain features like bright colors within ads and the inclusion of multiracial men were consistently the most important features within our population demographic.

Sending and Receiving Effect of Personality on Friendship: A Social Network Analysis Approach

Haiyan Liu*; University of California, Merced; US
Ren Liu; University of California, Merced; US

Social and personality psychologists have traditionally studied the relationship between personality similarity and friendship. Traditionally, researchers investigated how personality similarity impacts the present/absence of friendship. However, the self-reported or perceived friendship is often asymmetry. In the current study, we first discuss several similarity measures of personality similarity. We then propose a model to investigate the effect of personality similarity in asymmetry friendships. We demonstrate the application of the proposed model in analyzing a directional friendship network of college students.

Causal mediation analysis with the latent growth curve mediation model

Xiao Liu*; Department of Psychology University of Notre Dame; US
Lijuan Wang; Department of Psychology University of Notre Dame; US

Longitudinal mediation analysis is becoming increasingly popular (Mackinnon & Fairchild, 2009). Among different longitudinal mediation models, a popular one is the latent growth curve mediation model (LGCMM; Cheong et al., 2003). With the LGCMM, explicit models are specified for the change trajectories of time-varying variables, permitting the evaluation of complex mediation hypotheses such as treatment influences the level (i.e., intercept) and change (i.e., slope) of the outcome through influencing the level and change of the mediator. Despite the popularity, causal mediation analysis with the LGCMM has received limited methodological attention. To our best knowledge, the only exception is Sullivan et al. (2021). Using the potential outcomes framework, Sullivan et al. (2021) defined and identified the natural indirect effect of

treatment on outcome transmitted through the mediator intercept and slope jointly. However, for the indirect effects via the mediator intercept alone and via the mediator slope alone, which are often of researchers' interest, the causal interpretation and identification assumptions have not been previously investigated.

In this study, we develop methods for causal mediation analysis with the bivariate LGCMM, in which the treatment variable is time-invariant, whereas the mediator and outcome are both time-varying and are measured at multiple time points. Using the potential outcomes framework, we examine the causal interpretation of the indirect effects of the treatment via the mediator intercept/slope alone. A major challenge is that in the LGCMM, the mediator intercept (slope) can be a post-treatment confounder of the relationship between outcome and mediator slope (intercept), rendering the natural indirect effect through the mediator slope (intercept) unidentifiable. To overcome this challenge, we define the interventional (in)direct effects for the LGCMM based on Vansteelandt and Daniel (2017). Such interventional indirect effects have causal interpretation even in the presence of posttreatment confounding and can capture the effects along the pathway through each of the mediator intercept and slope. Furthermore, we define and identify an indirect effect not attributable to either the mediator intercept or slope alone but due to their mutual dependence. This interventional indirect effect would exist when the mediator intercept and slope interact to influence the outcome and have heterogeneous residual covariance between treatment vs. control groups.

Furthermore, for effect estimation, we develop the interaction LGCMM, allowing for the presence of interactions among treatment, mediator intercept, and mediator slope. In particular, for assessing the indirect effects due to the mediator intercept and slope's mutual dependence, extending the approach in Loh et al. (2021), we allow the residual covariance of the latent mediator intercept and slope to depend on the treatment variable. A Bayesian approach is developed to estimate parameters in the model. Preliminary simulation results show that when the number of time points is not too small (e.g., greater than 3), the Bayesian approach yielded satisfactory inference results for the interventional (in)direct effects from the interaction LGCMM. The results also demonstrate that when there are true interaction effects between the mediator intercept and slope, the traditional LGCMM ignoring the interactions can produce biased estimates and inaccurate inference results of the interventional (in)direct effects. The developed method is applied to data from an empirical longitudinal study for illustration.

In conclusion, employing the potential outcomes framework, the current study provides insights on the causal mediation effects in LGCMMs. Guided by the insights, a new approach – the interaction LGCMM – is developed for causal mediation analysis with this model, which offers researchers a useful tool for disentangling the longitudinal pathways underlying treatment effects.

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology, 51(6), 1173. https://doi.org/10.1037/0022-3514.51.6.1173

Cheong, J., MacKinnon, D. P., & Khoo, S. T. (2003). Investigation of mediational processes using parallel process latent growth curve modeling. Structural Equation Modeling: A Multidisciplinary Journal, 10(2), 238–262. https://doi.org/10.1207/s15328007sem1002

Loh, W. W., Moerkerke, B., Loeys, T., & Vansteelandt, S. (2021). Disentangling indirect effects through multiple mediators without assuming any causal structure among the mediators. Psychological methods.

MacKinnon, D. P., & Fairchild, A. J. (2009). Current directions in mediation analysis. Current Directions in Psychological Science, 18(1), 16–20. https://doi.org/10.1111/j.1467-8721.2009.01598.x

Sullivan, A. J., Gunzler, D. D., Morris, N., & VanderWeele, T. J. (2021). Longitudinal mediation analysis with latent growth curves. arXiv preprint arXiv:2103.05765. https://arxiv.org/abs/2103.05765

Practicalities of the power/sample size analysis of tabulated data

Oleksandr Ocheredko*; Vinnytsya National Medical University; UA

Power/sample size analysis is of particular interest in applied tools of researcher. Previously I delivered suggestion on enrichment of statistical tools by combination of bootstrap and MCMC modeling. Novelty suggests application of possible data generation mechanism using MCMC and power estimation in bootstrap procedure. I delineated further generalizations that are not incorporated in statistical software yet and demonstrated basic applications. Power/sample size analysis of tabulated data is of special value due to pervasive nature of grouped \ cross-classified data in support of hypotheses. The lack of available and reliable software is still disappointing. While building it I was confronted with challenges of the flexibility of model formulation and of reliability of covariance matrix estimates. In the paper I discuss GLM and NLS formulations, Levenberg–Marquardt algorithm that is used, practicalities and capacities of R package ltable.

A hybrid estimation of distribution algorithm for joint stratification and sample allocation

Mervyn OLuing*; Insight Centre for Data Analytics; IE

In this study, we propose a hybrid estimation of distribution algorithm (HEDA) to solve the joint stratification and sample allocation problem. EDAs are stochastic black-box optimization algorithms which can be used to estimate, build and sample probability models in the search for an optimal stratification. We enhance the exploitation properties of the EDA by adding a simulated annealing algorithm to make it a hybrid EDA. Results of empirical comparisons for atomic and continuous strata show that the HEDA attains the best results found so far, when compared to benchmark tests on the same data using a grouping genetic algorithm, simulated annealing or hill-climbing.

Household poverty levels in Namibia and their associated sociodemographic factors: A statistical analysis of the Namibia household income and expenditure survey

Opeoluwa Oyedele*; University of Namibia; NA

Despite the intervention strategies that have been put in place to fight poverty, Namibia continues to experience prevalence of poverty with large numbers of households still living in poverty conditions and unable to afford the minimum daily essentials for a decent life. In this quantitative cross-sectional study design, the impact of sociodemographic characteristics of households on their poverty levels was statistically analysed using an ordered probit regression on data from the 2015/16 Namibia household income and expenditure survey. Results showed that sociodemographic characteristics such as the types of household dwelling unit, highest education attainment of the head of household, household main language, household tenure and household main source of income had a significant impact on the household's poverty levels. Households living in a mobile home dwelling unit, whose heads had secondary education as their highest educational attainment as well as households that were mortgaged and whose main source of income were from other sources were less likely to be severely household poor and more likely to be household poor. Furthermore, households living in a single-quarters dwelling unit and whose main language were Setswana were more likely to be severely household poor and less likely to be household poor. It is therefore recommended that the Namibian government and policy makers put more efforts in improving the sociodemographic characteristics of households, particularly those living in a single quarter dwelling unit and whose main language were Setswana.

Calculating the Economic Impacts of Food Gentrification on Communities of Color in Portland

Karishma Shah*; University of Chicago; US

While there is much research about the extreme gentrification currently occurring in most major cities around the United States, the economic impacts of food gentrification remain unstudied. This thesis aimed to create a comprehensive definition of food gentrification in relation to Portland's rapidly changing neighborhoods. It explores the cultural and economic impacts of food gentrification in Portland using literature review, data collection, and data analysis. This data shows the quantitative impacts of gentrification in the food industry using statistical modeling and how it contributes to the displacement of communities of color in Portland. This paper intends to prove that food gentrification plays two roles: profiting from cultural appropriation and accelerating or triggering the gentrification of neighborhoods.

Modeling US-China Trade Relations: A Time Series Machine Learning approach using MNC stock data

Min Shi*; School of Economic, Political and Policy Sciences, The University of Texas at Dallas
; US
Karl Ho; School of Economic, Political and Policy Sciences, The University of Texas at Dallas
; US

The U.S.-China trade war has worsened U.S. and China relations in the past few years. The imposed tariffs from both sides impede the normal global value chains (GVCs) circulation, increase the manufacturing cost, and further reduce the profit margins of U.S. multinational corporations (MNCs). Moreover, worsened U.S.-China relations bring growing anti-U.S. sentiment in China and raise the risk of nationalistic reactions and boycotts, affecting the sales of U.S. MNC productions in China, which holds the world's largest retail market chemicals, chips, and other industries (Kapadia, 2021). This paper aims to examine the connection between the impact of the U.S.-China trade war on MNCs and how MNCs frame the U.S.-China trade war and their attitudes on U.S. trade policy towards China during the Trump era. To measure the effectiveness of the trade war on MNCs, stock price and revenue data during January 1, 2017, and January 20, 2021, are utilized. Then a quantitative content analysis based on a collection of company reports, articles, and statements related to the U.S.-China trade war and the Twitter posted by companies' official accounts is conducted to measure how U.S. MNCs frame the U.S.-China trade war.

Text Analytics Models of US news media: the case study of US-China Trade Relations

Wen Si*; The University of Texas at Dallas; US

This study aims to use unsupervised machine learning techniques such as Poisson scaling to examine attitudinal change towards trade relations between China and the U.S. over time. Poisson scaling model was developed in political science to measure ideological position of political parties (Slapin and Proksch, 2008). By using the Poisson scaling model, we identify the words frequently used in newspapers to represent their position towards China at a specific time. We collect 4,223 newspaper articles from the New York Times and 17,016 newspaper articles from the Wall Street Journal using the key search term "China Trade". The text dataset provides the corpora for studying trade-related topics not only limited to the U.S.-China trade war, which provides an advantage for exploring the general positioning of the newspaper article's attitudes towards trade relates to China. The text data can shed light on the mainstream public opinion towards trade with China during the U.S-China trade war era.

A note on Extreme value analysis of mortality at the oldest ages

Yuancheng Si*; Anhui Agricultural University and Bank of Huzhou; CN
Zezhi Tang; University of Sheffield; UK

Gbari, Poulain, Dal and Denuit [North American Actuarial Journal, 21, 2017, 397-416]fitted traditional extreme value models to oldest ages data in Belgium. We show that established extensions can provide much better fits. We illustrate this using two data sets available on-line.

Robust Bayesian Analysis of Longitudinal Data using Conditional Quantiles

Xin Tong*; University of Virginia; US

Longitudinal studies help us understand changes. Although longitudinal research has gained popularity in social and behavioral sciences, it often faces methodological challenges, such as handling nonnormal and/or missing data, small sample sizes, and population heterogeneity. In this study, I will introduce a robust Bayesian approach using conditional quantiles to address robustness and interpretability challenges in longitudinal studies. The new approach uses an asymmetric Laplace distribution to convert the problem of estimating a quantile-based model into a problem of obtaining the maximum likelihood estimator for a transformed model so that computationally powerful Bayesian methods can be applied conveniently and missing data can be flexibly addressed. The new approach has been applied to growth curve modeling. Real data examples will be provided to illustrate the application of the proposed robust Bayesian approach.

Comparison of anomaly detection methods for bot detection in online Likert-type questionnaires

Max Turgeon*; University of Manitoba; CA; US

Crowdsourcing platforms, like Amazon's Mechanical Turk, have been a cost-efficient way for researchers to administer surveys and obtain large sample sizes. For a small fee, registered users participate in the research project by filling the necessary questionnaires. However, this monetary incentive also creates an opportunity for users to answer randomly and collect the fee. This is especially problematic with Likert-type questionnaires, where each valid answer can only be an integer from 1 to d (typically d = 3 or 5). Identifying and removing these bot-generated answers is an important data processing step; failure to remove these observations can seriously bias the results of any downstream analysis.

In this work, we first discuss how the bot detection problem can be conceptualized as an anomaly detection (AD) problem. We then compare the performance of several standard AD methods (e.g. Isolation Forest, Local Outlier Factor, One-Class Support Vector Machine) for the bot detection task. We show that performance strongly depends on the ability of a method to detect **collective** anomalies, as opposed to point anomalies. We illustrate our results using freely available psychometric datasets.

Adaptive Respondent Driven Sampling of Social Networks: A Simulation based Study using Machine Learning

Josey VanOrsdale*; University of Nebraska-Lincoln; US

Respondent-driven sampling (RDS) was introduced by Dr. Douglas Heckathorn in 1997 as a means to survey the social and behavioral attributes of hidden and hard-to-reach networked populations. Within RDS, the first step is to select a few “seed” individuals from the population by using a convenience sample. After being interviewed, each seed is given a set number of “coupons” to recruit eligible peers, and this process continues recursively until the desired number of respondents have participated in the study (Salganik and Heckathorn 2004). In conducting RDS, certain norms have emerged, e.g. giving exactly three coupons to each respondent (Goel and Salganik 2010). This norm is problematic, as it often leads to a very clustered network sample, where it would be preferable to have long chains that reach mixing equilibrium deep in the network and beyond the social characteristics of the seeds (Heckathorn 1997).

The primary aim of this project is to create a variant of the RDS method that engages data science techniques during the sampling process to dynamically adjust the number of coupons given to each respondent, with the aim of producing long referral chains. Here we propose that, in practice, characteristics of the participants (e.g., race, ethnicity, gender, age, and other self-reported attributes) be used to build and continuously update predictive models of participants' propensity to generate future successful referrals. Forecasts of these predictive models can then be used to adjust the number of coupons allocated to each participant, in a manner that drives the production of long referral chains. The proposed adaptive RDS scheme is implemented as a simulation to explore its performance relative to the classical RDS technique. In this paper, I report that using an adaptive RDS is successful at producing long referral chains, and systematically describe the contextual factors that drive its relative advantage over the classical respondent driven sampling paradigm.

The Role of Personality in Trust in Public Policy Automation

Philip Waggoner*; YouGov America & Columbia University; US
Ryan Kennedy; University of Houston; US

Algorithms play an increasingly important role in public policy decision-making. Despite this consequential role, little effort has been made to evaluate the extent to which people trust algorithms in decision-making, much less the personality characteristics associated with higher levels of trust. Such evaluations inform the widespread adoption and efficacy of algorithms in public policy decision-making. We explore the role of major personality inventories – need for cognition, need to evaluate, the "Big 5" – in shaping an individual's trust in public policy algorithms, specifically dealing with criminal justice sentencing. Through an original survey experiment, we find strong correlations between all personality types and general levels of trusting automation, as expected. Further, we uncovered evidence that need for cognition increases the weight given to advice from an algorithm relative to humans, and "agreeableness" decreases the distance between respondents' expectations and advice from a judge, relative to advice from crowd.

The Big-Fish-Little-Pond Effect in Mathematics Classes across Nations and Over Years

Ze Wang*; University of Missouri; US

In educational psychology, the Big-Fish-Little-Pond Effect (BFLPE) refers to the phenomenon that students in high-achieving contexts tend to have lower self-concept than similarly abled students in a lower-achieving context due to social comparison. The theoretical explanation of the big-fish-little pond effect is the social comparison theory. Research has shown that when people evaluate themselves, they tend to compare themselves with their peers in the most immediate context they are in. Inherently, the BFLPE is a contextual effect. Extending the multilevel latent covariate model which uses implicit group mean centering, this study will examine and compare the BFLPE in 4th and 8th grade mathematics classes in many education

systems over years using data from the International Mathematics and Science Study (TIMSS). Methodological considerations include bias correction and variance estimation for complex surveys and the creation and use of plausible values to represent students' academic achievement. Examples of how to automate analysis using several R packages will be presented.

Impact of different school attendance modes on secondary school students' intention to pursue higher studies

Bernice Wong; Kolej Tuanku Ja'afar; MY
Joanne Yim*; Tunku Abdul Rahman University College; MY

Since 2020, Malaysian schools experienced closure, class rotation, and reopening at different times of the year, with different schools observing these different modes due to the severity of the Covid-19 virus infection. This study investigates the impact of these different modes of schooling among secondary school students on their perceived academic regression, test anxiety, motivation, fear of Covid-19, challenges faced, and their intention to stop school after secondary level. A total of 1207 secondary students responded to an anonymous online cross-sectional survey between January and March 2022. Data were analysed with Partial Least Squares-Structural Equation Modeling using SmartPLS software, while descriptive statistics were obtained with SPSS. Among the participants, only 23.3% attend school daily, 51.0% attend school on rotation basis, 19.1% attend school depending on the severity of Covid-19 infection rates, and 6.5% studied from home. In terms of the factors that posed difficulties to their studies, 70% reported challenges brought about by school attendance, 67.7% reported the reason of low motivation, 75.6% reported challenges in time management, 58.3% was due to fear of Covid-19. Structural Equation Modeling showed statistically significant effects between students' intention to discontinue studies after secondary level and low motivation, difficulties in school attendance, perceived academic regression, and test anxiety. Additionally, significant negative effects were found between academic engagement and intention to discontinue studies after secondary level. The study highlighted the negative impacts of switching different modes of schooling among secondary school students. The factor with the largest effect was test anxiety and academic engagement, implications of these associations will be discussed.

Stochastic Approximation Expectation-Maximization SAEM Algorithm for Fitting Differential Equation Models with Random Effects in dynr

Xiaoyue Xiong*; College of Health and Human Development, The Pennsylvania State University; US
Hui-Ju Hung; The Pennsylvania State University; US
Sy-Miin Chow; College of Health and Human Development, The Pennsylvania State University; US

A Stochastic Approximation Expectation-Maximization (SAEM) Algorithm for fitting multivariate nonlinear ordinary differential equation (ODE) models with random effects and unknown initial conditions to irregularly spaced longitudinal data was proposed by Chow, Lu, Zhu, and Sherwood (2016) by combining a Markov Chain Monte Carlo (MCMC) procedure, the

Metropolis Hasting algorithm, with the scoring algorithm to yield maximum likelihood point and standard error estimates of all time-invariant modeling parameters. Standard Expectation-Maximization (EM) procedure includes two steps in each iteration: an E-step that computes the conditional expectation of the complete-data loglikelihood function, and an M-step that updates the parameter estimates with the aim to maximize the conditional expectation from the E-step. The E-step in SAEM replaces analytic expectations with summary statistics of samples drawn from the Metropolis-Hasting procedure whereas the M-step utilizes the scoring procedure to update the parameter estimates using products from the E-step. A sequence of gain constant is used to control the degree to which new estimates are weighted at step to circumvent against settling into local minima too quickly in earlier iterations, and help speed convergence in later iterations A beta version of SAEM is now available in the R Package, Dynamic Modeling in R (dynr; Ou, Hunter, & Chow, 2019). Leveraging utility functions available from the R packages, Rcpp and RcppArmadillo, the SAEM option allows users to implement ODEs with random effects using standard model specification functions in dynr while capitalizing on computationally efficient matrix and numerical routines from Armadillo and C++. Step-by-step tutorial examples and benchmarking performance are provided in the context of three models: (1) a linear oscillator model; (2) a nonlinear extension of the Orstein-Uhlenbeck, and (3) a nonlinear Van der Pol oscillator model. The effects of the gain constants and heuristic approaches to help determine their values are demonstrated.

Comparison of Methods for Imputing Social Network Data

Ziqian Xu*; University of Notre Dame; US
Zhiyong Zhang; University of Notre Dame; US
Jiarui Hai; Tsinghua University; CN
Yutong Yang; Renmin University of China; CN

Social network data often contain missing values because of the sensitive nature of the information collected and the dependency among the network actors. As a response, network imputation methods including simple ones constructed from network structural characteristics and more complicated model-based ones have been developed. Although past studies have explored the influence of missing data on social networks and the effectiveness of imputation procedures in many missing data conditions, the current study aims to evaluate a more extensive set of eight network imputation techniques (i.e., null-tie, Reconstruction, Preferential Attachment, Constrained Random Dot Product Graph, Multiple Imputation by Bayesian Exponential Random Graph Models or BERGMs, k-Nearest Neighbors, Random Forest, and Multiple Imputation by Chained Equations) under more practical conditions through comprehensive simulation. A factorial design for missing data conditions is adopted with factors including missing data types, missing data mechanisms, and missing data proportions, which are applied to generated social networks with varying numbers of actors based on 4 different sets of coefficients in ERGMs. Results show that the effectiveness of imputation methods differs by missing data types, missing data mechanisms, the evaluation criteria used, and the complexity of the social networks. More complex methods such as the BERGMs have consistently good performances in recovering missing edges that should have been present. While simpler methods like Reconstruction work better in recovering network statistics when the missing proportion of present edges is low, the BERGMs work better when more present edges are missing. The

BERGMs also work well in recovering ERGM coefficients when the networks are complex and the missing data type is actor non-response. In conclusion, researchers analyzing social networks with incomplete data should identify the network structures of interest and the potential missing data types before selecting appropriate imputation methods.

Using Later Retrieval to Handle Missing Data in Ecological Momentary Assessments

Manshu Yang*; University of Rhode Island; US

Ecological momentary assessments (EMA) have been increasingly adopted in the past decade to help researchers understand how human behavior and experience unfold and interact in real time and in their natural context. While EMA uses intensive longitudinal measurements to maximize ecological validity, mitigate recall bias, and allow researchers to examine acute momentary factors, it also poses a major challenge -- missing responses are inevitable, often substantial, and not properly handled in data analysis. Furthermore, it is not uncommon that the missingness is driven by the unseen response itself, thereby resulting in missing not at random (MNAR) data. On the other hand, EMA brings a unique opportunity to address this challenge, by re-prompting participants shortly after they missed an EMA survey to retrieve their data. Analogous to the “double sampling” approach, such *later-retrieval* gives participants a second chance to report what they should have reported and provides *direct information* on missing responses. The current study compares four methods for handling missing data in EMA studies, including (1) maximum likelihood estimation (MLE) based on initially-observed data alone, (2) MLE based on both initially observed and later-retrieved data, (3) multiple imputation (MI) based on both initially observed and later-retrieved data, and (4) MI based on later-retrieved data alone. Monte Carlo simulations were conducted to compare the methods under different scenarios with varying sample sizes, numbers of repeated measurements, missing data patterns, as well as proportions of initially observed and later retrieved data. Preliminary results suggested that the MI method based on later-retrieved data outperformed the other methods when the later-retrieved data are a random sample of the initially unobserved data. Complete findings and implications will be discussed in the presentation.

Recent Advancements of Moderation and Mediation Analyses

Ke-Hai Yuan*; University of Notre Dame, USA; US
Hongyun Liu; Beijing Normal University, China; CN

Moderation and mediation analyses are essential methods for understanding the roles of variables in empirical research. In this talk, we review recent developments of the two methods with respect to more efficient parameter estimation and more accurate measures of effect sizes, as presented in Liu and Yuan (2021), Liu et al. (2021, in press, 2022) and Yuan et al. (2014).

Moderation occurs when the effect of the predictor on the outcome variable depends on a third variable, which is termed as the moderator. This effect is conceptually defined as the effect of the moderator on the path coefficient from the predictor to the outcome variable. The predictor and moderator are different not only in concept but also functionality. However, the effect of moderation has been implemented as an interaction effect in textbooks, software and tutorial articles. Such a treatment causes not only less efficient parameter estimates for the moderation effect itself but also wrong measures in quantifying the size of the effect. In addition, the treatment also causes problems in the study of the processes of moderated mediation and mediated moderation. The talk will cover recent methodological developments in addressing these issues by using a two-level model with single-level data. How moderated mediation is distinguished from mediated moderation will be illustrated.

Multiple effect-size measures were proposed in mediation analysis. However, these measures are either only restricted to the three-variable mediation model or lack of desirable properties. The talk will introduce a new framework for quantifying effect sizes in mediation analysis (Liu et al., 2022). Through contributions to the explained variance of the outcome variable, this framework not only facilitates easy formulation of effect-size measures for complex mediation processes but also permits more accurate moderation effects via particular paths.

References

Liu, H., & Yuan, K.-H. (2021). New measures of effect size in moderation analysis. Psychological Methods, 26(6), 680–700.
Liu, H., Yuan, K.-H., & Gan, K. (2021). Two-level mediated moderation models with single level data and new measures of effect sizes. Acta Psychologica Sinica, 53(3), 322–336.
Liu, H., Yuan, K.-H., & Li, H. (2022). A systematic framework for defining R-squaredmeasures in mediation analysis. Under review.
Liu, H., Yuan, K.-H., & Wen. Z. (in press). Two-level moderated mediation models withsingle-level data and new measures of effect sizes. Behavior Research Methods.
Yuan, K.-H., Cheng, Y., & Maxwell, S. (2014). Moderation analysis using a two-level regression model. Psychometrika, 79(4), 701–732.

Meta-Analysis of Correlation Coefficients: A Cautionary Tale on Treating Measurement Error

Qian Zhang*; Florida State University; US

A scale to measure a psychological construct is subject to measurement error. When meta-analyzing correlations obtained from scale scores, many researchers recommend correcting for measurement error. I considered three caveats when correcting for measurement error in meta-analysis of correlations: (1) the distribution of true scores can be non-normal, resulting in violation of the normality assumption for raw correlations and Fisher's z transformed correlations; (2) coefficient alpha is often used as the reliability, but correlations corrected for measurement error using alpha can be inaccurate when some assumptions of alpha (e.g., tau-equivalence) are violated; and (3) item scores are often ordinal, making the disattenuation formula potentially problematic. Via three simulation studies, I examined the performance of two meta-analysis approaches – with raw correlations and z scores. In terms of estimation accuracy and coverage probability of the mean correlation, results showed that (1) considering the true-score distribution alone, estimation of the mean correlation was slightly worse when true scores of the constructs were skewed rather than normal; (2) when the tau-equivalence assumption was violated and coefficient alpha was used for correcting measurement error, the mean correlation estimates can be biased and coverage probabilities can be low; and (3) discretization of continuous items can result in biased estimates and under-coverage of the mean correlations even when tau-equivalence was satisfied. With more categories and/or items on a scale, results can improve whether tau-equivalence was met or not. Based on these findings, I gave recommendations for conducting meta-analyses of correlations.

Social Network Analysis in the Framework of Structural Equation Modeling

Zhiyong Zhang*; University of Notre Dame Notre Dame, IN 46556 USA; US

Social network data are increasingly collected in many fields of research, business, and government. For example, to study student behaviors, it is important to understand the context of behaviors because students are not independent entities but are typically connected with one another, which naturally leads to the collection and analysis of network data. This study proposes to combine structural equation modeling (SEM) techniques and data science methods to model network data. It tackles the complex problems of network data by treating them as new types of variables in SEM. We will show how to construct a SEM with networks and how to estimate such a model.