Conference on Current Trends in Survey Statistics 2019

Institute for Mathematical Sciences

National University of Singapore, Singapore

13 - 16 August, 2019

A Satelite Conference to the 62nd ISI World Statistics Congress, to be held in Kuala Lumpur from 18 to 23 August 2019.

Scientific Advisory Board

Partha LAHIRI (Chair), University of Maryland, College Park
Raymond CHAMBERS, University of Wollongong
Malay GHOSH, University of Florida
Graham KALTON, Westat
Danny PFEFFERMANN, National Statistician of Israel and University of Southampton
J. N. K. RAO, Carleton University
Pedro Luis do Nascimento SILVA, IBGE, Brazil

Scientific Program Committee

Sanjay CHAUDHURI (Chair), National University of Singapore, Singapore
William BELL US Census Bureau, USA
Snigdhansu CHATTERJEE University of Minnesota, USA
Yang CHENG, US Census Bureau, USA
Cinzia CIRILLO, University of Maryland, College Park, USA
Gauri DATTA, University of Georgia, USA
Somnath DATTA, University of Florida, USA
Jörg DRECHSLER, Institute for Employment research, Germany
Arul EARNEST, Monash University, Australia
Michael ELLIOT, University of Michigan, USA
Scott HOLAN, University of Missouri, USA
Jiming JIANG, University of California, Davis, USA
Jae-Kwang KIM, Iowa State University, USA
Yan LI, University of Maryland, College Park, USA
Ralph MUNNICH, University of Trier, Germany
Santanu PRAMANIK, Delhi Centre of National Data Innovation, India
Rebecca STEORT, Duke University, USA
Dongchu SUN, University of Missouri, USA
Jiraphan SUNTORNCHOST, Chulalongkorn University, Thailand
Michael SVERCHKOV, Bureau of Labour Statistics, USA
Nikos TZAVIDIS, University of Southampton, UK
Li-Chun ZHANG, University of Southampton, Oslo, Stat Norway

Local Organizing Committee

Sanjay CHAUDHURI (Chair), National University of Singapore, Singapore
Emily Ee Cheng CHAN, Institute for Mathematical Sciences, National University of Singapore
Eileen TAN, Institute for Mathematical Sciences, National University of Singapore
Jia Ling LEE, Institute for Mathematical Sciences, National University of Singapore

IT Manager

Stephen San San AUYONG Institute for Mathematical Sciences, National University of Singapore

Sponsorships and Endorsements:

The Institute for Mathematical Science, National University of Singapore
The International Chinese Statistical Association
The International Association for Survey Statisticians.

Contents

Glossary

IS	Invited Session
CS	Contributed Session

Programme Summary

Program

Tuesday August 13

Plenary Lecture 1 *Venue:* IMS Auditorium

Chair : Sanjay CHAUDHURI, National University of Singapore

09:00 **Need for Data Integration and Statistical Modeling for Various Purposes** [**Abstract 44**]

Partha LAHIRI, *University of Maryland College Park*

Coffee Break 10:00 - 10:30 Lounge

IS 1 New Advances in Mixed Effects Models and Their Applications *Venue:* IMS Auditorium

Chair and Organizer : Jiming JIANG, University of California Davis

10:30 **Disparity subtyping: bringing precision medicine closer to disparity science** [**Abstract 55**]

J. Sunil RAO, *University of Miami*
Huilin YU, *University of Miami*
Jean-Eudes DAZARD, *Case Western Reserve University*

11:00 **Sumca: Simple, Unified, Monte-Carlo Assisted Approach to Second-order Unbiased MSPE Estimation** [**Abstract 33**]

Mahmoud TORABI, *University of Manitoba, Canada*

11:30 **Partially Observed Mixed Effects Model with Application to Missing Data Problems** [**Abstract 57**]

Thuan NGUYEN, *Oregon Health & Science University, USA*

IS 2 Data Integration in Survey Sampling *Venue:* Discussion Room

Chair : Takumi SAEGUSA, University of Maryland College Park

Organizer : Michael ELLIOTT, University of Michigan

10:30 **Total Error Frameworks for Found Data** [**Abstract 45**]

Paul BIEMER, *Distinguished Fellow, Statistics, RTI International*

11:00 **Statistical Data Integration and Inference via Multilevel Regression and Poststratification** [**Abstract 59**]

Yajuan SI, *University of Michigan*

11:30 **Using MRP to combine probability and non-probability sampling frames** [**Abstract 31**]

Lauren KENNEDY, *School of Social Work, Columbia University*

Lunch 12:00 - 14:00 Lounge

IS 3 Advancements in Entity Resolution

Venue: IMS Auditorium

Chair and Organizer : Rebecca C. STEORTS, Duke University

14:00 Distributed Markov chain Monte Carlo for Scalable Bayesian Entity Resolution [Abstract 42]

Neil MARCHANT, *School of Computing and Information Systems, University of Melbourne*

14:30 Simultaneous Bipartite Record Linkage and Linear Regression [Abstract 25]

Jiurui TANG, *Duke Statistical Science Department*

15:00 Distributed Entity Resolution for Computational Social Science [Abstract 49]

Rebecca C. STEORTS, *Duke University*

IS 4 Big Data, Record Linkage and Modelling in Official Statistics

Venue: Discussion Room

Chair and Organizer : Nikos TZAVIDIS, University of Southampton

14:00 Use of Record Linkage to Improve Statistics: methods and issues on using linked data [Abstract 1]

Andrea DINIZ DA SILVA, *Brazilian Institute of Geography and Statistics and National School of Statistical Sciences*

14:30 Big data in official statistics [Abstract 21]

Jan VAN DEN BRAKEL, *Statistics Netherlands and Maastricht University*

15:00 Regularized Area-level Modelling for Robust Small Area Estimation in the Presence of Unknown Covariate Measurement Errors [Abstract 23]

Jan Pablo BURGARD, *Trier University*
Joscha KRAUSE,
Dennis KREBER,

Coffee Break 15:30 - 16:00 Lounge

IS 5 Recent Developments in Small Area Estimation

Venue: IMS Auditorium

Chair : Michael SVERCHKOV, Bureau of Labor Statistics
Organizer : Sanjay CHAUDHURI, National University of Singapore

16:00 A General functional form of Mean Squared Error Estimator in Small Area Estimation [Abstract 37]

Masayo HIROSE, *The Institute of Statistical Mathematics*

16:30 Environmental bioassays and ecological risk assessments [Abstract 10]

Debashis MONDAL, *Oregon State University*

17:00 MULTIVARIATE SMALL AREA ESTIMATION UNDER NOT MISSING AT RANDOM NONRESPONSE [Abstract 39]

Michael SVERCHKOV, *Bureau of Labor Statistics*
Danny PFEFFERMANN, *National Statistician and CBS Director, Israel*

CS 1 Application of Data Integration Techniques in Real Life

Venue: Discussion Room

Chair : Indranil MUKHOPADHYAY, Indian Statistical Institute Calcutta

16:00 **Survey on Importance of Biostatistics in Medical Education and Research: Challenges and Experiences [Abstract 47]**

Pramod K. GUPTA, *PGIMER-Chandigarh*

16:20 **Latest development in the combined Census approach at Statistics Canada [Abstract 40]**

Michelle SIMARD, *Statistics Canada*

16:40 **Small Area Estimates for the Brazilian Annual Service Sector Survey [Abstract 12]**

Denise SILVA, *National School of Statistical Sciences*
André NEVES, *National School of Statistical Sciences*
Fernando MOURA, *Federal University of Rio de Janeiro*

Welcome Reception 17:30 - 19:00 Lounge

Wednesday August 14

Plenary Lecture 2

Venue: IMS Auditorium

Chair : Malay GHOSH, University of Florida

09:00 **Old and New Challenges in the Production of Official Statistics With Different Methods of Data Collection** [Abstract 8]

Danny PFEFFERMANN, *National Statistician and CBS Director, Israel*

Coffee Break 10:00 - 10:30 Lounge

IS 6 Recent Developments in Design-based Inference

Venue: IMS Auditorium

Chair : Danny PFEFFERMANN, National Statistician and CBS Director, Israel

Organizer : Li-Chun ZHANG, University of Southampton

10:30 **Exponential inequalities for sampling designs** [Abstract 16]

Guillaume CHAUVET, *Ensai - Irmar*

11:00 **Design-based indirect small area estimation** [Abstract 34]

Maria Giovanna RANALLI, *University of Perugia*

11:30 **Some new methods of graph sampling** [Abstract 32]

Li-Chun ZHANG, *University of Southampton*

IS 7 Spatial Methodology for Small Area Estimation and Official Statistics

Venue: Discussion Room

Chair : Gauri Sankar DATTA, U.S. Census Bureau and University of Georgia

Organizer : Scott HOLAN, University of Missouri

10:30 **Generating partially synthetic geocoded public use data with decreased disclosure risk using differential smoothing** [Abstract 17]

Harrison QUICK, *Drexel University*

11:00 **The Conjugate Multivariate Distribution with Application to Spatial Federal Datasets** [Abstract 27]

Jonathan BRADLEY, *Florida State University*

11:30 **A Multivariate Spatial Mixture Mixed Effects Model with Application to the American Community Survey** [Abstract 52]

Scott HOLAN, *University of Missouri*

Lunch 12:00 - 14:00 Lounge

IS 8 Survey Sampling in Transportaion

Venue: IMS Auditorium

Chair and Organizer : Cinzia CIRILLO, University of Maryland College Park

14:00 **Leveraging Social Capital Questions for Travel and Activity Surveys** [**Abstract 38**]

Michael MANESS, *University of South Florida*

14:30 **Collecting Stated Choice data on the adoption of Electric Vehicles with consideration of Social Conformity** [**Abstract 6**]

Cinzia CIRILLO, *University of Maryland College Park*
Javier BAS, *University of Maryland*
Elisabetta CHERCHI, *Newcastle University*

IS 9 Recent Topics on Data Integration

Venue: Discussion Room

Chair and Organizer : Jae-kwang KIM, Iowa State University

14:00 **Combining aggregate statistics from big data and probability survey: application to electronic health records and national health surveys** [**Abstract 50**]

Ryung S KIM, *Albert Einstein College of Medicine*

14:30 **Kernel-based Functional Calibration Estimation in Survey Sampling** [**Abstract 63**]

Zhengyuan ZHU, *Iowa State University*

15:00 **Mass imputation for combining non-probability and probability survey samples** [**Abstract 20**]

Jae-kwang KIM, *Iowa State University*

Coffee Break 15:30 - 16:00 Lounge

Special Invited Lecture

Venue: IMS Auditorium

Chair : Partha LAHIRI, University of Maryland College Park

16:00 **Data integration: the good, the bad and the ugly** [**Abstract 46**]

Pedro Luis do Nascimento SILVA, *IBGE - National School of Statistical Sciences, Brazil*

Thursday August 15

Plenary Lecture 3

Venue: `IMS Auditorium`

Chair : Debashis MONDAL, Oregon State University

09:00 **Statistical Methods To Survey Hidden Networked Populations** [Abstract 35]

Mark HANDCOCK, *University of California Los Angeles*

Coffee Break 10:00 - 10:30 Lounge

IS 10 Model Selection and Uncertainty Estimation for Area-Level Small Area Models

Venue: `IMS Auditorium`

Chair : Scott HOLAN, University of Missouri

Organizer : Gauri Sankar DATTA, U.S. Census Bureau and University of Georgia

10:30 **Small Area Estimation under Non-normal Random Effects** [Abstract 13]

Gauri Sankar DATTA, *U.S. Census Bureau and University of Georgia*
Kyle IRIMATA,
Jerry MAPLES,
Eric SLUD,

11:00 **Observed best selective prediction in small area estimation** [Abstract 62]

Yuki KAWAKUBO, *Chiba University*

11:30 **Estimation and Inference for Area-wise Spatial Income Distributions from Grouped Data** [Abstract 14]

Genya KOBAYASHI, *Chiba University*

IS 11 Inferences for Complex Statistical Problems

Venue: `Discussion Room`

Chair : Subharup GUHA, University of Florida

Organizer : Somnath DATTA, University of Florida

10:30 **Are We There Yet: Differential analysis of Single-cell RNA Sequencing Data?** [Abstract 56]

Susmita DATTA, *University of Florida*

11:00 **Inference for Generalized Linear Models with A Diverging Number of Covariates: A Refined De-Biased Approach** [Abstract 61]

Yi LI, *University of Michigan*

11:30 **Symmetric Geometric Skew Normal Regression Model** [Abstract 9]

Debashis KUNDU, *Indian Institute of Technology Kanpur*
Deepak PRAJAPATI, *Indian Institute of Technology Kanpur*

Lunch 12:00 - 14:00 Lounge

IS 12 Bayesian Approach for Poverty mapping and M-quantile Approach for Small Area *Venue:* `IMS Auditorium`
Chair : Masayo HIROSE, The Institute of Statistical Mathematics
Organizer : Jiraphan SUNTORNCHOST, Chulalongkorn University

14:00 Outlier robust area-level estimation using the M-quantile approach to small area [Abstract 53]

Stefano MARCHETTI, *University of Pisa*

14:30 An Approximate Bayesian Approach for Solving Different Inferential Problems in Poverty Research for Small Areas [Abstract 24]

Jiraphan SUNTORNCHOST, *Chulalongkorn University*
Partha LAHIRI, *University of Maryland College Park*

15:00 A Hierarchical Bayesian Approach for Addressing Multiple Objectives in Poverty Research for Small Areas [Abstract 41]

Monica PRATESI, *Department of Economics and Management, University of Pisa*
Partha LAHIRI, *University of Maryland College Park*
Gaia BERTARELLI, *University of Pisa*
Stefano MARCHETTI, *University of Pisa*
Nicola SALVATI, *University of Pisa*

IS 13 Data Integration in Bio and Health Statistics *Venue:* `Discussion Room`
Chair : Yan LI, University of Maryland College Park
Organizer : Sanjay CHAUDHURI, National University of Singapore

14:00 Bayesian Nonparametric Differential Analysis for Dependent Multigroup Data with Application to DNA Methylation Analyses [Abstract 54]

Subharup GUHA, *University of Florida*

14:30 Latent Multivariate Log-Gamma Models for High-Dimensional Multi-Type Responses with Application to Daily Fine Particulate Matter and Mortality Counts [Abstract 11]

DEBAJYOTI SINHA, *FLORIDA STATE UNIVERSITY*

15:00 A Kernel Weighting Approach to Improve Population Representativeness for Estimating Prevalence of Risk-factors and Diseases [Abstract 60]

Yan LI, *University of Maryland College Park*

Coffee Break 15:30 - 16:00 Lounge

CS 2 Likelihood Based Methods for Complex Survey Data *Venue:* `IMS Auditorium`
Chair : Andrea DINIZ DA SILVA, Brazilian Institute of Geography and Statistics and National School of Statistical Sciences

16:00 Bayesian Methods in the Governments Surveys Estimations of the U.S. Census Bureau [Abstract 3]

Bac TRAN, *U.S. Census Bureau*

16:20 **Simultaneous Inference for Empirical Best Predictor Under Generalized Linear Mixed Models** [**Abstract** 29]

Katarzyna RELUGA, *University of Geneva*

16:40 **A comparison of likelihood-based methods for size-biased sampling** [**Abstract** 58]

Victoria LEAVER, *University of Wollongong*

17:00 **Simulation Study to Compare AFT Model Parameter Estimations Based on Pure Model-Based Approach and Model Assisted Approach** [**Abstract** 18]

Imali JAYAMANNE, *Department of Statistics, University of Colombo, Sri lanka*
Asoka RAMANAYAKE, *Department of Statistics, University of Colombo, Sri lanka*

CS 3 Recent Developments in Data Integration and Prediction Techniques *Venue:* Discussion Room

Chair : Sanghamitra PAL, West Bengal State University

16:00 **Data integration using fractional hot deck imputation** [**Abstract** 28]

Jongho IM, *Yonsei University*
Inho PARK,

16:20 **Evaluating the effectiveness of using livestock guarding dogs as a method of conflict resolution** [**Abstract** 19]

Ivor CRIBBEN, *University of Alberta*

16:40 **An area model to estimate multidimensional poverty index in the municipalities with the worst social conditions in Colombia.** [**Abstract** 43]

Nicolas RAMIREZ, *Student*

17:00 **Predicting Panel Nonresponse with Machine Learning in a Longitudinal Framework** [**Abstract** 5]

Christoph KERN, *University of Mannheim*

Conference Dinner 18:30 - 21:30 NUS Guild House

Friday August 16

IS 14 Micro-Simulation Techniques *Venue:* IMS Auditorium

Chair and Organizer : Ralf MÜNNICH, Trier University

09:00 Micro-simulation in Official Statistics [Abstract 36]

Markus ZWICK, *Federal Statistical Office of Germany*

09:30 Analyzing and visualizing sources of uncertainly in dynamic microsimulations [Abstract 22]

Jan Pablo BURGARD, *Trier University*
Simon SCHMAUS,

10:00 Microsimulations between statistics and optimization [Abstract 48]

Ralf MÜNNICH, *Trier University*

IS 15 Data Integration Techniques in Epidemiology *Venue:* Discussion Room

Chair and Organizer : Arul EARNEST, Monash University

09:00 Exposure Assessment in Environmental Epidemiology [Abstract 15]

Geoff MORGAN, *University of Sydney*

09:30 Geo-spatial modeling of cancer registry data linked to administrative data [Abstract 2]

Arul EARNEST, *Monash University*

Coffee Break 10:30 - 11:00 Lounge

IS 16 Synthetic Data *Venue:* IMS Auditorium

Chair and Organizer : Jörg DRECHSLER, Institute for Employment Research

11:00 Validating outcomes from synthetic data: results from a multi-year implementation [Abstract 30]

Lars VILHUBER, *Cornell University*

11:30 MLDS Synthetic Data Project: An Evaulation [Abstract 7]

Daniel BONNERY, *University of Maryland College Park*

12:00 Synthetic Microdata for Establishment Surveys Under Informative Sampling [Abstract 26]

Jörg DRECHSLER, *Institute for Employment Research*
Hang J. KIM, *University of Cincinnati*
Katherine J. THOMPSON, *U. S. Census Bureau*

IS 17 Recent Advances in the Analysis of Complex Surveys *Venue:* Discussion Room

Chair : Balgobin NANDRAM, Professor of Statistics, Worcester Polytechnic Institute

Organizer : Snigdhansu CHATTERJEE, University of Minnesota

11:00 **A prediction approach in Adaptive Sampling** [**Abstract** 51]

Sanghamitra PAL, *West Bengal State University*

11:30 **Bayesian Nonparametric Models for Two-Stage Sample Surveys** [**Abstract** 4]

Balgobin NANDRAM, *Professor of Statistics, Worcester Polytechnic Institute*

Lunch 12:30 - 14:00 Lounge

Abstracts

1. Use of Record Linkage to Improve Statistics: methods and issues on using linked data

[IS 4, (page 4)]

Andrea DINIZ DA SILVA, *Brazilian Institute of Geography and Statistics and National School of Statistical Sciences*

Important indicators of living conditions of the population, such as gross domestic product, human development index and sustainable development indicators, still depend on improving statistics to be properly monitored. Traditional ways of producing statistics, such as census and surveys, are effective but costly for producers, users and respondents. The cost includes operation budget, loss of quality when operation size increases and respondent burden. Integrating data from already existing data sources can allow cost reduction. However, it depends on availability of data sources, setting a method up and making correspondent adjustments when using linked data to avoid potentially misleading analyses. Record Linkage methods are a key component of data integration so to contribute to the discussion results of an empirical study to identify the one providing comparatively better results when linking records of agricultural producers will be presented. The study includes eight methods, one based on the Fellegi-Sunter decision model, two based on clustering techniques using the K-Means algorithm, four using classification trees and one using the Support Vector Machine (SVM) algorithm. In addition, discussion on the use of linked data, including reference to solutions that correct for the bias caused by linkage errors, which have been proposed but not yet embedded in the various analysis procedures in common use, will be presented.

2. Geo-spatial modeling of cancer registry data linked to administrative data

[IS 15, (page 11)]

Arul EARNEST, *Monash University*

Geo-spatial modeling of cancer registry data linked to administrative data

3. Bayesian Methods in the Governments Surveys Estimations of the U.S. Census Bureau

[CS 2, (page 9)]

Bac TRAN, *U.S. Census Bureau*

The data on U.S. state and local government finances (ALFIN) and employment (ASPEP) are served as one of the primary sources for computations of critically important economic time series. The Bureau of Economic Analysis (BEA) uses the government statistics data as input to the gross domestic product (GDP) and other components of the national income and product accounts (NIPAs), as well as to the regional economic accounts. In this research we present the designs and estimation methodologies for those two surveys: ASPEP and ALFIN. ASPEP provides statistics on the number of federal, state, local government civilian employees and their gross payrolls. ALFIN provides statistics on revenue, expenditure, debt, and assets (cash and security holdings) for governments. There are statistics for the 50 state areas and the District of Columbia, as well as a national summary. The two surveys are designed to produce reliable estimates, for example, the number of full-time and part-time employees and payroll at the national level for large domains. However, it is also required to estimate the parameters for individual function codes (activities) within each state. This requirement prompted us (a) First, to develop estimation methods; especially, small area estimation to deal with sparse data due to the requirements for the estimates in small cells (b) Secondly, to develop a hierarchical Bayesian methodology that uses a unit level model in borrowing strength from previous census data as an alternative to collecting expensive additional data for small cells (c) Treatments for outliers. We evaluate design-based properties of the proposed hierarchical Bayesian methodology using a Monte Carlo simulation study. An important feature of this paper is evaluation of design-based properties of the proposed hierarchical Bayesian methodology through a Monte Carlo simulation study using a real finite population.

4. Bayesian Nonparametric Models for Two-Stage Sample Surveys

[IS 17, (page 12)]

Balgobin NANDRAM, *Professor of Statistics, Worcester Polytechnic Institute*

Typically survey data have continuous responses with outliers, gaps and ties, and the distributions of the responses might be skewed. Therefore, predictive inference using a two-stage hierarchical Bayesian model with normality at both stages (responses and random effects) might not be robust against these features. So we provide a two-level non-parametric

Bayesian model with a Dirichlet process at each stage with a normal baseline, thereby permitting a more robust predictive inference. This is a different hierarchical Dirichlet process model, particularly constructed for survey data from small areas, than the one available in the literature. We show how to fit the four versions of this model (e.g., one version has Dirichlet processes at both levels and another version has no Dirichlet process at any of the two stages) using Markov chain Monte Carlo methods. We also show how to assess model fit using Bayes factors, and point out a not so well-known difficulty using Bayes factors, thereby leading to alternative measures. An application on body mass index obtained from NHANES III and a simulation study are discussed to compare the four models using many different measures, which show that the full Dirichlet process model (Dirichlet processes at both levels) might be preferred. We also show how to extend our procedure to multi-stage surveys, and to include covariates, survey weights and other baseline distributions at different stages.

5. Predicting Panel Nonresponse with Machine Learning in a Longitudinal Framework

[CS 3, (page 10)]

Christoph KERN, *University of Mannheim*

Nonresponse in panel studies can lead to a substantial loss in data quality due to its potential to introduce bias and distort survey estimates. Recent work investigates the usage of machine learning methods to predict nonresponse in advance, such that predicted nonresponse propensities can be used to inform the data collection process. However, predicting nonresponse in panel studies requires accounting for the longitudinal data structure in terms of model building, tuning, and evaluation. This study proposes a new framework for predicting nonresponse with machine learning and multiple panel waves and illustrates its application. With respect to model building, this approach utilizes information from multiple panel waves by introducing features that aggregate previous (non)response patterns. Concerning model tuning and evaluation, temporal cross-validation is employed by iterating through pairs of panel waves such that the train and test set moves in time. This approach is exemplified with data from a German probability-based mixed-mode panel (GESIS Panel), with which multiple machine learning models were trained and tested for 20 panel waves, respectively. Particularly random forests and extremely randomized trees in combination with features that aggregate information over multiple previous waves resulted in competitive prediction performance over all hold-out sets.

6. Collecting Stated Choice data on the adoption of Electric Vehicles with consideration of Social Conformity

[IS 8, (page 7)]

Cinzia CIRILLO, *University of Maryland College Park*
Javier BAS, *University of Maryland*
Elisabetta CHERCHI, *Newcastle University*

Stated Preference (SP) Surveys collect data about future behavior often in terms of consumer choices. This technique is extensively used in economics, marketing, transportation and social sciences. In this talk we present a SP survey that was design to measure the effect of Social Conformity (SC) in the preference for electric vehicles (EV). To do so, we have designed an experiment that takes the classical approach a step further. SC is a type of social influence that involves changes in attitudes, beliefs, and behaviors, in order to fit in a group. Differently from most of the literature in the field, we include the variables controlling for the SC, randomized, along with the attributes of the alternatives (price, range, propulsion cost, etc.). This allows for a direct comparison of their effects. Three different measures of SC are considered and a different methodology is developed for each of them. The first measure pertains to what is known as Descriptive norms, which define when the behavior of an individual is affected by what other people do. We approached this imitation of others behavior by incorporating, as an attribute in the choice task, the number of EV sold in the last month. This figure provides information about how many individuals made the choice of purchasing an EV without linking it to a reference group. The second measure used to characterize the influence of SC is the Informational conformity. Informational conformity occurs when the person lacks knowledge about the alternative, or the consequences associated to the alternative, and she turns to the members of her group for guidance. In order to measure it, we included a sentence providing feedback, positive or negative, about one of two aspects of the EV; the need to charge the vehicle frequently, and the need to change ones activities due to range shortage. The survey has been designed following a D-optimum criterium, with 3 segments. Bayesian priors were used for the level of the attributes (2, 3, or

4 levels, depending on the attribute) based on similar studies, and then adjusted with the results of the pilot. The experiment was customized for 3 different car classes. Therefore, the final design included 9 utility functions; 3 alternatives for each of the 3 classes. 24 choices tasks for each segment were created, divided into 4 different blocks, so each respondent was presented a set of 6 choice tasks, randomly assigned, too.

The preliminary data exploration suggests an important role played by the SC elements. In a significant number of cases the interviewed changes her mind about the decision made. Results derived from discrete choice models will be presented and conclusions based on the empirical case study drawn about the role of SC on the adoption of EV.

7. MLDS Synthetic Data Project: An Evaulation

[IS 16, (page 11)]

Daniel BONNERY, *University of Maryland College Park*

For the last five years, the Maryland Longitudinal Data System (MLDS) has been serving as a central repository of highly confidential student and workforce data. The Institute of Educational Sciences funded a project to create, evaluate, and potentially release Synthetic versions of the MLDS data. We will present an evaluation of the synthetic data in terms of research validity and disclosure risk. The evaluation of research validity of the synthetic data will include efforts to assess the general utility (e.g., comparisons of variable distributions) and specific utility (e.g., comparisons of parameter estimates from analyses on real and synthetic data). We will also present another vital step in our synthetic data project: the assessment of disclosure risk of the synthetic data. That assessment is required to comply with laws governing the confidentiality of state held data and it will also be a necessary step in seeking permission from MLDS Governing Board to release the synthetic data. We will present our progress in the disclosure risk assessment process. Finally, we will discuss the benefits of the synthetic data for researchers who do not have access to the real data.

8. Old and New Challenges in the Production of Official Statistics With Different Methods of Data Collection

[Plenary Lecture 2, (page 6)]

Danny PFEFFERMANN, *National Statistician and CBS Director, Israel*

In this presentation I shall discuss three main methods of data collection for official statistics: traditional surveys based on probability sampling, use of administrative files, and the potential use of big data. I shall discuss the advantages and challenges of each method and propose possible solutions to some of the problems that I shall raise. Integration of data obtained from different sources will also be considered. Several numerical illustrations will be shown.

9. Symmetric Geometric Skew Normal Regression Model

[IS 11, (page 8)]

Debashis KUNDU, *Indian Institute of Technology Kanpur*

Deepak PRAJAPATI, *Indian Institute of Technology Kanpur*

Recently, Kundu (2014, Sankhya, Ser. B, 167 {189, 2014}) proposed a geometric skew normal distribution as an alternative to the Azzalini's skew normal distribution. The geometric skew normal distribution can be a skewed distribution, it can be heavy tailed as well as multimodal also, unlike Azzalini's skew normal distribution. In this paper we have proposed a symmetric geometric skew normal distribution as an alternative to the normal distribution and log Birnbaum-Saunders distribution. It is a very flexible class of distributions, of which normal distribution is a special case, and it can be heavy tailed also. The proposed model has three unknown parameters, and it is observed that the maximum likelihood estimators of the unknown parameters cannot be obtained in explicit forms. In this paper we have proposed a very efficient expecta- tion maximization (EM) algorithm, and it is observed that the proposed EM algorithm works very well. We have further considered a location shift symmetric geometric skew normal regression model. It is a more flexible than the standard Gaussian regression model. The maximum likelihood estimators of the unknown parameters are obtained based on expectation maximization algorithm. Extensive simulation experiments and the analyses of two data sets have been presented to show the effectiveness of the proposed model and the estimation techniques.

Key Words and Phrases: Absolute continuous distribution; singular distribution; Fisher information matrix; EM algorithm; joint probability distri-

bution function; joint probability density function.

AMS Subject Classifications: 62F10, 62F03, 62H12.

10. Environmental bioassays and ecological risk assessments

[IS 5, (page 4)]

Debashis MONDAL, *Oregon State University*

In recent years, we have witnessed a growing research focus on environmental toxicity sparked by concerns related to watershed protection in lakes, rivers, estuaries and coastal areas. Environmental bioassays are now routinely conducted by EPA and other regulatory bodies as part of numerous broad-range surveys on conditions of ecological resources in order to determine adverse biological effects of contaminants present in soil, water and sediments across a geographic region under study. However, till date, statistical analyses of these survey data and environmental bioassays have been limited. For example, the result of Ampeisca abdita toxicity test is often summarized by percentage of ratios of survivals in cases and controls, and toxicity is determined on the basis of whether or nor this percentage is below eighty. In this talk, I use NY-NJ harbor area as a case study to present some critical statistical evaluations of these bioassays. Furthermore, I show how spatial models allow us to map risk, predict levels of chemical contaminants and toxicity across the watershed, and identify statistical associations among particular chemical contaminants and degraded ecological community or toxic sediments. Part of the work is done in collaborations with Xiaohui Chang at Oregon State University.

11 . Latent Multivariate Log-Gamma Models for High-Dimensional Multi-Type Responses with Application to Daily Fine Particulate Matter and Mortality Counts

[IS 13, (page 9)]

DEBAJYOTI SINHA, *FLORIDA STATE UNIVERSITY*

Tracking and estimating Daily Fine Particulate Matter (PM2.5) is very important as it has been shown that PM2.5 is directly related to mortality related to lungs, cardiovascular system, and stroke. That is, high values of PM2.5 constitute a public health problem in the US, and it is important that we precisely estimate PM2.5 to aid in public policy decisions. Thus, we propose a Bayesian hierarchical model for high-dimensional ”multi-type” responses. By ”multi-type” responses we mean a collection of correlated responses that have different distributional assumptions (e.g., continuous skewed observations, and discrete counts observations). The Centers for Disease Control and Prevention (CDC) database provides mortality counts related to PM2.5 and daily averaged PM2.5 which are both treated as responses in our analysis. Our model capitalizes on the shared conjugate structure between the Weibull (to model PM2.5), Poisson (to model disease mortality), and multivariate log-gamma distributions, and we use dimension reduction to aid with computation. Our model can also be used to improve the precision of estimates and predict values at undisclosed/missing counties. We provide a simulation study to illustrate the performance of the model, and give an in-depth analysis of the data from CDC.

12. Small Area Estimates for the Brazilian Annual Service Sector Survey

[CS 1, (page 5)]

Denise SILVA, *National School of Statistical Sciences*
André NEVES, *National School of Statistical Sciences*
Fernando MOURA, *Federal University of Rio de Janeiro*

The Brazilian Institute of Geography and Statistics (IBGE) conducts an Annual Service Sector Survey (PAS) whose scope comprises segments of the tertiary sector in the whole country. However, sample estimates for some economic activities in the north, north-east and midwest regions of Brazil do not hold an adequate level of precision. Furthermore, one of the main variables of interest is considerably skewed with potential outliers. To surmount these problems, we applied skew normal and skew-t distributions to cross-sectional small domain estimation models that relate operating revenue variables to potential auxiliary variables (the number of employed people and wages) from a business register to obtain consistent and accurate estimates for specified domains (MOURA et al, 2017). In addition, models incorporating times series data of this repeated survey have also been tested. Preliminary results indicate that change over time patterns vary by domain and are related to auxiliary variables, therefore models that borrow information over time lead to estimates with better precision for most of domains .

13. Small Area Estimation under Non-normal Random Effects

[IS 10, (page 8)]

Gauri Sankar DATTA, *U.S. Census Bureau and University of Georgia*
Kyle IRIMATA,
Jerry MAPLES,
Eric SLUD,

The Fay-Herriot model is a popular linear mixed model which has been extensively used in small area estimation. This model treats the population mean of a small area or a domain as a sum of a fixed linear regression and a random effect. The empirical best linear unbiased predictor (EBLUP) of a small area mean is used as an alternative to the direct estimate. The EBLUP and its measure of uncertainty, usually the mean squared error (MSE), are generally obtained under the normality assumption for the random effects. Even under normality exact expressions of the MSEs are typically intractable, so their analytic approximations are obtained accurately up to the order of $o(m^{-1})$, when m, the number of small areas, is large. In this talk, we relax the normality assumption for the random effects. We estimate the model parameters by estimating equations and use asymptotic expressions to approximate the MSEs of the EBLUPs. A resampling-based nonparametric bootstrap method is used to estimate the MSEs accurately to the order $o(m^{-1})$. We investigate large sample and finite sample properties of our bootstrap estimators via asymptotic analysis and simulation study. We conducted a simulation study to compare the bootstrap MSE estimators with analytically derived second-order accurate estimators proposed in the literature. Both the methods perform satisfactorily when the sampling error variances do not display substantial heterogeneity and they are similar in magnitude to the model error variance. However, in practical applications, for example in some applications within the SAIPE program of the U.S. Census Bureau, the sampling error variances are usually quite heterogeneous and are larger than the model error variance by an order of magnitude. Such scenarios benefit more from small area estimation. Our simulations for the latter realistic scenarios show that the performance of the analytic MSE estimators deteriorates significantly compared to the bootstrap estimators.

14. Estimation and Inference for Area-wise Spatial Income Distributions from Grouped Data

[IS 10, (page 8)]

Genya KOBAYASHI, *Chiba University*

Estimating income distributions plays an important role in the measurement of inequality and poverty over space. The existing literature on income distributions predominantly focuses on estimating an income distribution for a country or a region separately and the simultaneous estimation of multiple income distributions has not been discussed in spite of its practical importance. In this work, we develop an effective method for the simultaneous estimation and inference for area-wise spatial income distributions taking account of geographical information from grouped data. Based on the multinomial likelihood function for grouped data, we propose a spatial state-space model for area-wise parameters of parametric income distributions. We provide an efficient Bayesian approach to estimation and inference for area-wise latent parameters, which enables us to compute area-wise summary measures of income distributions such as mean incomes and Gini indices, not only for sampled areas but also for areas without any samples thanks to the latent spatial state-space structure. The proposed method is demonstrated using the Japanese municipality-wise grouped income data. The simulation studies show the superiority of the proposed method to a crude conventional approach which estimates the income distributions separately.

15. Exposure Assessment in Environmental Epidemiology

[IS 15, (page 11)]

Geoff MORGAN, *University of Sydney*

Environmental epidemiological studies quantify the association between environmental hazards and human health and a central issue is the robustness of the exposure assessment. The type of study design influences the form of the exposure assessment along with the spatial and temporal resolution of the location information on the study subjects. Many time series studies use routinely collected administrative health data (eg: mortality, hospitalisation) on populations and only have location data at the suburb level to estimate exposure. While cohort studies often have detailed information on the current and historical residential address of individuals, and increasing more detailed data such as time activity pattern data that can be used to estimate exposure. This presentation will illustrate the importance of exposure assessment in epidemiological studies using examples

including: the health effects of both short term and long term exposure to air pollution; drinking water quality and adverse birth outcomes based on mothers residential exposure during pregnancy, and; the built environment and physical activity. The presentation will also give examples of new data sources for improved exposure assessment such as satellite data and low cost sensors, as well as data blending techniques that combine exposure data at different spatial and temporal scales.

16. Exponential inequalities for sampling designs

[IS 6, (page 6)]

Guillaume CHAUVET, *Ensai - Irmar*

Exponential concentration inequalities are helpful to guaranty that the difference between a parameter and its estimator is no greater than a given threshold, with probability tending exponentially fast to 1 as the sample size increases. Such inequalities are in particular helpful in streaming algorithms, when a sample is obtained in one pass of the file only, and when so-called epsilon-delta approximations are wanted. In a recent work, Bertail and Clemencon (2019) obtained an exponential inequality for negatively associated sampling designs, a family including rejective sampling, Rao-Sampford sampling and pivotal sampling. However, the property of negative association may be difficult to prove for other sampling designs, such as Chao's procedure with is currently used in streaming algorithms. In this work, we define what we call the generalized Sen-Yates-Grundy conditions. Making use of a martingale characterization, we prove that under these conditions the Horvitz-Thompson estimator satisfies a version of the Azuma-Hoeffding therem. These conditions hold true for rejective sampling, Chao's sampling, Tille's eliminatory procedure and the generalized Midzuno method, for example. This is joint work with Mathieu Gerber (University of Bristol).

17 . Generating partially synthetic geocoded public use data with decreased disclosure risk using differential smoothing

[IS 7, (page 6)]

Harrison QUICK, *Drexel University*

When collecting geocoded confidential data with the intent to disseminate, agencies often resort to altering the geographies prior to making data publicly available. An alternative to releasing aggregated and/or perturbed data is to release synthetic data, where sensitive values are replaced with draws from models designed to capture distributional features in the collected data. The issues associated with spatially outlying observations in the data, however, have received relatively little attention. Our goal here is to shed light on this problem, propose a solution – referred to as "differential smoothing" – and illustrate our approach using sale prices of homes in San Francisco.

18. Simulation Study to Compare AFT Model Parameter Estimations Based on Pure Model-Based Approach and Model Assisted Approach

[CS 2, (page 10)]

Imali JAYAMANNE, *Department of Statistics, University of Colombo, Sri lanka*
Asoka RAMANAYAKE, *Department of Statistics, University of Colombo, Sri lanka*

When model parameters are the target of inference in survey data, the inference on the model can be based on a pure model-based approach or a model-assisted approach which is a hybrid approach combining design-based and model-based methods. However, perhaps due to lack of proper guidance and awareness, most of the data analysts apply model-based methods without considering the complex design structure of the survey data such as stratification, clustering, unequal probability of selection. In the model-assisted approach, the model specifies the parameter of interest while all inference is based on survey design. This study aims to compare these methods for survival data that are gathered from a complex survey design with equal or unequal probabilities from a stratified random sample. Accelerated Failure Time (AFT) model was fitted to describe the relationship between the interval censored response variable and the explanatory variables. Samples from the dataset were drawn using resampling methods with different sample sizes and different sampling designs. AFT models were fitted to each of the samples using three approaches to estimate model parameters. Firstly, using model- based without considering survey design and weights and secondly using model-based only considering survey weights and finally using model assisted method by considering both survey design and weights. Squared Bias, Variance and Mean squared error (MSE) were used to compare the three approaches. Consequently, for the full model, the second approach resulted in the least MSE for all scenarios. Besides, changes in the distribution of the AFT model and changes in the best set of covariates were analyzed. Average squared bias, average standard error and average MSE were used to compare three methods of analysis for the best model selected using AIC. Even though it was challenging to select the best approach for all scenarios, the second and third approach worked well for small samples than the first.

19. Evaluating the effectiveness of using livestock guarding dogs as a method of conflict resolution

[CS 3, (page 10)]

Ivor CRIBBEN, *University of Alberta*

The Cheetah Conservation Fund (CCF) is the global leader in research and conservation of cheetahs. The CCF began a livestock guarding dog program in Namibia in order to save the cheetah in the wild. Surveys were conducted on 419 livestock guarding dogs placed on farms between 1994 and 2018. The performance of the dog was evaluated on the basis of 334 questionnaires conducted on 117 livestock guarding dogs. The dogs were scored for attentiveness, trustworthiness, protectiveness, care given and farmer satisfaction. Using multiple imputation techniques and several types of logistic regression models (e.g., ordinal, mixed effects, GEE) we were able to determine the age/stage at which livestock guarding dogs were most effective in reducing livestock losses and to investigate farmer perception on whether the dog brought economic benefit influenced how well they took care of the dog. We also considered if age after placement influenced loss of livestock and whether there was a significant difference in the types of death the dogs dies from.

20. Mass imputation for combining non-probability and probability survey samples

[IS 9, (page 7)]

Jae-kwang KIM, *Iowa State University*

This paper presents theoretical results on combining non-probability and probability survey samples through mass imputation, an approach originally proposed by Rivers (2007) as sample matching without rigorous theoretical justification. Under suitable regularity conditions, we establish the consistency of the mass imputation estimator and derive its asymptotic variance formula. Variance estimators are developed using either linearization or bootstrap. Finite sample performances of the mass imputation estimator are investigated through simulation studies and an application to analyzing a non-probability sample collected by the Pew Research Centre.

21. Big data in official statistics

[IS 4, (page 4)]

Jan VAN DEN BRAKEL, *Statistics Netherlands and Maastricht University*

National statistical institutes are under increasing pressure to reduce administration costs and response burden for the production of official statistics. This could potentially be accomplished by using large data sets - so called big data. However, there are problems that must be addressed when using such data source for the production of official statistics. In these sessions, two different research lines will be presented on how big data sources can be used in the production of official statistics. They will be illustrated with research results from projects conducted at Statistics Netherlands. The first approach to be presented is to combine big data sources with sample data in a model-based inference approach. This implies that big-data sources are used as covariates in models used for small area estimation, for example in an area level model where cross-sectional correlation between areas are used to improve the effective sample size of the domains. The second approach is to use big data sources as a primary data source for the compilations of official statistics. This can be considered if a big data source covers the intended target population and not suffer to much from under- and over-coverage, e.g. the use of satellite and areal images for deriving statistical information on land use. In most cases, however, adjustments for selection bias are required.

22. Analyzing and visualizing sources of uncertainly in dynamic microsimulations

[IS 14, (page 11)]

Jan Pablo BURGARD, *Trier University*
Simon SCHMAUS,

With the incorporation of microsimulations into the Federal Statistics Act BStatG in Germany the Bundestag encourages the use of microsimulation in political decision processes. In many areas of political decision making, e.g. in transport infrastructure, microsimulation has been a strongly applied methodology. Usually such microsimulation make use of agents or representatives to incorporate differing behavior between individuals. With nowadays computer power, also demography based microsimulations come into perspective. Instead of using a subset of the population or just a few agents, with demographic based microsimulations the change is measured for a total population and for each unit within. This enables the modeling of much finer and more diverse behavioral attitudes. However, dynamic demography based microsimulations encounter many sources of errors. Whereas some of these error sources can be controlled for, others still remain without appropriate consideration. For example, the different transition processes in the population such as births and deaths, relocations, and change in household characteristics can be seen as a realization of a model or a sequence of realizations of models. Here, the uncertainty induced by the random realizations can be controlled by Monte Carlo simulation. In contrast, errors that are occurring at an earlier stage, such as the construction of the base population, or the selection and estimation of models for the transition process typically are hard to be controlled for. If possible, different scenarios can be constructed to cover the range of plausible models or estimates. One example could be the selectivity of the sampling design, where the weighted and unweighted parameter estimates of the transition model differ. This can lead to a very different interpretation of the microsimulation outcome if one or the other parameter estimation is used. It is presented how sensitivity analysis can be used to assess the uncertainty induced in the output by single parts of this error sources. Furthermore, it is discussed, how the sensitivity analysis can be used to illustrate and evaluate the output of dynamic microsimulations under consideration of the temporal component.

23. Regularized Area-level Modelling for Robust Small Area Estimation in the Presence of Unknown Covariate Measurement Errors

[IS 4, (page 4)]

Jan Pablo BURGARD, *Trier University*
Joscha KRAUSE,
Dennis KREBER,

An approach to model-based small area estimation under covariate measurement errors is presented. Using a min-max approach, we proof that regularized regression coefficient estimation is equivalent to robust optimization under additive noise. Applying this equivalence, the Fay-Herriot model is extended by l1-norm, squared l2-norm and elastic net regularizations as robustification against design matrix perturbations. This allows for reliable area-statistic estimates without distributive information about the measurement errors. A best predictor and a Jackknife estimator of the mean squared error are presented. The methodology is evaluated in a simulation study under multiple measurement error scenarios to support the theoretical findings. A comparison to other robust small area approaches is conducted. An empirical application to poverty mapping in the US is provided. Estimated economic figures fromthe US Census Bureau and crime records from the Uniform Crime Reporting Program are used to model the number of citizens below the federal poverty threshold.

24. An Approximate Bayesian Approach for Solving Different Inferential Problems in Poverty Research for Small Areas

[IS 12, (page 9)]

Jiraphan SUNTORNCHOST, *Chulalongkorn University*
Partha LAHIRI, *University of Maryland College Park*

Poverty mapping that displays spatial distribution of various poverty indices is most useful to policymakers and researchers when they are disaggregated into small geographic units, such as cities, municipalities or other administrative partitions of a country. Typically, national household surveys that contain welfare variables such as income and expenditures provide limited or no data for small areas. It is well-known that while direct survey-weighted estimates are quite reliable for national or large geographical areas they are unreliable for small geographic areas. If the objective is to find areas with extreme poverty, these direct estimates will often select small areas due to the high variabilities in the estimates. Empirical best prediction and Bayesian methods have been proposed to improve on the direct point estimates. However, these estimates are not appropriate for different inferential purposes. For example, for identifying areas with extreme poverty, these estimates would often select areas with large sample sizes. In this paper, using databases used by the Chilean Ministry for their Small Area Estimation production, we illustrate how appropriate Bayesian methodology can be developed to address different inferential problems.

25 . Simultaneous Bipartite Record Linkage and Linear Regression

[IS 3, (page 4)]

Jiurui TANG, *Duke Statistical Science Department*

Bipartite record linkage is the task of merging two data files, often in the absence of a unique identifier. Performing this task is an important step in many problems, such as healthcare, finance, human rights and other application areas, where one can obtain larger and more aggregated data sets. After the record linkage step, researchers often wish to perform more powerful statistical analysis, such as linear regression. In this talk, we assume the response variable and explanatory variable are in two files respectively and propose a Bayesian record linkage procedure to borrow information from the potential correlation in the data set if a linear regression model could be built. We will illustrate the performance of this method via simulation studies.

26. Synthetic Microdata for Establishment Surveys Under Informative Sampling

[IS 16, (page 11)]

Jörg DRECHSLER, *Institute for Employment Research*
Hang J. KIM, *University of Cincinnati*
Katherine J. THOMPSON, *U. S. Census Bureau*

Many agencies are currently investigating whether releasing synthetic microdata could be a viable dissemination strategy for highly sensitive data, such as business data, for which disclosure avoidance regulations otherwise prohibit the release of public use microdata. However, existing methods assume that the original data either cover the entire population or comprise a simple random sample from this population, which limits the application of these methods in the context of survey data with unequal survey weights. This paper discusses synthetic data generation under informative sampling. To utilize the design information in the survey weights, we rely on the pseudo likelihood approach when building a hierarchical Bayesian model to estimate the distribution of the finite population. Synthetic populations are randomly drawn from the estimated finite population density. Using simulation studies, we show that the suggested synthetic data approach offers high utility for design-based and model-based analyses while offering a high level of disclosure protection. We apply the proposed method to a subset of the 2012 U. S. Economic Census and evaluate the results with utility metrics and disclosure avoidance metrics under data attacker scenarios commonly used for business data.

27. The Conjugate Multivariate Distribution with Application to Spatial Federal Datasets

[IS 7, (page 6)]

Jonathan BRADLEY, *Florida State University*

We introduce a Bayesian approach for analyzing (possibly) high-dimensional dependent data that are distributed according to a member from the natural exponential family of distributions. This problem requires extensive methodological advancements, as jointly modeling high-dimensional dependent data leads to the so-called “big n problem.” The computational complexity of the “big n problem” is further exacerbated when allowing for non-Gaussian data models, as is the case here. Thus, we develop new computationally efficient distribution theory for this setting. In particular, we introduce something we call the “conjugate multivariate distribution,” which is motivated by the univariate distribution introduced in Diaconis and Ylvishaker (1973). Furthermore, we provide substantial theoretical and methodological development including: results regarding conditional distributions, an asymptotic relationship with the multivariate normal distribution, conjugate prior distributions, and full-conditional distributions for a Gibbs sampler. To demonstrate the wide-applicability of the proposed methodology, we provide two simulation studies and an application to data obtained by the US Census Bureau’s American Community Survey.

28 . Data integration using fractional hot deck imputation

[CS 3, (page 10)]

Jongho IM, *Yonsei University*
Inho PARK,

Jongho Im and Inho Park Data integration is a popular technique for combining multiple data sources when the outcome variables and covariates of interests are not observed together within the same dataset but are all available across datasets. As data integration can be viewed as a missing data analysis, in this study, we propose a new data integration method using fractional hot deck imputation to integrate two different datasets obtained from the same target population but have different survey items. We extended Kim, Berg and Park (2016)'s statistical matching idea in a nonparametric way. All variables are first converted to categorical types and then the joint probabilities are computed using then EM method. The estimated cell probabilities are used to compute fractional weights and those fractional weights are used to create new variables into a fixed dataset. The simulation results demonstrates that the proposed method works well for general cases and has strength in presenting of the relationship between the target variables and others. The proposed method is applied to combine two datasets collected for investigating consumer attitudes of processed foods and food consumption behavior.

29. Simultaneous Inference for Empirical Best Predictor Under Generalized Linear Mixed Models

[CS 2, (page 10)]

Katarzyna RELUGA, *University of Geneva*

We consider a simultaneous inference for the empirical best predictor (EBP) under a generalized linear mixed model (GLMM). In particular, we propose a method to construct simultaneous prediction intervals (SPIs). To the best of our knowledge, SPIs have not been developed under this modelling framework. SPIs allow researchers and practitioners to carry out statistically valid multiple comparisons of all or several parameters of interest. Aforementioned analysis can be desirable within certain domains such as small area estimation (SAE), which is often applied in, among others, studies measuring poverty, policy making or ecological and demographic projects. We focus on the maximum likelihood based estimation and we advocate a usage of the adaptive Gauss-Hermite quadrature when an analytical expression is not easily available. We provide details of derivations for two extensively employed examples of GLMM, namely the unit-level binomial mixed model and the area-level Poisson mixed model. The theoretical results are accompanied by an extensive simulation experiment and a data example. The latter reveals an advantage of SPIs in the simultaneous study of the estimators. On the other hand, in this situation, the pointwise confidence intervals do not account for the variability arising from the joint statements and may lead to completely erroneous conclusions.

30. Validating outcomes from synthetic data: results from a multi-year implementation

[IS 16, (page 11)]

Lars VILHUBER, *Cornell University*

The SDS at Cornell University was set up to provide early access to new synthetic data products by the U.S. Census Bureau. These datasets are made available to interested researchers in a controlled environment, prior to a more generalized release. Over the past 7 years, 4 synthetic datasets were made available on the server, and over 120 users have accessed the server over that time period. This paper reports on outcomes of the activity: results of validation requests from a user perspective, functioning of the feedback loop due to validation and user input, and the role of the SDS as a access gateway to and educational tool for other mechanisms of accessing detailed person, household, establishment, and firm statistics.

31. Using MRP to combine probability and non-probability sampling frames

[IS 2, (page 3)]

Lauren KENNEDY, *School of Social Work, Columbia University*

An example with the Poverty Tracker Multiple survey frames are a challenge for traditional survey weights, but the inclusion of non-probability frames increase this challenge considerably. However, non-probability frames often provide access to cases difficult to access through traditional probabilistic sampling techniques. In this talk we consider the New York City: Longitudinal Study of Wellbeing, which combines two probability samples and one non-probability sample. We demonstrate that multilevel regression and poststratification (MRP) can be used to impute non-census variables using auxiliary information and thus adjust for differences in the sample demographics between the three samples.

32. Some new methods of graph sampling

[IS 6, (page 6)]

Li-Chun ZHANG, *University of Southampton*

Finite graph sampling is a relatively new and under-developed topic area for design-based inference. Zhang and Patone (2017) synthesise and extend the existing literature. A formal definition of sample graph is proposed, and the Horvitz-Thompson (HT) estimator for arbitrary T-stage snowball sampling is developed. This talk represents some recent developments. Bipartite incidence graph (BIG) is defined as a framework of estimation for graph sampling, which allows one to explore infinitely many exactly design-unbiased estimators in addition to the HT estimator. A family of adaptive weighting estimators are developed, which includes as special cases all the design-unbiased non-HT estimators that exist in the relevant literature, such as indirect sampling, "network" sampling, adaptive cluster sampling and snowball sampling.

33. Sumca: Simple, Unified, Monte-Carlo Assisted Approach to Second-order Unbiased MSPE Estimation

[IS 1, (page 3)]

Mahmoud TORABI, *University of Manitoba, Canada*

We propose a simple, unified, Monte-Carlo assisted approach to second-order unbiased estimation of mean squared prediction error (MSPE) of a small area predictor. The proposed MSPE estimator is easy to derive, has a simple expression, and applies to a broad range of predictors that include the traditional empirical best linear unbiased predictor (EBLUP), empirical best predictor (EBP), and post model selection EBLUP and EBP as special cases. Furthermore, the leading term of the proposed MSPE estimator is guaranteed positive; the lower-order term corresponds to a bias correction, which can be evaluated via a Monte-Carlo method. The computational burden for the Monte-Carlo evaluation is much lesser, compared to other Monte-Carlo based methods that have been used for producing second-order unbiased MSPE estimators, such as double bootstrap and Monte-Carlo jackknife. Theoretical and empirical results demonstrate properties and advantages of the proposed MSPE estimator. This is a joint work with Jiming Jiang from University of California, Davis.

34. Design-based indirect small area estimation

[IS 6, (page 6)]

Maria Giovanna RANALLI, *University of Perugia*

Consider the typical problem in small area estimation of estimating the population mean of a variable of interest in each of a set of domains, as well as the overall population mean. The traditional design-based domain estimation provides a *direct* estimator of such means using values of the variable of interest only from the sample units in the domain. In the absence of auxiliary information, an Hajek-type estimator can be used that is the ratio of the Horvitz-Thompson (HT) estimator of the domain total of the variable of interest, divided by the HT estimator of the domain size. These HT estimators can be seen as using a weighting system in which (i) there are as many sets of weights as the number of domains (each set produces the estimate for a particular domain), (ii) these weights are given by the inverse of the inclusion probability, and (iii) for each set, they take value zero for units outside the domain for which the estimate is produced. We will consider this as the *basic* weighting system. It is well-known that estimates obtained using this basic weighting system can have unacceptably large variance if the domain sample size is small. In this talk, we discuss extensions to the basic weighting system that allow one to trade variance off with bias. In particular, we allow for inclusion of auxiliary information and/or for *indirect* estimation by allowing each set of weights to take non-null values for units outside the domain for which the estimate is produced. Well-known direct and indirect small area estimators fall in this class of estimators that use an *extended* weighting system. The properties of the proposed class are discussed and investigated in terms of the design-based mean squared error.

35. Statistical Methods To Survey Hidden Networked Populations

[Plenary Lecture 3, (page 8)]

Mark HANDCOCK, *University of California Los Angeles*

In many situations, standard survey sampling strategies fail because the target populations cannot be accessed through well-defined sampling frames. Typically, a sampling frame for the target population is not available, and its members are rare or stigmatized in the larger population so that it is prohibitively expensive to contact them through the available frames. We discuss statistical issues in studying hard-to-reach or otherwise "hidden" populations. These populations are characterized by the difficulty in survey sampling from them using standard probability methods. Examples in a demographic setting include unregulated workers and migrants. Examples of such populations in a behavioral and social setting include injection drug users, men who have sex with men, and female sex workers.

Hard-to-reach populations are under-served by current sampling methodologies mainly due to the lack of practical alternatives to address these methodological difficulties. We will focus on populations where some form of social network information can be used to assist the data collection. In such situations sophisticated statistical methods are needed to allow the characteristics of the population to be inferred from the collected data.

We review adaptive network sampling, including respondent-driven sampling. We develop model-assisted methods and and capture-recapture ideas.

This is joint work with Krista J. Gile, Katherine R. McLaughlin and Brian Kim.

36. Micro-simulation in Official Statistics

[IS 14, (page 11)]

Markus ZWICK, *Federal Statistical Office of Germany*

The main task of National Statistical Institutes (NSIs) is to collect data for describing the economic and social reality. Some NSIs also perform data analyses. It is a traditional discussion if NSIs should be only independent data producers or if they also should analyse their own data. Nowadays some NSIs carry out analyses combining administrative or new digital data. There are relevant arguments for both views. For Germany the answer is easier because data-analyses as well as micro-simulation are by law tasks of the Federal Statistical Office (Destatis). The German Federal Statistics Act - BStatG gives Destatis the duty to upon request by supreme federal authorities, carry out follow-up processing for federal purposes, including the development and application of micro-simulation models and micro-econometric analyses (3 passage (a) No 6 BStatG). Using micro-simulation models or performing micro-econometric analyses is part of this paragraph since 2016. Currently Destatis has few experiences to do micro-simulation and this is one of the reasons for the German NSI to participate in different research projects concerning micro-simulation so it can build up more skills in this direction. However micro-simulation has a longer tradition in Germany and Destatis. From 1979 until 1992 the German Research Foundation financed the project Microanalytical foundations of social policy. This project had a very strong influence on the economic and social empirical science for decades in Germany. Most of the used micro data in this project were official statistics. During the timeframe of the project, a lot of cooperation ran between the scientific community and official statisticians. At the beginning of the 2000s tax micro-simulation models were particularly important in Germany. Based on a change of the tax statistic law in 1996, the individual income tax declarations were accessible for Destatis and a lot of models were created in cooperation with the research community especially for policy consulting. During this time a new middle-left government changed some income tax regulations and Destatis analysed parts of this reform steps for the Federal Ministry of Finance. Since 2018 Destatis is engaged in the project MikroSim which is a common project with the universities in Trier and Duisburg and is financed by the German Research Foundation. The presentation will give an insight into the motivation of a NSI to run micro-simulation models, on their own survey data and in combination with non-traditional data as well. The focus will be a report of the experience built up in the past and the experience gained from the participation in MikroSim and Eurostat or OECD projects.

37. A General functional form of Mean Squared Error Estimator in Small Area Estimation

[IS 5, (page 4)]

Masayo HIROSE, *The Institute of Statistical Mathematics*

There are high demand to estimate small areas characteristics via the empirical best linear unbiased prediction approach. To assess such approach, it is well known to utilize the second-order unbiased estimator of mean squared error. And it is also generally known that the functional form of MSE estimation depends on the existing estimation method of model unknown parameters. In this study, under Fay-Herriot model, we seek adequate functional form of the MSE estimator for achieving several desirable properties in terms of MSE estimation, when the estimation method of model unknown parameter is changed. To attain this, we consider a general functional form of Mean squared error estimation. Moreover, we will also report the results of our simulation study for showing overall superiority of our class of MSE estimator over the others.

38. Leveraging Social Capital Questions for Travel and Activity Surveys

[IS 8, (page 7)]

Michael MANESS, *University of South Florida*

Activity participation provides opportunities to create and maintain social ties thus promoting bridging and brokerage as well as social support critical functions of social networks for promoting social mobility. Prior research has found evidence of a link between strong and weak social connections and variations in activity behavior but was limited by short activity lists and the lack of travel availability and time-use data. This study will design and analyze a web-based survey of activity preferences and social capital. The survey includes an activity list of about 85 activities and asks respondents to recall their frequency of activity participation over three months. Social capital indicators of social resource access will be obtained through position and resource generators. Additionally, the survey will include sections on in-home time use (including household chores and child care duties), travel patterns, and respondents and household demographics. A trial survey is currently running, and a full launch is expected in April 2019 with a national sample of 1500 households. This study will also include analysis of the data using machine learning and econometric techniques. Integrated choice and latent variable model (and potentially integrated count and latent variable) will be used to relate the social resources available to social capital constructs and these techniques will be compared with machine learning techniques for clustering, classification, and regression.

39. MULTIVARIATE SMALL AREA ESTIMATION UNDER NOT MISSING AT RANDOM NONRESPONSE

[IS 5, (page 4)]

Michael SVERCHKOV, *Bureau of Labor Statistics*
Danny PFEFFERMANN, *National Statistician and CBS Director, Israel*

We consider multivariate small area estimation under not missing at random (NMAR) nonresponse. We define a response model that accounts for the different patterns of the observed outcomes, (which values are observed and which ones are missing), and estimate the response probabilities by application of the Missing Information Principle (MIP). By this principle, we first define the likelihood score equations for the case where the missing outcomes are actually observed, and then integrate out the unobserved outcomes from the score equations with respect to the distribution holding for the missing data. The latter distribution is obtained from the distribution fitted to the observed data and the response model. The integrated score equations are then solved with respect to the unknown parameters underlying the response model. Once the response probabilities are estimated, we impute the missing outcomes from their appropriate distribution, yielding a complete data set with no missing values, which is used for estimating the target area means. We illustrate our approach by a small simulation study.

40. Latest development in the combined Census approach at Statistics Canada

[CS 1, (page 5)]

Michelle SIMARD, *Statistics Canada*

Statistics Canada is undergoing a vast modernization exercise; putting users at the core of its activities, making greater use of various source of data and using leading edge methods. The objective is to place greater reliance on data integration and modelling. One of the key project is the Census Program Transformation Project (CPTP) . The CPTP was launched in 2016. The purchase of the project was to develop and test a new and effective model for the Canadian Census of Population that will make optimal use of Canadian data. The transition from a traditional census to a combined census will unequivocally bring and raise new new issues about the methods for estimating coverage errors. This presentation will describe the latest development on the modernization, the CPTP, and the potential options to explore for estimating coverage errors.

41. A Hierarchical Bayesian Approach for Addressing Multiple Objectives in Poverty Research for Small Areas

[IS 12, (page 9)]

Monica PRATESI, *Department of Economics and Management, University of Pisa*
Partha LAHIRI, *University of Maryland College Park*
Gaia BERTARELLI, *University of Pisa*
Stefano MARCHETTI, *University of Pisa*
Nicola SALVATI, *University of Pisa*

Data availability often is limited by cost, by sensitivity of the questions, or for other reasons. Still, because the information extracted from data can be key to good policy and to good decisions, analysts must make the best possible use of all available information. When a hierarchical model can be found to represent the situation properly, analysis of that model often can be used to extract most or all of the relevant information and so provide the best possible estimates. Our goals are to develop appropriate large parametric models and hierarchical models and to demonstrate their ability to improve inferences about quantities for which there are meager data. The application considered will include small area estimation in the context of the European Union Statistics on income and living conditions. In developing the hierarchical model, considerable care will be taken that the analytical needs of the public policy analysts be met. The methods provided must be understandable and the fitted data must be usable by a range of social science researchers. Different levels of the model and specification of the covariates used at both levels will be justified with statistical tests of fit appropriate for multi-level models. As for the implementation of the hierarchical model, we propose to use Bayesian methodology assisted by Monte Carlo Markov Chain.

42. Distributed Markov chain Monte Carlo for Scalable Bayesian Entity Resolution

[IS 3, (page 4)]

Neil MARCHANT, *School of Computing and Information Systems, University of Melbourne*

Entity resolution (also record linkage or deduplication) seeks to identify records across and within data sources that refer to the same statistical units. It allows statisticians to leverage information that may not be available from a single source, and is also useful as a data cleansing tool. A recent development in entity resolution methodology has been the application of Bayesian generative models. These models offer several advantages over conventional methods, namely: (i) they do not require labeled training data; (ii) they treat linkage as a clustering problem which preserves transitivity; (iii) they propagate uncertainty; and (iv) they allow for flexible modeling assumptions. However, due to difficulties in scaling, these models have so far been limited to small data sets of around 1000 records. In this talk, I will propose the first scalable implementation of a popular Bayesian model for entity resolution called "blink. Our extension brings together several key ideas, including probabilistic blocking, indexing, and efficient sampling algorithms.

43. An area model to estimate multidimensional poverty index in the municipalities with the worst social conditions in Colombia.

[CS 3, (page 10)]

Nicolas RAMIREZ, *Student*

Trough history, Colombian population vulnerability has been increased by internal armed conflict. These conditions are associated with the life quality gap between urban and rural population, high poverty indexes, low educational levels, poor living conditions, and the lack of agricultural, commercial and social development. In order to face this social problem, in 2016 the national government and FARC-EP guerrilla group, signed a peace agreement to transform the territory through a comprehensive rural reform. To obtain a diagnosis of the situation prior to this transformation in a joint manner is possible to use the Multidimensional Poverty Index as a measure of the living conditions and thus have a point of reference that allows estimating the impact of the intervention. So, in this study, it is possible to obtain an accurate estimation for the Multidimensional Poverty Index for each one of the 170 municipalities that make up the territory prioritized in the implementation of the peace agreement. These results could be convenient to execute focused public policies that allow to improve the life quality of the Colombian population and reduce the vulnerability of the inhabitants of these areas in the country.

44. Need for Data Integration and Statistical Modeling for Various Purposes

[Plenary Lecture 1, (page 3)]

Partha LAHIRI, *University of Maryland College Park*

We present a few illustrative examples to explain potential needs for integrating information contained in multiple databases. For each example, we argue the usefulness of statistical modeling and the associated statistical methodology. We borrow our first example from Hansen and Hurwitz (1953), who showed how a simple regression model can be used to integrate information contained in two surveys in reducing various nonsampling errors. Our second example concerns an extension of this simple regression model to combine information from different sample surveys, administrative, and census data in improving direct estimation at granular levels. As for our third example, we consider a situation when the variables to be used in the intended data analysis are stored in multiple databases that do not have error-free unique identifiers. We will explain how models can be used to probabilistically link multiple databases and to provide valid inferences from such linked data that contains potential mismatch errors. Our fourth example concerns monitoring socio-economic and health statistics at shorter interval (e.g., every year) in order to evaluate the effectiveness of planning. Lack of available annual detailed data is the main impediment to undertake such an exercise. This needs massive resources in terms of cost and time to conduct annual survey to produce reliable estimates and it is difficult, especially for developing and underdeveloped countries, to set aside such massive resources for this purpose, given other prioritized vital socio-economic needs for their people. We will discuss how modeling can help to integrate multiple data sources to provide a viable solution to this problem. Our final example is related to traffic projection in real time. Traditional time series methods are not suitable for this purpose because of the considerable time they take for building and fitting models for traffic projections. We will discuss how historical big data on traffic patterns can be used to first build and fit a reasonable working model for short time traffic projection and then use the fitted model and current time data as an input for fast real time traffic projection. We will end the talk with a few concluding remarks.

45. Total Error Frameworks for Found Data

[IS 2, (page 3)]

Paul BIEMER, *Distinguished Fellow, Statistics, RTI International*

The survey world is relying more heavily on found data for inference and decision making rather than survey or design data. Found data are data that are not primarily collected for statistical purposes but contain information that might be useful for inference. The data become found when they are used to achieve some statistical purpose through data mining or analysis. This paper focuses on the accuracy of data produced by integrating two or more datasets, particularly when one of those datasets is from a survey and the other is found. First it considers the processes by which two or more data sources are integrated as well as the processes by which (hybrid) estimates are derived from integrated datasets. It then reviews several total error frameworks that have been proposed for evaluating the quality of the integrated dataset itself and several related frameworks for evaluating the quality of the hybrid estimates that may be produced from such datasets. The application of a total error framework is illustrated with an administrative data set that is currently being integrated with a national survey data set.

46. Data integration: the good, the bad and the ugly

[Special Invited Lecture, (page 7)]

Pedro Luis do Nascimento SILVA, *IBGE - National School of Statistical Sciences, Brazil*

Statistical data integration is here to stay: the ubiquitous abundance of data, the affordability and accessibility of computing power, the availability of statistical methods, the need to mitigate single data source defects or quality issues, the inadequate protection of privacy offered by most current legal frameworks, and the tremendous incentives and promises of yield from data integration all conspire to ensure that data from different sources will be explored for integration exercises aiming to address critical questions on economic, societal, health, environmental and other issues. In this talk, I review these pressures and the opportunities afforded by relentless technology development and innovation. I also argue for the urgent development and adoption of both ethical and legal frameworks that can support the productive and ethically acceptable data integration exercises that provide the basis for many statistical applications now and in the future. The challenge is immense: the seal has already been broken, and the tsunami of data sources available to both harmless exploration, principled research, and reckless profiteering is sweeping the world.

47 . Survey on Importance of Biostatistics in Medical Education and Research: Challenges and Experiences

[CS 1, (page 5)]

Pramod K. GUPTA, *PGIMER-Chandigarh*

This research was motivated by my experience working as faculty in medical institution, where use of biostatistics seems indispensable in the education and research. However, inappropriate application of statistics and misinterpretation of statistical inference were big challenges. The research proposal to conduct a pan India study at each medial school was granted for 2- states and 1-union territory of northern part of India by Indian Council of Medical Research, New Delhi, a central government organization. The survey areas include Punjab, Haryana and Chandigarh and secondary units were all medical colleges and institutions providing PG and higher degree in medical specialties. The primary units for data collection were final year PG and PhD residents/students and their submitted theses/dissertations. The survey strategies appeared to be straightforward as it was related to educational institutions and also due to digital development of information technology. However, it turned to be dramatically complicated and tedious that forced us for intermediate change in the underline study protocol and thus readjust the sampling proportion. The finding of the study is hereby restricted to complexities and challenges faced during data collection and adopted procedures to overcome it.

48. Microsimulations between statistics and optimization

[IS 14, (page 11)]

Ralf MÜNNICH, *Trier University*

Dynamic microsimulations are a powerful tool for making scenario-based predictions based on microdata. The basic idea goes back to Orcutt (1957) who criticized that the heterogeneity of individuals and the complexity of mutual influences often cannot be aggregated and analyzed on a higher level. Microsimulations take up this problem and offer approaches for forecasting and scenario-based analysis of social, political and economic issues at the micro level. In dynamic microsimulations, individual behavior is simulated over time using transition probabilities or hazard rates. Therefore, two components are essential for the creation of a microsimulation model: a microdata set as the base population and a method for estimating the transitions. An appropriate base population is crucial for the quality of microsimulations. In most cases, survey data is used due to its availability and a large number of observed variables. For regional analyses, however, these data can only be used to a very limited extent due to its small regional sample size and lack of geographical information. It is, therefore, evident to create and expand a suitable database. The aim of the MikroSim research project is to generate a synthetic but realistic population that includes a full German dataset based on 82 mio. individuals and households, geographically located at real addresses using census grid cells and maps. Taking into account individual and family relationships, the population is updated as a discrete-time dynamic microsimulation at annual intervals. Various administrative datasets and different statistical surveys such as the microcensus and the socio-economic panel serve as the basis for the aging of the population. Model-based transition probabilities regulate the occurrence of certain events, such as births, partnerships, and regional mobility. The given setting urges the need of using many different statistical methods including estimation and prediction considering numerous survey and administrative data. However, finally households have to be allocated in real houses. The final stage urges the application of sophisticated integer optimization methods. The presentation discusses the integrated approach of regional multi-source estimation techniques using benchmarked optimization routines.

49. Distributed Entity Resolution for Computational Social Science

[IS 3, (page 4)]

Rebecca C. STEORTS, *Duke University*

Entity resolution (ER) is the process of merging together noisy databases, often in the absence of a unique identifier. One major advancement in ER is Bayesian models, where records are clustered to unobserved (latent) entities, while providing exact uncertainty quantification and tight performance bounds. Despite these advancements, such models do not scale to realistically-sized databases as they do not incorporate probabilistic blocking. In this talk, we propose “distributed blink” or / dblink—the first scalable and distributed end-to-end Bayesian model for ER, which propagates uncertainty in blocking, matching, and merging. We make several novel contributions, including: (i) incorporating probabilistic blocking directly into the model through auxiliary partitions; (ii) support for missing values; (iii) a partially-collapsed Gibbs samper; and (iv) a novel perturbation sampling algorithm that enables fast updates of the entity attributes. We demonstrate empirically that / dblink/ can achieve efficiency gains of over 100 × when compared with the state-of-the art Bayesian ER literature.

This is joint work with Neil Marchant (PhD candidate, University of Melbourne), Andee Kaplan (Duke University), Ben Rubinstein (Associate Professor, University of Melbourne), and Daniel Elzar (ABS).

50. Combining aggregate statistics from big data and probability survey: application to electronic health records and national health surveys

[IS 9, (page 7)]

Ryung S KIM, *Albert Einstein College of Medicine*

There are increasing use of big data such as electronic health records (EHR) as a source for public health surveillance to report disease prevalence. However, often the prevalence estimates based on big data can be biased due to measurement errors or use of proxy measures. We first review existing methods to pool multiple possibly biased surveys. We note that existing methods gain in accuracy of prevalence estimates was achieved by pooling information across domains. And the existing estimators are inverse-variance weighted average of unbiased estimators. In this work, we propose a prevalence estimator that combine aggregate domain specific estimators from possibly biased big data and a gold-standard probability survey without pooling information across domains. We provide the condition when the mean squared error of the proposed estimator is less than the estimators based only on either of the sources. We performed the simulation studies to compare statistical properties of the proposed estimator with existing methods. Finally, we apply the method to Korean National Health and Nutrition Examination Survey and EHR from Korean National Health Insurance Service.

51. A prediction approach in Adaptive Sampling

[IS 17, (page 12)]

Sanghamitra PAL, *West Bengal State University*

Adaptive Cluster Sampling (ACS) due to Steven K Thompson (1990) is a tool to survey rare and hidden elements in a population as an improvement over traditional survey procedures. In ACS, if an observed sample unit satisfies the given criterion of rarity, its neighboring units are added to the sample and this is continued until one is detected with no rarity. In practice, often cluster sizes turn out too big demanding high cost and time. So, Chaudhuri, Bose, and Ghosh (2004) gave a subsampling technique to restrict the sample size in Adaptive Sampling, which was extended to unequal probability sampling. A sample size restriction technique was also developed later. But in ACS, capturing neighboring rare units turn out difficult because of various hazards. We propose to try Royalls (1970) prediction approach here to model features of uncaptured cluster elements. Considering intraclass correlation among units of a particular network in ACS, we employ Brewers (1979) model-assisted approach to derive a predictor with asymptotic design unbiasedness based on unequal probability samples and examine its efficacy by simulations.

Keywords: Adaptive sampling; Asymptotic; Model assisted approach; Prediction approach; Unequal probability.
AMS classification: 62D05

52 . A Multivariate Spatial Mixture Mixed Effects Model with Application to the American Community Survey

[IS 7, (page 6)]

Scott HOLAN, *University of Missouri*

Leveraging multivariate spatial dependence to improve the precision of the American Community Survey and other surveys has been a topic of recent interest among data-users and federal statistical agencies. One strategy is to use a multivariate spatial mixed effects model with a Gaussian observation model and latent Gaussian process model. In practice, this works well for a wide variety of tabulations. Nevertheless, in situations that exhibit heterogeneity among geographies and/or sparsity in the data, the Gaussian assumptions may be problematic and lead to underperformance. To remedy these situations, we propose a hierarchical Bayesian Multivariate Spatial Mixture Mixed Effects Model (MSMM) to increase model flexibility. The number of clusters is chosen automatically in a data-driven manner. The effectiveness of our approach is demonstrated through simulation and application to the American Community Survey.

53. Outlier robust area-level estimation using the M-quantile approach to small area

[IS 12, (page 9)]

Stefano MARCHETTI, *University of Pisa*

Decision makers and stakeholders to maximize the effect of policies need information at disaggregated geographical level. For many variables of economic interest, such information can be obtained, in principle, by producing estimates using data collected in national surveys. Often, there is lack of data for the desired level of disaggregation, thus direct survey estimates (estimates that use only the sample data from the domain of interest) are unreliable. Small area estimation methods combine survey data with auxiliary data available for example from Censuses and/or administrative registers. If the auxiliary information is available also at the statistical unit level, unit level models can be specified, otherwise, area level models are the only viable option. The most popular area-level models are linear mixed models that include area level random effects to account for between area dissimilarities beyond that explained by the auxiliary information included in the model. However, estimators based on these models can deteriorate their properties when normality assumptions of area level random effects and unit level errors do not hold. Different robust alternatives are available in the literature, such as using influence functions for estimating robust parameters in a linear mixed model and obtain robust small area estimators. We propose an alternative robust method, based on an area level M-quantile regression models of the conditional distribution of the direct estimates given the covariates. Its properties are evaluated using a Monte Carlo model-based simulation.

54. Bayesian Nonparametric Differential Analysis for Dependent Multigroup Data with Application to DNA Methylation Analyses

[IS 13, (page 9)]

Subharup GUHA, *University of Florida*

Cancer ’omics datasets involve widely varying sizes and scales, measurement variables, and correlation structures. An overarching scientific goal in cancer research is the development of general statistical techniques that can cleanly sift the signal from the noise in identifying genomic signatures of the disease across a set of experimental or biological conditions. We propose BayesDiff, a nonparametric Bayesian approach based on a novel class of first order mixture models, called the Sticky Poisson-Dirichlet process or multicuisine restaurant franchise. The BayesDiff methodology flexibly utilizes information from all the measurements and adaptively accommodates any serial dependence in the data, accounting for the interprobe distances, to perform simultaneous inferences on the variables. The technique is applied to analyze the motivating DNA methylation gastrointestinal cancer dataset, which displays both serial correlations and complex interaction patterns. In simulation studies, we demonstrate the effectiveness of the BayesDiff procedure relative to existing techniques for differential DNA methylation. Returning to the motivating dataset, we detect the genomic signature for four types of upper gastrointestinal cancer. The analysis results support and complement known features of DNA methylation as well as gene association with gastrointestinal cancer.

55. Disparity subtyping: bringing precision medicine closer to disparity science

[IS 1, (page 3)]

J. Sunil RAO, *University of Miami*
Huilin YU, *University of Miami*
Jean-Eudes DAZARD, *Case Western Reserve University*

The genomics revolution also spawned the dawn of precision medicine. Disparities researchers have also begun looking to the precision medicine paradigm with the hope that some incorporation of its principles will allow for a more focused and precise path forward to reduce population disparities. Central to the paradigm is the ability to classify individuals into subpopulations who differ in meaningful ways with respect to underlying biology and outcomes. Identification of these subpopulations is an active area of precision medicine research. How to do such a thing in disparity science has proven to be elusive since it requires identifying disparity subpopulations which is a somewhat abstract concept. In this paper we present two different strategies - level set estimation and peeling. Both are based on a recursive partitioning algorithms. The former is combined with clustering of similar partitions; the latter adopts a strategy of sequentially searching for and then extracting extreme difference subgroups in a population. Using series of simulation studies and then an analysis of ovarian cancer survival in patients from The Cancer Genome Atlas (TCGA) repository, we demonstrate that such disparity subtypes can indeed be found, characterized, and then validated on test data. Implications for further use of these subtypes will also be discussed.

56 . Are We There Yet: Differential analysis of Single-cell RNA Sequencing Data?

[IS 11, (page 8)]

Susmita DATTA, *University of Florida*

Traditionally, transcriptomic studies have examined transcript abundance measurements averaged over bulk populations of thousands (or even millions) of cells. While these bulk RNA-sequencing (RNA-Seq) measurements have been valuable in countless studies, they often conceal cell-specific heterogeneity in expression signals that may be paramount to new biological findings. Fortunately, with single cell RNA-sequencing (scRNA-Seq), transcriptome data from individual cells are now accessible, providing opportunities to investigate functional states of cells, identify rare cell populations and uncover diverse gene expression patterns in cell populations that seem homogeneous. However, there are challenges in analyzing such scRNA-Seq data. Amongst many challenges the most significant are the bimodal or multimodal distribution, sparsity and tremendous heterogeneity in the data. Consequently, we will describe potential ways of statistical modeling of such data, finding differentially expressed genes and possible ways of constructing gene-gene interaction network using this data. Moreover, we will compare the performance of our modeling and differential analysis with respect to some other existing methods.

57 . Partially Observed Mixed Effects Model with Application to Missing Data Problems

[IS 1, (page 3)]

Thuan NGUYEN, *Oregon Health & Science University, USA*

Mixed effects models typically involve random effects which, by definition, are unobserved. In some practical situations, however, some of these random effects are observed, or can be treated as observed. This occurs, for example, in the analysis of longitudinal data when some of the covariates are missing. One natural idea of handling the missing covariates is to treat them as unobserved random effects; then, the observed covariates correspond to random effects that are observed. We propose a nonlinear mixed effects model, in which part of the random effects are observed. The model, called partially observed mixed effects model, may be viewed as a nonlinear mixed effects model, except that some of the random effects are observed. As a result, the method of inference is different from the standard procedure of mixed model analysis. We derive maximum likelihood (ML) estimation and establish asymptotic properties of the estimators. Furthermore, we derive an empirical best predictor (EBP) of the missing covariates as well as a measure of uncertainty for the EBP. Finite-sample performance of the proposed estimators and predictors is evaluated via simulation studies. A real data application is considered.

58. A comparison of likelihood-based methods for size-biased sampling

[CS 2, (page 10)]

Victoria LEAVER, *University of Wollongong*

Three likelihood approaches to estimation under informative sampling are compared using a special case for which analytic expressions are possible to derive. The sampling method is selection probability proportional to a power of size from a population of known size drawn from a gamma distribution with known shape parameter. The sampling is with replacement, so that duplicate sample units are possible. Mean and variance estimators are derived for maximum likelihood estimation, sample likelihood and pseudo-likelihood. Theoretical properties of the variance estimators and simulation results show that the efficiency of the sample likelihood approaches that of full maximum likelihood estimation when the population size is large compared with the sample size. However, when the sample size is close to the population size, the MLE approach is more efficient than the other methods because it takes the possibility of duplicate sample units into account. The efficiency of the sample likelihood approach is particularly affected by duplicate units. For the special case when the superpopulation is exponential and the selection is probability proportional to size, the model-based expectation of the variance for the pseudo-likelihood estimate is infinite.

59. Statistical Data Integration and Inference via Multilevel Regression and Poststratification

[IS 2, (page 3)]

Yajuan SI, *University of Michigan*

Rapidly decreasing response rates and increasing measurement error rates motivate statistical agencies to put research priorities on combining multiple data sources, such as administrative records and surveys. We develop a unified framework under multilevel regression and poststratification (MRP) for data integration and inferences and handle the methodology issues such as the selection of auxiliary variables, unknown control information, and combination of probability and nonprobability-based surveys. In contrast with the superpopulation and weighting framework, MRP combines prediction and weighting as a hybrid approach. Leveraging the flexible Bayesian paradigm, MRP propagates all sources of uncertainty and stabilizes small area estimation while accounting for sample selection and response mechanisms into modeling. We use simulation studies to evaluate the frequentist properties and compare with alternative methods. The proposal is demonstrated and applied to real-life survey studies.

60 . A Kernel Weighting Approach to Improve Population Representativeness for Estimating Prevalence of Risk-factors and Diseases

[IS 13, (page 9)]

Yan LI, *University of Maryland College Park*

Epidemiological (EPI) cohorts are often collected from convenient samples, suffering from selection bias and coverage issue. Accordingly, the analyses are lack of external validity, leading to biased point estimation and invalid variance estimation. We develop an efficient kernel weighting approach that treats a survey sample as a reference to create pseudo weights for the cohort by utilizing propensity scores and kernel smoothing techniques. The proposed jackknife (JK) variance estimator, in addition to accounting for the correlation induced by the homogeneity of the participants from the same study center, considers the variability due to estimating propensity scores. Comparing existing inverse propensity score- (IPS-) and the proposed kernel-weighted estimators, Monte Carlo simulation studies are conducted and show that the proposed kernel-weighed estimators reduce the bias and increase the efficiency of the estimated disease prevalence. The developed approach is further demonstrated using National Health Interview Survey and the Prostate, Lung, Colorectal, and Ovarian cohort to estimate the prevalence of emphysema.

61 . Inference for Generalized Linear Models with A Diverging Number of Covariates: A Refined De-Biased Approach

[IS 11, (page 8)]

Yi LI, *University of Michigan*

Statistical inference for models with large numbers of covariates has sparked much interest, and the de-sparsified lasso approach has emerged as a useful tool. Extensive numerical simulations indicate that for the generalized linear models, however, the existing de-sparsified lasso approach cannot satisfactorily remove biases, thus yields unreliable confidence intervals largely because the sparsity condition on the inverse Hessian matrix does not hold in these settings. To properly remove the biases, we devise an alternative approach to the original de-sparsified lasso by directly inverting the Hessian matrix of the concerned model in the large n, diverging p scenario, and develop asymptotic theory that relaxes the stringent assumptions on signal and matrix sparsity to draw inference on any linear combinations of the regression coefficients. Extensive simulations reveal that the refined approach satisfactorily eliminates biases thus results in better confidence interval coverage than the original de-sparsified lasso approach. We also apply the proposed approach to estimate the effects of candidate genes on the lung cancer risk among the Boston Lung Cancer Survival Cohort, a prospective hospital-based cohort study.

62. Observed best selective prediction in small area estimation

[IS 10, (page 8)]

Yuki KAWAKUBO, *Chiba University*

In small area estimation methodology, selection of the suitable covariates and estimation in the selected model are usually considered separately. In this talk, we consider variable selection and estimation simultaneously to minimize the total mean squared prediction errors (MSPE) for estimation of small area means. The derived method, which we call observed best selective prediction (OBSP), can be regarded as an extension of the observed best prediction (OBP) method by Jiang et al. (2011). When the true model is included in the largest model, the resulting OBSP estimator is consistent. Based on the asymptotic result, we derive an estimator of MSPE by applying the parametric bootstrap method. Through simulation experiments, we investigate the finite sample performance of OBSP together with OBP in which the variable selection is carried out by using AIC and BIC, and OBP using all the covariates. As an example, we applied OBSP to the synthetic income data in Spain.

63 . Kernel-based Functional Calibration Estimation in Survey Sampling

[IS 9, (page 7)]

Zhengyuan ZHU, *Iowa State University*

Calibration estimation is a commonly used technique in survey sampling where the sampling weights are adjusted to match the population total of a set of auxiliary variables. When the auxiliary variables are observed for all units in the population, one can calibrate by matching the population total of functions of auxiliary variables. In this paper we develop a kernel-based method to do calibration by matching functions of auxiliary variables for all functions in an infinite dimensional reproducing kernel Hilbert space (RKHS). Numerical algorithms are developed and implemented to solve the optimization problem, and theoretical results on the consistency of such estimates will be presented as well. Simulation results indicate that our approach performs better than a number of widely used alternatives.

Directory

ANGKUNSIT, Annop
Chulalongkorn University
annop_ngoun@hotmail.com

BAS, Javier
University of Maryland
jbas@umd.edu
Coauthor: IS 8, p. 7, § 6, p. 16

BERTARELLI, Gaia
University of Pisa
gaia.bertarelli@ec.unipi.it
Coauthor: IS 12, p. 9, § 41, p. 27

BIEMER, Paul
Distinguished Fellow, Statistics, RTI International
ppb@rti.org
Speaker: IS 2, p. 3, § 45, p. 29

BONNERY, Daniel
University of Maryland College Park
dbonnery@umd.edu
Speaker: IS 16, p. 11, § 7, p. 17

BRADLEY, Jonathan
Florida State University
jrbradley@fsu.edu
Speaker: IS 7, p. 6, § 27, p. 23

BURGARD, Jan Pablo
Trier University
burgardj@uni-trier.de
Speaker: IS 14, p. 11, § 22, p. 21,
Speaker: IS 4, p. 4, § 23, p. 22

CHATTERJEE, Snigdhansu
University of Minnesota
chatt019@umn.edu
Organizer: IS 17, p. 12

CHAUDHURI, Sanjay
National University of Singapore
stasc@nus.edu.sg
Organizer: IS 13, p. 9

CHAUVET, Guillaume
Ensai - Irmar
chauvet@ensai.fr
Speaker: IS 6, p. 6, § 16, p. 20

CHERCHI, Elisabetta
Newcastle University
Elisbetta.Cherchi@newcastle.ac.uk
Coauthor: IS 8, p. 7, § 6, p. 16

CIRILLO, Cinzia
University of Maryland College Park
ccirillo@umd.edu
Chair and organizer: IS 8, p. 7,
Speaker: IS 8, p. 7, § 6, p. 16

COELHO, Isabela Bertolini
NIC.BR / CETIC.BR
isabelabcoelho@gmail.com

CRIBBEN, Ivor
University of Alberta
cribben@ualberta.ca
Speaker: CS 3, p. 10, § 19, p. 21

DAS, Snigdha
University of Calcutta
snigdha.stats@gmail.com

DATTA, Gauri Sankar
U.S. Census Bureau and University of Georgia
gaurisdatta@gmail.com
Chair: IS 7, p. 6,
Organizer: IS 10, p. 8,
Speaker: IS 10, p. 8, § 13, p. 18

DATTA, Somnath
University of Florida
somnath.datta@ufl.edu
Organizer: IS 11, p. 8

DATTA, Susmita
University of Florida
susmita.datta@ufl.edu
Speaker: IS 11, p. 8, § 56, p. 32

DAZARD, Jean-Eudes
Case Western Reserve University

Coauthor: IS 1, p. 3, § 55, p. 32

DEY, Anurag
Indian Statistical Institute, Calcutta
deyanuragsaltlake64@gmail.com

DINIZ DA SILVA, Andrea
Brazilian Institute of Geography and Statistics and National School of Statistical Sciences
andrea.diniz100@gmail.com
Chair: CS 2, p. 9,
Speaker: IS 4, p. 4, § 1, p. 15

DRECHSLER, Jörg
Institute for Employment Research
joerg.drechsler@iab.de
Chair and organizer: IS 16, p. 11,
Speaker: IS 16, p. 11, § 26, p. 23

EARNEST, Arul
Monash University
arul.earnest@monash.edu
Chair and organizer: IS 15, p. 11,
Speaker: IS 15, p. 11, § 2, p. 15

GUHA, Subharup
University of Florida
s.guha@ufl.edu
Chair: IS 11, p. 8,
Speaker: IS 13, p. 9, § 54, p. 32

GUPTA, Pramod K.
PGIMER-Chandigarh
guptapkg@gmail.com
Speaker: CS 1, p. 5, § 47, p. 29

HANDCOCK, Mark
University of California Los Angeles
handcock@ucla.edu
Speaker: Plenary Lecture 3, p. 8, § 35, p. 25

HIROSE, Masayo
The Institute of Statistical Mathematics
masayo@ism.ac.jp
Chair: IS 12, p. 9,
Speaker: IS 5, p. 4, § 37, p. 26

HOLAN, Scott
University of Missouri
holans@missouri.edu
Organizer: IS 7, p. 6,
Chair: IS 10, p. 8,
Speaker: IS 7, p. 6, § 52, p. 31

IM, Jongho
Yonsei University
ijh38@yonsei.ac.kr
Speaker: CS 3, p. 10, § 28, p. 23

IRIMATA, Kyle

Coauthor: IS 10, p. 8, § 13, p. 18

JAYAMANNE, Imali
Department of Statistics, University of Colombo, Sri lanka
imali@stat.cmb.ac.lk
Speaker: CS 2, p. 10, § 18, p. 20

KAWAKUBO, Yuki
Chiba University
kawakubo@chiba-u.jp
Speaker: IS 10, p. 8, § 62, p. 34

KENNEDY, Lauren
School of Social Work, Columbia University
lauren.kennedy729@gmail.com
Speaker: IS 2, p. 3, § 31, p. 24

KERN, Christoph
University of Mannheim
c.kern@uni-mannheim.de
Speaker: CS 3, p. 10, § 5, p. 16

KIM, Hang J.
University of Cincinnati

Coauthor: IS 16, p. 11, § 26, p. 23

KIM, Jae-kwang
Iowa State University
jkim@iastate.edu
Chair and organizer: IS 9, p. 7,
Speaker: IS 9, p. 7, § 20, p. 21

KIM, Ryung S
Albert Einstein College of Medicine
ryung.kim@einstein.yu.edu
Speaker: IS 9, p. 7, § 50, p. 30

KOBAYASHI, Genya
Chiba University
gkobayashi@chiba-u.jp
Speaker: IS 10, p. 8, § 14, p. 19

KRAUSE, Joscha

Coauthor: IS 4, p. 4, § 23, p. 22

KREBER, Dennis

Coauthor: IS 4, p. 4, § 23, p. 22

KUNDU, Debashis
Indian Institute of Technology Kanpur
Kundu@iitk.ac.in
Speaker: IS 11, p. 8, § 9, p. 17

LAHIRI, Partha
University of Maryland College Park
plahiri@umd.edu
Chair: Special Invited Lecture, p. 7,
Coauthor: IS 12, p. 9, § 24, p. 22,
Coauthor: IS 12, p. 9, § 41, p. 27,
Speaker: Plenary Lecture 1, p. 3, § 44, p. 28

LAPHUDOMSAKDA, Kodchapown
Chulalongkorn University
kodchapown.l@gmail.com

LEAVER, Victoria
University of Wollongong
victoria.leaver@abs.gov.au
Speaker: CS 2, p. 10, § 58, p. 33

LI, Yan
University of Maryland College Park
yli6@umd.edu
Chair: IS 13, p. 9,
Speaker: IS 13, p. 9, § 60, p. 34

LI, Yi
University of Michigan
yili@med.umich.edu
Speaker: IS 11, p. 8, § 61, p. 34

LUNA HERNANDEZ, Angela
University of Southampton
a.luna.hernandez@soton.ac.uk

MÜNNICH, Ralf
Trier University
muennich@uni-trier.de
Chair and organizer: IS 14, p. 11,
Speaker: IS 14, p. 11, § 48, p. 30

MANESS, Michael
University of South Florida
manessm@usf.edu
Speaker: IS 8, p. 7, § 38, p. 26

MAPLES, Jerry

Coauthor: IS 10, p. 8, § 13, p. 18

MARCHANT, Neil
School of Computing and Information Systems, University of Melbourne
n.marchant@student.unimelb.edu.au
Speaker: IS 3, p. 4, § 42, p. 28

MARCHETTI, Stefano
University of Pisa
stefano.marchetti@unipi.it
Coauthor: IS 12, p. 9, § 41, p. 27,
Speaker: IS 12, p. 9, § 53, p. 31

MONDAL, Debashis
Oregon State University
debashis@stat.oregonstate.edu
Chair: Plenary Lecture 3, p. 8,
Speaker: IS 5, p. 4, § 10, p. 18

MONDAL, Pronoy Kanti
Indian Statistical Institute, Kolkata
pronoykantimondal@gmail.com

MORGAN, Geoff
University of Sydney
geoffrey.morgan@sydney.edu.au
Speaker: IS 15, p. 11, § 15, p. 19

MOURA, Fernando
Federal University of Rio de Janeiro

Coauthor: CS 1, p. 5, § 12, p. 18

NANDRAM, Balgobin
Professor of Statistics, Worcester Polytechnic Institute
balnan@wpi.edu
Chair: IS 17, p. 12,
Speaker: IS 17, p. 12, § 4, p. 15

NEVES, André
National School of Statistical Sciences

Coauthor: CS 1, p. 5, § 12, p. 18

NGUYEN, Thuan
Oregon Health & Science University, USA
nguythua@ohsu.edu
Speaker: IS 1, p. 3, § 57, p. 33

OKRASA, Wlodzimierz
Statistics Poland
W.Okrasa@stat.gov.pl

PAL, Samhita
University of Calcutta
samhitapal3896@gmail.com

PAL, Sanghamitra
West Bengal State University
mitrapal2013@gmail.com
Chair: CS 3, p. 10,
Speaker: IS 17, p. 12, § 51, p. 31

PARK, Inho

Coauthor: CS 3, p. 10, § 28, p. 23

PARK, Seunghwan
Kangwon National University
stat.shpark@kangwon.ac.kr

PFEFFERMANN, Danny
National Statistician and CBS Director, Israel
MsDanny@cbs.gov.il
Speaker: Plenary Lecture 2, p. 6, § 8, p. 17,
Coauthor: IS 5, p. 4, § 39, p. 27

PITTA, Marcelo
Brazilian Network Information Center (NIC.br)
marcelopitta@nic.br

PRAJAPATI, Deepak
Indian Institute of Technology Kanpur
deepakdp@iitk.ac.in
Coauthor: IS 11, p. 8, § 9, p. 17

PRATESI, Monica
Department of Economics and Management, University of Pisa
monica.pratesi@unipi.it
Speaker: IS 12, p. 9, § 41, p. 27

QUICK, Harrison
Drexel University
harryq@gmail.com
Speaker: IS 7, p. 6, § 17, p. 20

RAMANAYAKE, Asoka
Department of Statistics, University of Colombo, Sri lanka

Coauthor: CS 2, p. 10, § 18, p. 20

RAMIREZ, Nicolas
Student
nicolasramirezv@usantotomas.edu.co
Speaker: CS 3, p. 10, § 43, p. 28

RANALLI, Maria Giovanna
University of Perugia
giovanna.ranalli@unipg.it
Speaker: IS 6, p. 6, § 34, p. 25

RAO, J. Sunil
University of Miami
JRao@med.miami.edu
Speaker: IS 1, p. 3, § 55, p. 32

RELUGA, Katarzyna
University of Geneva
Katarzyna.Reluga@unige.ch
Speaker: CS 2, p. 10, § 29, p. 24

SALVATI, Nicola
University of Pisa

Coauthor: IS 12, p. 9, § 41, p. 27

SAWONGNAM, Siwarak
Chulalongkorn University
siwarak.sawongnam@gmail.com

SCHMAUS, Simon

Coauthor: IS 14, p. 11, § 22, p. 21

SEN, Anisha
University of Calcutta
anishasen1709@gmail.com

SI, Yajuan
University of Michigan
yajuan@umich.edu
Speaker: IS 2, p. 3, § 59, p. 33

SILVA, Denise
National School of Statistical Sciences
denisebritz@gmail.com
Speaker: CS 1, p. 5, § 12, p. 18

SILVA, Pedro Luis do Nascimento
IBGE - National School of Statistical Sciences, Brazil
pedronsilva@gmail.com
Speaker: Special Invited Lecture, p. 7, § 46, p. 29

SIMARD, Michelle
Statistics Canada
michelle.simard@canada.ca
Speaker: CS 1, p. 5, § 40, p. 27

SINHA, DEBAJYOTI
FLORIDA STATE UNIVERSITY
sinhad@stat.fsu.edu
Speaker: IS 13, p. 9, § 11, p. 18

SLUD, Eric

Coauthor: IS 10, p. 8, § 13, p. 18

SOMJETTANA, Sarasinee
Chulalongkorn University
sarasineesom@hotmail.com

STEORTS, Rebecca C.
Duke University
beka@stat.duke.edu
Speaker: IS 3, p. 4, § 49, p. 30

SUNTORNCHOST, Jiraphan
Chulalongkorn University
Jiraphan.S@chula.ac.th
Organizer: IS 12, p. 9,
Speaker: IS 12, p. 9, § 24, p. 22

SVERCHKOV, Michael
Bureau of Labor Statistics
Sverchkov.Michael@bls.gov
Speaker: IS 5, p. 4, § 39, p. 27

TANG, Jiurui
Duke Statistical Science Department
jiurui.tang@duke.edu
Speaker: IS 3, p. 4, § 25, p. 23

THOMPSON, Katherine J.
U. S. Census Bureau

Coauthor: IS 16, p. 11, § 26, p. 23

TORABI, Mahmoud
University of Manitoba, Canada
Mahmoud.Torabi@umanitoba.ca
Speaker: IS 1, p. 3, § 33, p. 25

TRAN, Bac
U.S. Census Bureau
Bac.Tran@census.gov
Speaker: CS 2, p. 9, § 3, p. 15

VAN DEN BRAKEL, Jan
Statistics Netherlands and Maastricht University
jbrl@cbs.nl
Speaker: IS 4, p. 4, § 21, p. 21

VILHUBER, Lars
Cornell University
Lars.vilhuber@cornell.edu
Speaker: IS 16, p. 11, § 30, p. 24

YU, Huilin
University of Miami

Coauthor: IS 1, p. 3, § 55, p. 32

ZHANG, Li-Chun
University of Southampton
L.Zhang@soton.ac.uk
Organizer: IS 6, p. 6,
Speaker: IS 6, p. 6, § 32, p. 24

ZHU, Zhengyuan
Iowa State University
zhuz@iastate.edu
Speaker: IS 9, p. 7, § 63, p. 34

ZWICK, Markus
Federal Statistical Office of Germany
markus.zwick@destatis.de
Speaker: IS 14, p. 11, § 36, p. 26