Year of Data Science Workshop on Small Area Data Analytics

University of Maryland, College Park
March 30 - April 3, 2020

Organizers:

Partha Lahiri Department of Mathematics & Joint Program in Survey Methodology (JPSM)

Cinzia Cirillo Department of Civil & Environmental Engineering

Sponsors:

Office of the Provost, University of Maryland College Park

Maryland Transportation Institute (MTI), University of Maryland College Park

Department of Mathematics, University of Maryland College Park

Joint Program in Survey Methodology (JPSM)

United Nations Statistical Institute for Asia and the Pacific

Speakers and discussion leaders of this workshop represent the following agencies, organizations and universities:

University of Maryland College Park

Baanda Inc., Los Angeles

BPS-Statistics Indonesia

Bureau of Labor Statistics

Census Bureau

Department of Energy

General Statistics Office of Vietnam

HydroQuebec, Canada

Indian Statistical Institute

IREQ, Science des donnees and calcul haute performance, Varennes, Canada

Italian National Institute of Statistics

Johns Hopkins Bloomberg School of Public Health

National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture

National Cancer Institute

The National Center for Health Statistics

Qatar Computing Research Institute QCRI Social Computing

United Nations Economic Commission forLatin America and the Caribbean (ECLAC)

University of Montreal, Canada

University of Pisa, Italy

University of Trier, Germany

Westat

The World Bank

Glossary

AS	Application Session
DS	Discussion Session

Contents

Programme Summary

Tentative Program

Monday March 30

Registration 8:00 - 8:30

Welcome Session
Time : **8:30 - 9:00**

AS 1 Challenges in Cancer Surveillance Research with Small Populations

Time : **9:00 - 10:30**

Overview of challenges in cancer control research with small population [Abstract 6]

David BERRIGAN, *National Cancer Institute*

Small area estimation and issues for measures related to cancer surveillance [Abstract 5]

Benmei LIU, *National Cancer Institute, National Institutes of Health*

Modeling Approach for Cancer Incidence Completeness Measure [Abstract 16]

Li ZHU, *National Cancer Institute*

Coffee Break 10:30 - 11:00

Special Invited Talk 1

Time : **11:00 - 12:00**

Small Domain Analytics: Goals, Issues and Methods [Abstract 28]

Thomas A. LOUIS, *Department of Biostatistics, Johns Hopkins Bloomberg School of Public Heath, Baltimore MD USA*

Lunch 12:00 - 13:15

DS 1 Discussion of AS 1

Time : **13:15 - 14:00**

AS 2 Small area estimation with data from complex health surveys and vital statistics

Time : **14:00 - 15:30**

Modeling county-level drug overdose death rates [Abstract 15]

Lauren ROSSEN, *Division of Research and Methodology,National Center for Health Statistics at the Centers for Disease control and Prevention*

County-level Estimates of Cancer Screening and Cancer Risk Factors in the United States: Combining Information for the National Health Interview Survey (NHIS) and the Behavioral Risk Factors Surveillance Survey(BRFSS), 2011 2016 [Abstract 12]

John R. PLEIS, *National Center for Health Statistics, U.S. Centers for Disease Control and Prevention*

Computational methods in Bayesian disease mapping: an application to Chronic Lower Respiratory Disease (CLRD) mortality in the U.S., 2000-2017 [Abstract 9]

Diba KHAN, *Division of Research and Methodology, National Center for Health Statistics at the Centers for Disease control and Prevention*

Issues concerning direct state-level estimation on the National Health Interview Survey [Abstract 23]

Robin COHEN, *National Center for Health Statistics at the Centers for Disease control and Prevention*

Coffee Break 15:30 - 16:00

DS 2 Discussion of AS 2

Time : **16:00 - 16:45**

Reception 16:45 - 18:00

Tuesday March 31

Registration 8:30 - 9:00

AS 3 Ever Greater Detail: Modeling to Provide More to Consumers of Agricultural Estimates

Time : **9:00 - 10:30**

A Preface on Small Area Analytic Approaches for Crop and Economic Estimatesat the United States Department of Agriculture National Agricultural Statistics Service(USDA NASS) [Abstract 20]

Nathan B. CRUZE, *USDA NASS*

Research Challenges in Model-Based Estimation ofAgricultural Cash Rental Rates [Abstract 18]

Michael E. BELLOW, *USDA NASS*

Preserving Acreage Relationships in Small Area Agricultural Models [Abstract 17]

Lu CHEN, *USDA NASS*

Kriging and Co-Kriging approaches for Representing Crop Progress and Conditionat Small Domains [Abstract 3]

Arthur ROSALES, *USDA NASS*

Coffee Break 10:30 - 11:00

Special Invited Talk 2

Time : **11:00 - 12:00**

Challenges with Statistical Studies of Small Populations [Abstract 11]

Graham KALTON, *Westat*

Lunch 12:00 - 13:15

DS 3 Discussion of AS 3

Time : **13:15 - 14:00**

AS 4 Data driver methods for Energy related problems

Time : **14:00 - 15:30**

Short-term Demand Forecasting on the Quebec Power Grid: Challenges Ahead [Abstract 27]

Stéphane DELLACHERIE, *Hydro-Québec TransÉnergie, Montréal (Québec), Canada*
Arnaud ZINFLOU, *IREQ, Science des données and calcul haute performance, Varennes, Canada*

Coffee Break 15:30 - 16:00

DS 4 Discussion of AS 4

Time : **16:00 - 16:45**

Wednesday April 1

Registration 8:30 - 9:00

AS 5 Micro-simulation, Record Linkage and Big Data

Time : **9:00 - 10:30**

TBA [Abstract 19]

Monica PRATESI,

Casualties in road accidents: challenges and opportunities with small area estimation for linked data sets [Abstract 29]

Tiziana TUOTO, *ISTAT, Italy*

TBA [Abstract 22]

Ralf MUNNICH,

Coffee Break 10:30 - 11:00

Special Invited Talk 3

Time : **11:00 - 12:00**

TBA

Lunch 12:00 - 13:15

DS 5 Discussion of AS 5

Time : **13:15 - 14:00**

AS 6 Sustainable Development Goals Indicators

Time : **14:00 - 15:30**

Small Area Estimation using Fay-Herriot model: Household poverty rate for districtlevel in Vietnam [Abstract 21]

Nguyen THI THANH TAM, *General Statistics Office of Vietnam*

DISAGGREGATING AVERAGE HOURY EARNINGS ONDISTRICT LEVEL IN INDONESIA USING SMALL AREA ESTIMATION [Abstract 8]

Dhiar NIKEN LARASATI, *BPS-Statistics Indonesia*

Leave No One Behind: A Continuous Challenge on SDG Monitoring Through Small Area Estimation in Latin-America [Abstract 2]

Andrs GUTIRREZ, *Statistics Division, United Nations Economic Commission forLatin America and the Caribbean (ECLAC)*

Coffee Break 15:30 - 16:00

DS 6 Discussion of AS 6

Time : **16:00 - 16:45**

Thursday April 2

Registration 8:30 - 9:00

AS 7 Challenges for Small Area Methodology based on Data Requirements at the US Census Bureau

Time : **9:00 - 10:30**

Scalability issues for SAE methods [Abstract 31]

Wesley BASEL, *U.S. Census Bureau*

Small Area Estimation in Government Surveys (U.S. Census Bureau) [Abstract 4]

Bac TRAN, *US Census Bureau*

Small Area Estimation in the Census Bureau 2016 Determinations for Multiple Language Ballots under the Voting Rights Act [Abstract 10]

Eric SLUD, *University of Maryland College Park and US Census Bureau*

Coffee Break 10:30 - 11:00

Special Invited Talk 4

Time : **11:00 - 12:00**

Administrative Data and Evidence-Based Policymaking [Abstract 14]

Katherine ABRAHAM, *University of Maryland College Park*

Lunch 12:00 - 13:15

DS 7 Discussion of AS 7

AS 8 Poverty and Employment Statistics at Granular Levels

Time : **14:00 - 15:30**

Can a Poverty Map Based on Remote Sensing Data Replicate One Based on Census Data? An Assessment for Malawi [Abstract 24]

Roy VAN DER WEIDE, *The World Bank*

Big data for small area estimation [Abstract 7]

David NEWHOUSE, *The World Bank*

Recent Work in Small Domain Modeling for the Current Employment Statistics Survey [Abstract 13]

Julie GERSHUNSKAYA, *Bureau of Labor Statistics*

Coffee Break 15:30 - 16:00

DS 8 Discussion of AS 8
Time : **16:00 - 16:45**

Workshop Dinner 18:00 - 20:00

Friday April 3

Registration 8:30 - 9:00

AS 9 Research challenges and opportunities in emerging problems of small area data analytics

Time : **9:00 - 10:30**

Solving Global Socio-Economic Issues with Humanoid AI Engine [Abstract 25]

Sarbojit MUKHERJEE, *Baanda Inc, Los Angeles*

Spatial Sensitivity Analysis for Urban Hotspots using Cell Phone Traces [Abstract 30]

Vanessa FRIAS-MARTINEZ,

Data-driven research for transportation [Abstract 26]

Sofiane ABBAR, *Qatar Computing Research Institute QCRI Social Computing*

Coffee Break 10:30 - 11:00

Special Invited Talk 5

Time : **11:00 - 12:00**

Statistical Machine Learning [Abstract 1]

Amita PAL, *Indian Statistical Institute, Kolkata*

Lunch 12:00 - 13:15

DS 9 Discussion of AS 9

Time : **13:15 - 14:00**

Concluding Remarks

Time : **14:00 -**

Abstracts

1. Statistical Machine Learning

[Special Invited Talk 5, (page 11)]

PAL, **Amita**,*Indian Statistical Institute, Kolkata*

Statistical Machine Learning is an integral component of Data Analytics. It embodies an algorithmic approach derived from statistical models, for solving certain problems generally encountered in the domain of Artificial Intelligence, that can be implemented through computers. Machine learning algorithms build a mathematical model from the available data, known as "training data", in order to make predictions/decisions or to explore the underlying structure. Depending on whether training data is labeled/ unlabeled, a variety of supervised/unsupervised Statistical Machine Learning methods exist thatare respectively predictive or exploratory in nature.

The supervised approach is extremely useful for predicting categories to which test data samples belong. On the other hand, unsupervised methods are more appropriate for exploring the dataset to identify possible categories in it. Machine learning is the essential tool by which meaningful categorical information on a given variable of interest can be inferred from new and unconventional sources of datathat are available in abundance, like satellite data and cell-phone data. These new sources of data are particularly relevant where reliable data on the variable of interest is not easily available. For example, socioeconomic categories (e.g., very poor, poor, a verage, prosperous and very prosperous) can be defined for a given population on the basis of certain parameters, which can then be learned from easily available data like cell-phone usage of some individuals (training samples) from each category using unsupervised methods. The learned categories can then be used to predict the socioeconomic category that a particular member of the population belongs to, through the application of supervisedmethods, using his/her cell-phone usage data.

An overview of the most widely used statistical machine learning approaches will be provided in this talk, and application to the problems of automatic speaker recognition (ASR) and content-based image retrieval (CBIR) will be briefly described.

2 . Leave No One Behind: A Continuous Challenge on SDG Monitoring Through Small Area Estimation in Latin-America

[AS 6, (page 7)]

GUTIRREZ, **Andrs**,*Statistics Division, United Nations Economic Commission forLatin America and the Caribbean (ECLAC)*

The United Nations Economic Commission for Latin America and the Caribbean (ECLAC) regularly analyses and presents data based on household surveys. However, traditional data sources used to produce official statistics face several limitations to provide information for disaggregated population groups.Moreover, the implementation of the Sustainable Development Goals (SDG) requires addressing the challenge of leaving no one behind and overcoming disparities, an important topic that have historically characterized the Latin American region. To tackle these limitations, ECLAC is encouraging countries with the use of methods that combine information from different data sources (household surveys, population censuses, and administrative records), such as those provided by the Small Area Estimation (SAE) approach. In this talk, we show some real applications of SAE methods and the challenges that come with these efforts when estimating disaggregated indicators for poverty, contraception, nutrition, and access to justice inLatin American countries.

3. Kriging and Co-Kriging approaches for Representing Crop Progress and Conditionat Small Domains

[AS 3, (page 5)]

ROSALES, **Arthur**,*USDA NASS*

The USDA National Agricultural Statistics Service (NASS) provides crop progress and condition estimates for selected crops on a weekly basis during the crop specific growing season. These estimates are based on data from the non-probability crop progress and condition survey, which targetsrespondents whose occupations allow them to observe crop activity in their countyand bring them in contact with their local farmers. Approximately 3600 respondents complete weekly questionnaires, where they report on progress and conditions of crops in their counties. Crop progress reporting is based on standard definitionsof phenological stages, and crop condition is based on subjective evaluationsby the respondents.As a rule, any crop can progress from 0% planted to 100% harvested. The condition of any crop for any week is represented in 5 exhaustive categories, ranging at the extremes from 100% very poor to 100% excellent.Weekly reports arereviewed for reasonableness and consistency. Aggregation of the data to state and national levels relies on weights derived from historical NASS acreage estimates.

The survey is designed to provide national and state estimates, even though data are collected from respondents at the county level. It is common for only one or two respondents to complete a questionnaire for any given county for any given week(as there are over 3100counties in the lower 48 states). However, thisaggregationresults in loss of information about spatial trends at the county level.

NASS has received requests for crop progress and condition data at the county level and smaller domains. NASS already publishes annual county acreage and production estimates where possible, and the Census of Agriculture includes county datadescribing many aspects of US agriculture. The push for more agricultural estimates at smaller domains coincides with the increased availability of high resolution geospatial datasets containing information about vegetation, soil, weather, and climate. Greater availability of county level data such as crop progress and condition would be aboon to agricultural researchers who are integrating these varying data types, and to market participants with an interest in forecasting crop production.

The purpose of this discussion is to explore ways to provide reliable estimates of crop progress and condition at the county level. Twogeospatialapproaches will be discussed. First, ordinary kriging is considered. Kriging is a spatial interpolation method that predicts variable values in unobserved locations based on variable values in observed locations. Kriging relies on spatial autocorrelation to fit a model which is then applied to the existing data to create a prediction surface. For this discussion, GIS software is used to derive approximate county centroid locations based on county boundary data. The compositional, county-level crop progress and condition categories arereinterpreted into continuous, range-limited variables. This numeric data is then attached to the county centroids, creating a dataset of observedpoints. Kriging is then used to create the prediction surface covering the lower 48 states. This does incorporateexternal information besidesthe spatial relationship between observations.

Next, a co-kriging approach is considered which includes auxiliary information in the form of the Normalized Difference Vegetation index (NDVI). Cokriging is a multivariate extension of kriging which uses cross-correlation between multiple input variables to predict a surface representing a single, target variable. NDVI is a representation of the vigor of vegetation, and is usually derived from multispectral reflectance data collected bysatellites. Operating under the assumption the plant vigor and crop progressand conditionare related, their cross-correlation may be useful for informing a co-kriging model. In both the kriging and co-kriging approaches, the resulting prediction surfaces can be summarized to the county level.

We have discussed novel approachesfor representing crop progress and condition asgeostatistical surfaces. While crop progress and condition data has previously been available at the state and national level, kriging and co-kriging methods provide an opportunity to provide totally synthetic data at finer resolutions than the US county level. Additional work is needed to solidify these approaches, particularly to devise good practices for characterizing the uncertainty of resulting county estimates. Other geospatial and non-geospatial methods for deriving county estimates of crop progress and condition may be considered for future work.

4. Small Area Estimation in Government Surveys (U.S. Census Bureau)

[AS 7, (page 9)]

TRAN, **Bac**,*US Census Bureau*

The Annual Survey of Public Employment & Payroll (ASPEP), conducted by the U.S. Census Bureau, provides statistics on the number of federal, state, and local government civilian employees and their gross payrolls. The universe of ASPEP is about 90,000+ state and local government units. Every five years (year ending with 2 and 7, e.g., (2007 and 2012) Census Bureau conducts a Census of Governments, Survey of Public Employment & Payroll (CoG:E). Between censuses, Census Bureau conducts the ASPEP, a nationwide sample survey covering all state and local governments in the United States. The ASPEPsurvey is 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 (e.g., government functions such as elementary and secondary education, higher education, police protection, financial administration, judicial and legal, etc., at the national level, and states aggregates of all function codes). However, it is also required to estimate the parameters for individual function codes within each state. This requirement prompted us to develop a methodology that employs Small Area Estimation (SAE) using unit-level covariate models in order to borrow strength from previous census data as an alternative to collecting expensive additional data for small cells. In this paper we summarize our applications of the estimators over the years for the ASPEP. The outlier treatments (Trinh & Tran, JSM 2016 & 2017) will also be discussed in this research to improve the quality of the estimates. The data we used in this research are the two CoG:E of the years 2007 and 2012.

5. Small area estimation and issues for measures related to cancer surveillance

[AS 1, (page 3)]

LIU, **Benmei**,*National Cancer Institute, National Institutes of Health*

TBA

6 . Overview of challenges in cancer control research with small population

[AS 1, (page 3)]

BERRIGAN, **David**,*National Cancer Institute*

TBA

7. Big data for small area estimation

[AS 8, (page 9)]

NEWHOUSE, **David**,*The World Bank*
MASAKI, Takaaki,
SILWAL, Ani Rudra,
BEDADA, Adane,
CORRAL, Paul,
ENGSTORM, Ryan,
NGUYEN, Minh,

This paper uses household survey, remote sensing, and census data from Sri Lanka and Tanzania to evaluate the efficiency and accuracy of small area estimates of non-monetary poverty rates. The estimates are generated by employing village-level remote sensing indicators in a household-level Empirical Bayes Linear Unbiased Predictor (EBLUP) model with anormalized welfare measure. Standard errors are estimated using a parametric bootstrap approach. In both countries, the estimates are highly correlated with non-monetary poverty directly calculated from the full census and the gain in precision is comparable to nearly quadrupling the size of the sample. The estimates are substantially more precise than those obtained from an area-level Fay-Herriot model. Standard errors are underestimated, but coverage rates are comparable to direct survey estimates that assume independent disturbances across clusters. The results demonstrate that combining household survey data with village-level remote sensing indicators can greatly increase the precision of non-monetary poverty estimates at modest cost.

8 . DISAGGREGATING AVERAGE HOURY EARNINGS ONDISTRICT LEVEL IN INDONESIA USING SMALL AREA ESTIMATION

[AS 6, (page 7)]

NIKEN LARASATI, **Dhiar**,*BPS-Statistics Indonesia*

One of Sustainable Development Program goals is to promote sustained, inclusive and sustainable economic growth, full and productive employmentand decent work for all. One of its targets is to achieve full and productive employment and decent work for all women and men,including for persons with disabilities, and equal pay for work of equal value in 2030.

Indonesia is trying to reach the target by providing decent workfor all population. It should be supported by appropriate indicators that help policy maker, especially the government, to make policies that are right on target. One of them is to provide average hourly earnings(AHE)indicator to the smallest level.

As a preliminary study, weestimatedthe AHEindicator disaggregated by sex and disability status for 2017. The Small Area Estimation (SAE) methodis applied.The main data source used wasthe National Labor Force Survey (SAKERNAS) in August 2017. Population Census 2010 (SP2010) and Village Potential Data Collection (PODES) 2014 were used as sources of auxiliary variables. Proportion of people graduated fromhigh school or above, proportion of literate people, and proportion of people working on agriculture sector were auxiliary variables from SAKERNAS 2017, whereas proportion of village with production skills improvement program, proportion of village with sales skills improvement program, and proportion of village with strengthening social institutions program were auxiliay variables from PODES

2014.

The result showed that,based on AIC scores, proportion of village with sales skill improvement program and proportion of literate people were not selected in AHE model for all disaggregation categories in districts level. Western Indonesia districts (Sumatera and Java) have the lowest AHE, while Eastern Indonesia districts (Sulawesi, Maluku, Papua) have the highest AHE.

9. Computational methods in Bayesian disease mapping: an application to Chronic Lower Respiratory Disease (CLRD) mortality in the U.S., 2000-2017

[AS 2, (page 4)]

KHAN, **Diba**,*Division of Research and Methodology, National Center for Health Statistics at the Centers for Disease control and Prevention*
PLEIS, John R., *National Center for Health Statistics, U.S. Centers for Disease Control and Prevention*
ARIAS, Elizabeth, *Division of Vital Statistics, National Center for Health Statistics at the Centers for Disease control and Prevention*

Several computational methods exist in literature for Bayesian disease mapping on small geographic scales. Hierarchical Bayesian space time models are often used along with a set of associated covariates to describe geographic variation and trends for small areas.Due to the methodological and computational complexity in dealing with large datasets (in space and time) several methods have been proposed in literature. Posterior sampling methods such as Markov Chain Monte Carlo (MCMC methods) which are computationally intensive and time consuming have been primarily used to approximate the posterior distribution of model parameters in Bayesian Hierarchical models.With the advent of the Integrated Nested Laplace Approximation (INLA) technique, the posterior approximation is achieved by applying numerical integrations for fixed effects and Laplace integral approximation to the random effects. Appropriate selection of method and software in analyzing spatiotemporal variations in health outcomes results in less computation time, flexibility in implementation of appropriate models and precise estimates. In this study we propose to investigate and compare small area models often used in Bayesian disease mapping in the software NIMBLE(Numerical Inference for Statistical Models using Bayesian and Likelihood Estimation)and INLA via the software R for county level Chronic Lower Respiratory Disease (CLRD) mortalitydata in the U.S., 2000-2017.

10. Small Area Estimation in the Census Bureau 2016 Determinations for Multiple Language Ballots under the Voting Rights Act

[AS 7, (page 9)]

SLUD, **Eric**,*University of Maryland College Park and US Census Bureau*

Section 203(b) of the Voting Rights Act of 1965 (amended in 1982 and 2006) requires the Director of the Census Bureau every 5 years to determine states and political subdivisions that must make voting materials available in languages other than English. The Determinations follow fixed rules involving the sizes and proportions of nested subgroups of the voting age population who are citizens, limited English-proficienct, and illiterate. These populations and proportions must be estimated, from the most current available American Community Survey (ACS) and decennial census data, in approximately 8000 jurisdictions, 570 American Indian and Alaska Native Areas (AIA/ANAs), and 12 Alaska Native Regional Corporations (ANRCs), for 68 Language Minority Groups (LMGs). Many of the required estimates concern geographic areas with extremely small subpopulations. As a result, the sampling variability of direct estimates from the 5-year ACS data is large.

Special tabulations of weighted sample survey ”direct estimates” of state, jurisdiction, and AIA/ANA voting-age populations cross-classified by citizenship, limited English proficiency, illiteracy, and LMG are produced from American Community Survey 5-year data. These tabulations could be used to create direct estimates of the elements of the Determinations, as was done prior to 2011, along with (fairly unreliable) estimated variances. Since 2011, the Census Bureau has estimated the populations and subpopulation proportions needed for VRA Sec. 203(b) Determinations using statistical models and ”small area estimation” techniques. This talk describes some of the features of the data and methodology used in 2016.

Models were fitted separately within each LMG, and separately for jurisdictions and for American Indian and Alaska Native Areas, based on the ACS 5-year 2010-2014 data. The form of model is Dirichlet-multinomial regression, a random-effects generalization of logistic regression, for the incidence of citizenship, LEP and illiteracy among voting-age persons within a domain (the intersection of LMG with jurisdiction or AIA/ANA). Under this model, the observed data for the voting-age ACS sampled persons within each domain, conditional on covariates and on the random-effect parameters, are modeled as independent multinomial trials. Covariates explored included educational level, age, proportion foreign born, and average time in US, separately calculated for all adults in the domain. The parameters of these models are shared across geographic areas within each fixed LMG. Models are estimated using maximum likelihood. Subpopulation shares are predicted as weighted combinations of the direct ACS survey-weighted ratio estimates and those from the regression model. The weights heavily favor the direct estimators in large-population domains, where direct estimates are precise, and give substantial weight to model-predicted values when the direct estimates are unstable.

The models considered were evaluated extensively using ACS 5-year 2008-2012 data in the same way that the selected model ultimately employed the ACS 5-year 2010-2014 data. The models had generally the same form in the different LMGs, but fewer (or no) covariates were used when almost all domains in the LMG had tiny sample sizes. In the smallest LMGs, no model of the selected form could be fitted with parameters in reasonable ranges, in which case direct survey-weighted ACS estimates were used.

Mean-squared errors of predicted populations and proportions were estimated by a novel hybrid technique combining direct balanced replications with model-based estimates and generally were much smaller than the estimated direct-method variance.

This talk is based on joint work with Robert Ashmead and Patrick Joyce of the US Census Bureau. The data described here and the extended technical summary can be found at https://www.census.gov/programs-surveys/decennial-census/technical-documentation/voting-rights-determination-methodology.html

11. Challenges with Statistical Studies of Small Populations

[Special Invited Talk 2, (page 5)]

KALTON, **Graham**,*Westat*

Recently there has been a great expansion in the demand for statistical studies of small populations across a range of different disciplines. The term small population covers a diverse set of populations including, for example, persons residing in small ruralareas, children of single mothers living in poverty, small racial minorities, rape victims, recent immigrants, illegal immigrants, illegal drug users, men who have sex with men, street prostitutes, and the homeless. Standard probability sampling methods are impractical for many of these populations, and particularly so for hidden populations with amembershipthat is sensitive. This presentation will review the challenges in trying to applyprobability sampling in studies of small populations. It will outline some of the alternative designs that are being usedand it will review the issues involved in making valid statisticalinferences about small populationsfrom samplesgenerated by suchdesigns. It will considerthe use of data collected from such samples toestimatethe size of the population, to estimatecharacteristics of the members of the population, and to estimatedifferences in characteristics between different subgroups of the population.

12. County-level Estimates of Cancer Screening and Cancer Risk Factors in the United States: Combining Information for the National Health Interview Survey (NHIS) and the Behavioral Risk Factors Surveillance Survey(BRFSS), 2011 2016

[AS 2, (page 4)]

PLEIS, **John R.**,*National Center for Health Statistics, U.S. Centers for Disease Control and Prevention*
LIU, Benmei, *National Cancer Institute, National Institutes of Health*
FEUER, Eric, *National Cancer Institute, National Institutes of Health*
HE, Yulei, *National Center for Health Statistics, U.S. Centers for Disease Control and Prevention*
PARSONS, Van, *National Center for Health Statistics, U.S. Centers for Disease Control and Prevention*
CAI, Bill, *National Center for Health Statistics, U.S. Centers for Disease Control and Prevention*
TOWN, Machell, *U.S. Centers for Disease Control and Prevention*

Small-area estimates (e.g., county-level) for measures of health continue to be of interest for public health researchers and policy makers. Raghunathan et al. (2007) previously published methodological work for producing county-level estimates of risk factors and cancer screening rates bycombining information from the BRFSS and NHIS for the time periods 1997 -2000. For this previous work, BRFSS conducted interviews via landline telephone and the NHIS utilized face-to-face household interviews. However, since 2011, BRFSS has also used acell phone frame which was not part of the previous developments. As a result of this difference and to provide updated estimates, researchers from the National Cancer Institute (NCI), BRFSS, and the National Center for Health Statistics (NCHS) have beenworking on updating the previous methodology as well as extending the work to more outcomes. For this presentation, we will illustrate some of the methodological challenges faced including the use truncated bivariate distributions. For illustration, models will be presented using PROC MCMC in SAS.

13 . Recent Work in Small Domain Modeling for the Current Employment Statistics Survey

[AS 8, (page 10)]

GERSHUNSKAYA, **Julie**,*Bureau of Labor Statistics*

The U.S. Bureau of Labor Statistics Current Employment Statistics (CES) survey publishes monthly estimates of employment and other major indicators of the U.S. economy at various national and state levels, as well as for numerous domains defined by intersection of industry and geography. At a finer level of detail, where the sample is sparse, small area models areused to improve the estimates. I will briefly review the CES estimation setup, introduce the classical Fay-Herriot model and describe several modifications that address some of the problems encountered with the classical formulation.

14. Administrative Data and Evidence-Based Policymaking

[Special Invited Talk 4, (page 9)]

ABRAHAM, **Katherine**,*University of Maryland College Park*

Making effective use of government resources requires evidence both on social and economic conditions and on the effects of specific policies. The Foundations of Evidence-Based Policymaking Act, passed by the U.S. Congress at the end of 2018 in response to the recommendations of the Commission on Evidence-Based Policymaking, has created new opportunities to use administrative data the government already holds to provide evidence that policy officials need. This talk will illustrate some of the many ways in which administrative data can be helpful for generating policy-relevant evidence and discuss the challenges that will need to be confronted to achieve the Commissions vision of “a future in which rigorous evidence is created efficiently, as a routine part of government operations, and used to construct effective public policy.

15 . Modeling county-level drug overdose death rates

[AS 2, (page 4)]

ROSSEN, **Lauren**,*Division of Research and Methodology,National Center for Health Statistics at the Centers for Disease control and Prevention*
KHAN, Diba, *Division of Research and Methodology, National Center for Health Statistics at the Centers for Disease control and Prevention*
HE, Yulei, *Division of Research and Methodology,National Center for Health Statistics at the Centers for Disease control and Prevention*

Drug overdose mortality has been increasing rapidly over the pastseveral decades and there is a critical need for timely and accurate local-level data to inform efforts to address this epidemic. Data on drug overdose mortality are available from several online dissemination toolssuch as CDC WONDER, but rates are suppressed for any geographic areas with fewer than 20 deaths. These data suppression rules result in a lack of information for most rural counties across the US. To overcome these limitations, methods to produce county-level estimates of drug overdose mortality using data from the National Vital Statistics System (NVSS) have been implemented, evaluated, and published in several journal articlesand data visualizations.

Several extensions to these models are currently being explored. These extensions include: 1) modeling age-specific rates to generate age-adjusted estimates; 2) exploring whether these methods can be applied to provisional mortality data to facilitate the publication of more timely informationon drug overdose mortality; 3) implementing methodsto detect spatial and temporal outliers (i.e., outbreaks); and,4) determiningthe feasibility of accounting for missing information on specific drugs or drug classes on the death certificate in order to produce county-level estimates of death rates associated with specific drugs (e.g., opioids).

This presentation will provide an overview ofthe current approach for modeling county-level drug overdose death rates and associated trends, and then focus on a discussion of various problems and future extensions to these methods.

16. Modeling Approach for Cancer Incidence Completeness Measure

[AS 1, (page 3)]

ZHU, **Li**,*National Cancer Institute*

TBA

17. Preserving Acreage Relationships in Small Area Agricultural Models

[AS 3, (page 5)]

CHEN, **Lu**,*USDA NASS*
B. CRUZE, Nathan, *USDA NASS*
NANDRAM, Balgobin, *Worcester Polytechnic Institute*

When the US Department of Agricultures (USDA) National Agricultural Statistics Services (NASS) publishes county-level crop totals, certain inequality constraints and benchmarking must be satisfied. For example, the county-level estimates of planted acreages should "cover" the corresponding available administrative data while also satisfying benchmarking constraints so that county-level estimates add up to the state-level estimates. The acreages of crop area harvested within any given geographic boundary (e.g. county or state) should not exceed the acreages that were planted. In addition, the published estimates of failed acreage (planted acreage -harvested acreage) should "cover" the corresponding administrative data of failed acreage.

NASS conducts the County Agricultural Production Survey (CAPS) annually to provide accurate county-level estimates of planted acreage, harvested acreage, yield and production. The current method of producing these official estimates is an expert assessment incorporating multiple sources of information, including the CAPS estimates and other auxiliary information whenever it is available. Key administrative data sources are from two other USDA agencies: the Farm Service Agency (FSA) and the Risk Management Agency (RMA). Recent studies and papers have shown that the hierarchical Bayesian small area models can incorporate auxiliary source of data to improve county-level survey estimation of crop totals with measures of uncertainty. While Cruze et al. (2019) identified satisfaction of these constraints as a necessity, attempts to do so have not been addressed in previous literature.

In this talk, the challenges and methodologiesfor addressing constraintestimates problemsin small area agricultural modelsare discussed.The strategies are to modelplanted and harvested acreage estimatesseparately while preserving important relationships. First, a sub-area Bayesian hierarchical model with inequality constraints is proposed for the planted acreage estimates. The proposed model integrates the CAPS data with other administrative data to produce reliable county-level estimates that satisfy important relationships, along with associated measures of uncertainty. Inequality constraints add complexity to fitting the model and present a computational challenge to a full Bayesian approach, so improvedperformance is needed to justify the additional computational burden. The external ratio benchmarking is applied to the county-level estimates so that they add up to state targets.

Second, as an alternative to modeling harvested acreages directly, we proposed modeling the proportions of crop area harvested, again using the hierarchical Bayesian model with inequality constraints. The proportion estimates of harvested acreages can promise the relationship between planted and harvested acreages. The inequality constraints help to preserve the failed acreage relationship. We comment on the challenges of specifying a fully joint model and comment on the effects on precision of estimates when conditioning harvested area on planted area, and incorporating other publication requirements.

The development and implementation of two models for estimating planted acreages and harvested acreages respectively are illustrated based on 2014 corn in Illinois, one of the largest producers of corn in US. The example shows the difficulties encountered while preserving all relationships and several challenges that NASS faces in order to adopt model-based estimates as the official statistics. On the other hand, to evaluate the inclusion of all inequality constraints, the proposed models and the models without inequality constraints were compared via several modeling diagnostics, as well as external checks with published estimates. The performance of the model with inequality constraints illustrates the improvement of county-level estimates in accuracy and precision while preserving required relationships.

18 . Research Challenges in Model-Based Estimation ofAgricultural Cash Rental Rates

[AS 3, (page 5)]

E. BELLOW, **Michael**,*USDA NASS*

Estimates of agricultural cash rental rates at different geographic levels are used by landowners and farm operators to formulate rental agreements and by policymakersto administer state and federal programs. The USDAs National Agricultural Statistics Service (NASS) began publishing estimates of average cash rental rates at the state level in 1997. In 2009, the agency fully implemented an annual Cash Rents Survey (CRS) to estimate average cash rental rates at the state, agricultural statistics district and county level in three land use categories: 1) non-irrigated cropland, 2) irrigated cropland, and 3) pastureland. The design is stratified systematic sampling where the sampling units are farm operators identified as cash renters in the NASS list frame from the previous CRS and the Census of Agriculture(for all U.S. states except Alaska).An experienced team of NASS regional field office statisticians (coordinated by the agencys headquarters in Washington, DC) establish official county-level estimates of cash rental ratesbased on survey responses and other information, with benchmarking applied to ensure consistency with state and district-level estimates.

Since realized survey sample sizes for most counties are too small for reliable direct estimation, NASS was motivated to explore the potential of model-based small area estimation approaches that borrow strength from auxiliary data sources. After initial consideration of bivariate modeling, a procedure known as the Berg-Cecere-Ghosh (BCG) method was developed which involvesseparate univariate area level models for the average and difference of rates between two survey years in a Fay-Herriot type formulation. The covariate is a county-level single number index obtained by combining several available supplementarydata sources. The BCG method respects the CRS sample design by incorporating the direct estimators of means and sampling variances. An assumption of orthogonality between the residuals of thetwo submodels simplifies the computation of estimates and mean-squared errors. The intended deliverables are the county-level cash rental rate estimates and coefficients of variation for the three land use practices in NASSs 49-state cash rents estimation program.Development and refinement of various aspects of the BCG methodology (with operational considerations always the foremost priority) has been a focus of research since the model was first proposed. This presentation deals specifically with the following two subtopics: 1) impact of skipping a year of survey data on estimation effectiveness, and 2) benchmarking options.

The BCG method was first applied to operational county-level estimation in 2013. With the 2014 Farm Bill specifying thatthe survey be conducted ”no less frequently than every other year”, there was a need to evaluate how this estimator would perform with CRS data from two years apart (with the survey skipped in the intervening year). In an empirical study, BCG estimates of 2014 cash rental rates were compared under two scenarios: 1) consecutiveyears of CRS data (2013 and 2014) as used in production,and 2) non-consecutive years (2012 and 2014). Some noticeable departures from the assumption of uncorrelated residuals for the average and difference were observed,while in terms of estimatorperformance scenario 1 trackedofficial NASS estimates more closely overall than scenario 2 for non-irrigated and irrigated cropland but less closely for pastureland. Subsequent to this study, a new Farm Bill mandated that the survey be conducted on an annual basisin the foreseeable future.

An important issue is efficient benchmarking to ensure consistency among state, district and county-level cash rents estimates. NASS currently usesthe Ghosh-Steorts(GS)method which involves two-stage difference benchmarking with a single weighted squared-error loss function combining county-level and district-level loss without making any specific distributional assumptions. A study was conducted comparing the GS method with two alternative ratio-based benchmarking methods known as RB1(employing a single state-to-county adjustment factor) and RB2(employing separate state-to-district and district-to-county adjustment factors)using CRS data from 2013-14 in states having complete data (in all counties) for either non-irrigated cropland or pastureland. Exploratory analyses suggested that the relative effectiveness of the three benchmarking methods depends on the land use practice. Although GS appeared least prone to excessive adjustments, RB2 tended to have lower (magnitude) correlations between change in estimate and CRS sample size than either GS or RB1.

We conclude with a discussion on the current status of implementation of the BCG model in operational cash rents estimation and possible directions for future research.

19. TBA

[AS 5, (page 7)]

PRATESI, **Monica**,

20. A Preface on Small Area Analytic Approaches for Crop and Economic Estimatesat the United States Department of Agriculture National Agricultural Statistics Service(USDA NASS)

[AS 3, (page 5)]

B. CRUZE, **Nathan**,*USDA NASS*
J. YOUNG, Linda,

This session of contributed talks fromsubject matterexperts at USDA NASS will highlight approaches to small area estimation and analysis focused on applications in the agencyscrop and economic estimates programs. The collection of presentations showcases research transitioning into production as well as research still in progress. This talk serves as a preface describing three particular NASS data products:

- *Rosales on Crop Progress and Condition*–Crops are living things. The planting and harvesting of a crop occur at two different points in time, and a number of growth and development stages (phenological stages) can be impacted byconditions in between those two events, determining the crop that remains to be harvested at the end of the season as well as the total volume of output that isharvested. Published NASS crop progress statistics describe, at national and state levels, the proportion of the crop that is transitioning into each of these milestone growth stages. Proportions representing subjective crop condition ratingsare published in five exhaustive categoriesthese data are compositional in nature, summing to 100% of the crop of interest. In this talk, early research on kriging-based approaches for deriving detail at county or finer levels from the same survey data are explored.

- *Chen on Constrained Estimation of Crop Areas*–NASS publishes estimates of planted area, harvested area, total production, and yieldat the national, state, and county level. Coherence is a necessary property of the official county-level crop estimates. Consequently, official statistics of yield must be equal to the ratio of estimated production to estimated harvested areaat all geographic levels. Furthermore, estimates of harvested area (measured in acres in the US) must not exceed planted area at any level of aggregation. Importantrelationships with administrative data are expected to be satisfied. Administrative data collected byother USDA agencies help inform lower bounds on crop acreage totals. That is, they help NASS statisticians determine a least amount of activity known to have taken place in the county. While unconstrained acreage models show remarkable reductions in the estimated uncertainty surroundingthe point estimate, the resulting point estimates sometimes imply that large quantities of administrative data have shifted across county lines.

- *Bellow on the Berg-Cecere-Ghosh Cash Rental Rates Model*–Land rental agreements between landlords and tenants exist in the agricultural sector. As opposed to a share rent agreement(where the landlord is entitled to a proportion of the proceeds of the tenants realized production), cash rental agreementsentail a lease wholly on a cash per acre basis.Both types of agreements are different strategies used by tenants and landlords for managing risk, but only the latter is in scope of NASSs Cash Rents Survey, which is designed to help support county-levelestimates of cash rental rates by irrigated, nonirrigated, and pastureland practices.This talk describes a methodology developed to provide more reliable estimates of cash rental rates for all three land-use practices.

In this overview talk, the motivations for providing more detailed estimates are described, and brief statements of the problems to be addressed and the required properties of the official statistics are discussed. A synopsis of available data inputs for each program is provided, and more detail on each problem is provided in the subsequent talks.

21. Small Area Estimation using Fay-Herriot model: Household poverty rate for districtlevel in Vietnam

[AS 6, (page 7)]

THI THANH TAM, **Nguyen**,*General Statistics Office of Vietnam*

General Statistics Office of Vietnam produces statistical figures on poverty annually forprovinces. However, to end poverty in all its form everywhere, data should be disaggregated by as smaller population groupas good.

For the study, we applied Small Area Estimation(SAE)technique to estimate poverty rate for district level. We used two datasets: (1) Vietnam Household Living Standard Survey(VHLSS) in 2012, and (2) Population Census in 2009.While Population Census data is representative for district level, VHLSS sample size is designed to producepoverty rate at provincial level. Withparameterscalculated from these two datasets, usingFay-Herriot modelfor small area estimation,we preliminarily estimated-poverty rate for district level.SAEtechniquehelps us estimatepoverty ratefor more detail disaggregation-based on available data source. It is important for policy makers to provide with effective policies and programson poverty reduction.

22. TBA

[AS 5, (page 7)]

MUNNICH, **Ralf**,

23. Issues concerning direct state-level estimation on the National Health Interview Survey

[AS 2, (page 4)]

COHEN, **Robin**,*National Center for Health Statistics at the Centers for Disease control and Prevention*

For thepast fifteen years there has been much interest in generating state-level estimates from the National Health Interview Survey (NHIS). Since 2004, NHIS has generated health insurance coverageestimatesfor selected states. The methodology used to generate state-level estimates was developed more than 15 years ago and is outlined in a report by Cohen and Makuc (https://www.cdc.gov/nchs/data/nhsr/nhsr001.pdf). This report discusses the criteria used for a states eligibility for a state-level estimate as well as methodology for adjusting standard errors of included estimates to adjust for bias. However, the NHIS survey design has changed since the publication of this report.Based on thischange for the2016NHIS, threemainissues areaddressed. Firstly, should the standard error adjustment methodology outlined previouslybe modified? Secondly, should the criteria for inclusion of state-level direct estimatesbe refined or modified? And lastly, does the methodology for assessing reliability of an estimate need to differ from that as outlined in National Center for Health Statistics Data Presentation Standards for Proportions (https://www.cdc.gov/nchs/data/series/sr_02/sr02_175.pdf) The NHIS is a cross-sectional household interview surveythat has monitored the health of the Nation since 1957. The NHIS is a multipurpose health survey, based on a complex survey design,of the U.S. civilian noninstitutionalized population and conducted continuously throughout the year by the National Center for Health Statistics. NHIS consists of both a core set of questions that remain relatively unchanged from year to year as well as rotating questions that are not asked every year. Each year an estimated 27,000 adults and 9,000 child interviews are conducted.NHIS sample is drawn from each state and the District of Columbia. However, the NHIS sample is too small to provide state-level data with acceptable precision for each state.

24 . Can a Poverty Map Based on Remote Sensing Data Replicate One Based on Census Data? An Assessment for Malawi

[AS 8, (page 9)]

VAN DER WEIDE, **Roy**,*The World Bank*

We assess the reliability of poverty maps derived from remote-sensing data. Employing data for Malawi, we first obtain small area estimates of poverty using the widely implemented small-area estimation approach introduced by Elbers, Lanjouw and Lanjouw, 2003). This combines the Malawi household expenditure survey from 2010/11 with unit record population census data from 2008. We then ignore the population census data and obtain a second poverty map for Malawi by combining the survey data with predictors of poverty derived from remote sensing data. We thereby allow for a clean comparison between the two poverty maps; one obtained with and the other without population census data. Our findings are encouraging -although that assessment does depend somewhat on the evaluation criteria employed. The two approaches reveal the same patterns in the geography of poverty in Malawi; the statistical correlation between the two different small area estimates of poverty is above 90 percent. However, there are instances where the two approaches obtain markedly different estimates of poverty. We conclude that poverty maps obtained using remote sensing data do well when the decision maker is interested in comparisons of poverty between assemblies of areas. However, the approach may be less reliable when the focus is on estimates for specific small areas.

25. Solving Global Socio-Economic Issues with Humanoid AI Engine

[AS 9, (page 11)]

MUKHERJEE, **Sarbojit**,*Baanda Inc, Los Angeles*

Experts predict that in ten years, the global population will exceed 10 billion people and 65

Its no wonder that human interconnection activities such as shared economy and co-living are very appealing these days. However, such togetherness is only meaningful when there is harmony among participants. Baanda has created a humanoid AI driven technology solution to help people find those with whom they are likely to have meaningful interactions. Baanda will serve as a catalyst for trust among strangers and help them cooperate. It will improve the wellbeing of individuals, society, reduce energy consumption, and thus aid in the reduction of manmade carbon emissions.

Cooperation hinges on trust. Baanda fast-tracks trust in order to build emotional harmony and mutual chemistry among participants. When you find your people and youre doing things together, then life is good. The focus of this workshop will be around the humanoid AI engine and its components.

The humanoid AI engine has three main interrelated components. They are, a) Understanding an individual b) Forging immutable and dynamic agreements among strangers, and c) Calculating DCCS (Dynamic Co-operation Chemistry Score).

Though the engine uses aspects of computer science, computational psychology, statistics, economics, and social science, Baanda puts them together in a box as tools necessary for building the solution. The core of the AI engine relies on a commonsense variation of Causal Inference technique and converts a limitation of machines to an opportunity.

All lifeforms, including humans, explore life via feelings. In standard Causal Inference, we need to find the cause event of some outcome. In Baanda, we find the causal feeling (not event) of some outcome. Also, machines are limited by lack of feelings, but Baanda uses this limitation to its advantage and neutralizes human biases.

In this talk, while we will explore the problem space, the overall solution, the application and technical architecture at a high level, the focus will be on how we get feeling-data, convert into uniform scaled digital measurable data, and the process data in the core engine under the computational architecture.

26. Data-driven research for transportation

[AS 9, (page 11)]

ABBAR, **Sofiane**,*Qatar Computing Research Institute QCRI Social Computing*

According to the last report of the United Nations on development, two thirds of the whole worlds population will live in urban areas by 2050. Doha, one of the fastest growing cities in the world, with a population that almost doubled in the last 10 years is a good example that showcases the several challenges inherent to this massive urbanization, including urban mobility, traffic congestion, infrastructure expansion, air pollution, and reachability. In this talk, we will cover some applied data-driven initiativesto tackle these challenges using big urban data, and cutting-edge ML and AI technologies. We will also show how some of our solutions are benefiting local stakeholders including Ministry of Transport and Communication, taxi and delivery companies, as well as policy making groups.

27 . Short-term Demand Forecasting on the Quebec Power Grid: Challenges Ahead

[AS 4, (page 6)]

DELLACHERIE, **Stéphane**,*Hydro-Québec TransÉnergie, Montréal (Québec), Canada*
ZINFLOU, Arnaud, *IREQ, Science des données and calcul haute performance, Varennes, Canada*

In order to balance supply and demand at all time, Hydro-Québec (HQ) must produce at all time the electrical load forecast on the Qubec grid as accurate as possible. To do this, HQ has developed expertise and reliable internal tools to estimate this load with an excellent level of accuracy.Nevertheless, the recent changes in society (teleworking, variable rates, etc.) and in the future (transport electrification, behind the meter production, storage, smart grids, active role of the consumer, etc.) are current and incoming challenges for the parametric forecasts HQ developed and exploits, since the load will undoubtedly be more difficult to modelize with no clear physical phenomena and measures to explain it.Additionally, the large number of manually adjusted parameters greatly complicates the maintenance and evolution of these models, which ultimately make them unsuitable for capturing the changes induce by the energy transition.It is therefore necessary to get prepared for this energy transition by testing potentially more flexibleapproaches, that can a prioricapture rapid and deep behavior changes and that also beable to predict the networks load at the consumer level (bottom-up approach).Artificial intelligence techniques, particularly in deep learning, are possible complementary or alternative approachesto parametric models. They can give useful and additional insights as they arenot built from a more or less precise understanding of load relative physical phenomena (such as the use of a heating versus an air conditioner) and they are moresuitable to deal with verylarge input signals.After introducing the main trends and tools in production in the field of load forecastingin Quebec, this talkaims to show preliminary results using deep learning approachesto predict the electrical load on the network. Some mathematical results highlighting the interest of the neural approach in the context of the bottom-up (egdensity result in L_{inf} [1], optimality of the approximation error [2]) will also be recalled. Finally, we will conclude the presentation by sharing other forecasting challenges (egsolar forecasting) and some computer challenges (big data management, new languages, cybersecurity, etc.).

References

G. Cybenko. Approximation by Superpositions of a Sigmoidal Function, Math. Control Signals System, 2, pp. 303-314, 1989.

A. R. Barron. Universal Approximation Bounds for Superpositions of a Sigmoidal Function, IEEE Transactions on Information Theory, 39(3), pp. 930-945, 1993.

28. Small Domain Analytics: Goals, Issues and Methods

[Special Invited Talk 1, (page 3)]

LOUIS, **Thomas A.**,*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Heath, Baltimore MD USA*

The direct estimate of an underlying feature based on a small sample size from a geographic,demographic or institutional domain may be unbiased, but its variance and mean squarederror (MSE) are high. A regression-based estimate has low variance, but its bias and MSEfor a specific domain are generally high. A weighted average of the direct and the regression-based values (i.e., shrinkage towards the regression) strikes a compromise that reduces MSE.The Bayesian formalism is very effective in optimizing the weights, and confers other benefitsincluding stabilizing estimates from non-linear models; efficient ranking; estimating the thehistogram of parameters; computing the probability of exceeding a threshold; and borrowinginformation in spatial and other contexts. Modern computing activates these models.

If the model is well-specified, the Bayesian approach is effective, however a mis-specifiedmodel can perform poorly, especially for a domain with an unstable direct estimate. Inproducing the domain-specific estimate the regression gets a big weight, but in estimatingthe regression the domain gets relatively low weight. Use of a flexible prediction model, overweighting the relatively high variance domains when estimating the regression, or using aflexible prior distributions can improve performance.

These modeling issues should not distract from careful attention to the sampling plan,confounding and other traditional issues, especially in the big data context. Against thisbackground, I highlight a subset of inferential goals and issues, and illustrate approaches viaexamples from the clinical, epidemiologic and policy arenas.

29. Casualties in road accidents: challenges and opportunities with small area estimation for linked data sets

[AS 5, (page 7)]

TUOTO, **Tiziana**,*ISTAT, Italy*

Road accidents are the leading cause of death for young adults in industrialised countries. One of the most important challenges facing the World Health Organization (WHO) is the prevention of accidents. Data available for road traffic accidents often come from different sources, each of which, if considered separately, has limits. The different nature of the data often highlights difficulties of comparability, due to different definitions. However, the integration of health and non-health data is essential for building an adequate surveillance system to guide both preventive and repressive actions. In the talk, we focus on the potentialities of the analysis based on integrated datasets, which report information on casualties due to road accidents. The data come from the linkage of two separate archives collected by the Italian National Institute of Statistics (Istat), the survey of road accidents with injuries on road accidents, provided to Istat by Police authorities, and the register on deaths and causes of death, reporting health data collected by doctors in the hospitals. In this way, with the use of probabilistic record linkage, it is possible to provide a set of integrated information for each individual victim of a road accident. The joint analysis between the data on the nature of injuries and those on the dynamics and characteristics of accidents and individuals constitutes a plus value compared to the traditional dissemination of the results from the two surveys. The integration of different data sources is in line with the common strategy put in place by several National Institute of Statistics, especially in recent years, for the enhancement of existing administrative archives, with the aim of both reducing the statistical burden on respondents and eliminating redundancies in data dissemination. The analysis on the integrated dataset is particularly interesting when refined at a very detail level, so to allow monitoring the most dangerous circumstances and planning specific preventive actions. To this purpose, small area estimation methodology provides interesting solution that highlights differences and peculiarities at geographic level as well as for other small domains. However, the analyses based on integrated data set have to take into account that results can be affected by linkage errors, i.e. false links and missed links, with a typical trade-off between these two types of errors. Statistical analyses may be compromised and the results of standard statistical techniques can be misleading, if the resulting integrated data contain a substantial proportion of false links or if a significant proportion of true links are erroneously left apart. Data integration and subsequent analyses on integrated data should be considered as part of a single statistical activity and the appropriate strategies have to be designed accordingly. However, this ideal situation is rather difficult to achieve in practice, as it is very costly in terms of time and resources. Often, the analysis on integrated data takes on a secondary user perspective; one does not have full access to the linkage key variables nor to the details or tools of the actual linkage procedure, but at most one is only provided with some non-disclosive linkage comparison data about the record linkage precision or how the records compare to each other. In this talk, we discuss some existing approaches to statistical analysis, and their respective theoretical and practical implications.

30. Spatial Sensitivity Analysis for Urban Hotspots using Cell Phone Traces

[AS 9, (page 11)]

FRIAS-MARTINEZ, **Vanessa**,

Urban hotspots can be used to model the structure of urban environments and to study or predict various aspects of urban life. An increasing interest in the analysis of urban hotspots has been triggered by the emergence of pervasive technologies that produce massive amounts of spatio-temporal data including cell phone traces (or Call Detail Records). Although hotspot analyses using cell phone traces are extensive, there is no consensus among researchers about the process followed to compute them in terms of four important methodological choices: city boundaries, spatial units, interpolation methods and hotspot variables. Using a large scale CDR dataset from Mexico, we are working on a systematic spatial sensitivity analysis of the impact that these methodological choices might have on the stability of the hotspot variables at both inter-city and intra-city levels.

31. Scalability issues for SAE methods

[AS 7, (page 9)]

BASEL, **Wesley**,*U.S. Census Bureau*

General methods in use at for published SAE estimation at government agencies primarily focus on one or two response variables, disaggregated perhaps into many age/sex/race and geographic domains. Such projects involve a narrowly targeted GLMM-style model for optimal predictions. This general method is very good for one or two response variables of key importance, but increasingly, we are being asked about the possibility of modeling larger sets of responses. This may be from one sponsor, or a combination of multiple stakeholders for the same survey. The current paradigm is not scalable past a few response fields. Big data methods, such as partitioned trees, neural nets, etc., are adept at handling very large rows and columns in the auxiliary data. But as currently understood not so much in the target matrix, unless one structures as a very high degree multinomial, which may not be feasible under given methods. Furthermore, such methods do not include estimates of the local area bias, i.e. random effects, which comprise such a large part of the optimality of SAE methods. So how do we proceed for efficient modeling of very large response vectors?

1) High dimensional multivariate GLMM. Maybe hierarchical, but it would need automated to a large degree.

2) Difference between big data classification models and SAE methods. Is it just the random effect, i.e. local area bias estimate. Corrollary: Optimality of classification models relative to GLMM theory.

3) Difference between model-assisted (design-based) vs. model-based. One of the methods used here for large scale survey improvement is multivariate imputation, which is basically design-based modeling.

4) Partial-coverage data sets. Increasingly common for us to have access to very detailed data sets (unit-level) that only cover a portion of the geographic universe. Examples include SNAP participation, TANF participation, and electronic health records. What are the methods these can be incorporated in an efficient scalable manner?

Directory

ABBAR, Sofiane
Qatar Computing Research Institute QCRI Social Computing

Speaker: AS 9, p. 11, §26, p. 25

ABRAHAM, Katherine
University of Maryland College Park

Speaker: Special Invited Talk 4, p. 9, §14, p. 20

ARIAS, Elizabeth
Division of Vital Statistics, National Center for Health Statistics at the Centers for Disease control and Prevention

Coauthor: AS 2, p. 4, §9, p. 18

B. CRUZE, Nathan
USDA NASS

Coauthor: AS 3, p. 5, §17, p. 21,
Speaker: AS 3, p. 5, §20, p. 23

BASEL, Wesley
U.S. Census Bureau

Speaker: AS 7, p. 9, §31, p. 27

BEDADA, Adane

Coauthor: AS 8, p. 9, §7, p. 17

BERRIGAN, David
National Cancer Institute

Speaker: AS 1, p. 3, §6, p. 17

CAI, Bill
National Center for Health Statistics, U.S. Centers for Disease Control and Prevention

Coauthor: AS 2, p. 4, §12, p. 19

CHEN, Lu
USDA NASS

Speaker: AS 3, p. 5, §17, p. 21

COHEN, Robin
National Center for Health Statistics at the Centers for Disease control and Prevention

Speaker: AS 2, p. 4, §23, p. 24

CORRAL, Paul

Coauthor: AS 8, p. 9, §7, p. 17

DELLACHERIE, Stéphane
Hydro-Québec TransÉnergie, Montréal (Québec), Canada

Speaker: AS 4, p. 6, §27, p. 25

E. BELLOW, Michael
USDA NASS

Speaker: AS 3, p. 5, §18, p. 21

ENGSTORM, Ryan

Coauthor: AS 8, p. 9, §7, p. 17

FEUER, Eric
National Cancer Institute, National Institutes of Health

Coauthor: AS 2, p. 4, §12, p. 19

FRIAS-MARTINEZ, Vanessa

Speaker: AS 9, p. 11, §30, p. 27

GERSHUNSKAYA, Julie
Bureau of Labor Statistics

Speaker: AS 8, p. 10, §13, p. 20

GUTIRREZ, Andrs
Statistics Division, United Nations Economic Commission forLatin America and the Caribbean (ECLAC)

Speaker: AS 6, p. 7, §2, p. 15

HE, Yulei
Division of Research and Methodology,National Center for Health Statistics at the Centers for Disease control and Prevention

Coauthor: AS 2, p. 4, §15, p. 20

HE, Yulei
National Center for Health Statistics, U.S. Centers for Disease Control and Prevention

Coauthor: AS 2, p. 4, §12, p. 19

J. YOUNG, Linda

Coauthor: AS 3, p. 5, §20, p. 23

KALTON, Graham
Westat

Speaker: Special Invited Talk 2, p. 5, §11, p. 19

KHAN, Diba
Division of Research and Methodology, National Center for Health Statistics at the Centers for Disease control and Prevention

Speaker: AS 2, p. 4, §9, p. 18,
Coauthor: AS 2, p. 4, §15, p. 20

LIU, Benmei
National Cancer Institute, National Institutes of Health

Speaker: AS 1, p. 3, §5, p. 17,
Coauthor: AS 2, p. 4, §12, p. 19

LOUIS, Thomas A.
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Heath, Baltimore MD USA
`louis@jhu.edu`
Speaker: Special Invited Talk 1, p. 3, §28, p. 26

MASAKI, Takaaki

Coauthor: AS 8, p. 9, §7, p. 17

MUKHERJEE, Sarbojit
Baanda Inc, Los Angeles

Speaker: AS 9, p. 11, §25, p. 25

MUNNICH, Ralf

Speaker: AS 5, p. 7, §22, p. 24

NANDRAM, Balgobin
Worcester Polytechnic Institute

Coauthor: AS 3, p. 5, §17, p. 21

NEWHOUSE, David
The World Bank

Speaker: AS 8, p. 9, §7, p. 17

NGUYEN, Minh

Coauthor: AS 8, p. 9, §7, p. 17

NIKEN LARASATI, Dhiar
BPS-Statistics Indonesia
dhiarniken@bps.go.id
Speaker: AS 6, p. 7, §8, p. 17

PAL, Amita
Indian Statistical Institute, Kolkata

Speaker: Special Invited Talk 5, p. 11, §1, p. 15

PARSONS, Van
National Center for Health Statistics, U.S. Centers for Disease Control and Prevention

Coauthor: AS 2, p. 4, §12, p. 19

PLEIS, John R.
National Center for Health Statistics, U.S. Centers for Disease Control and Prevention

Coauthor: AS 2, p. 4, §9, p. 18,
Speaker: AS 2, p. 4, §12, p. 19

PRATESI, Monica

Speaker: AS 5, p. 7, §19, p. 23

ROSALES, Arthur
USDA NASS

Speaker: AS 3, p. 5, §3, p. 15

ROSSEN, Lauren
Division of Research and Methodology,National Center for Health Statistics at the Centers for Disease control and Prevention

Speaker: AS 2, p. 4, §15, p. 20

SILWAL, Ani Rudra

Coauthor: AS 8, p. 9, §7, p. 17

SLUD, Eric
University of Maryland College Park and US Census Bureau

Speaker: AS 7, p. 9, §10, p. 18

THI THANH TAM, Nguyen
General Statistics Office of Vietnam

Speaker: AS 6, p. 7, §21, p. 23

TOWN, Machell
U.S. Centers for Disease Control and Prevention

Coauthor: AS 2, p. 4, §12, p. 19

TRAN, Bac
US Census Bureau
Bac.Tran@Census.gov
Speaker: AS 7, p. 9, §4, p. 16

TUOTO, Tiziana
ISTAT, Italy

Speaker: AS 5, p. 7, §29, p. 26

VAN DER WEIDE, Roy
The World Bank

Speaker: AS 8, p. 9, §24, p. 24

ZHU, Li
National Cancer Institute

Speaker: AS 1, p. 3, §16, p. 21

ZINFLOU, Arnaud
IREQ, Science des données and calcul haute performance, Varennes, Canada

Speaker: AS 4, p. 6, §27, p. 25