**Abstracts**

**A Review of the Problem of Seasonal Adjustment Variances**

*William R. Bell*

*U.S. Census Bureau*

Many economic indicators (e.g., retail sales, housing starts, labor force statistics) are obtained from monthly or quarterly sample surveys. Almost all such indicators exhibit seasonal behavior, hence, conventional survey estimation is followed by seasonal adjustment, and it is the seasonally adjusted estimates that are of primary interest. While statistical agencies routinely provide variance estimates for the direct survey estimates, this has generally not been the case for the seasonally adjusted estimates due to certain technical, conceptual, and practical difficulties. We shall review the general problem of seasonal adjustment variances and the proposed solutions, including discussion of substantial work done in this area by Danny Pfeffermann and his collaborators at the U.S. Bureau of Labor Statistics.

**Small Area Challenges in the Conservation Effects Assessment Project**

*Emily Berg,* *Iowa State University*

 The Conservation Effects Assessment Project (CEAP) is a large-scale survey that measures several types of erosion on cropland and pastureland in the United States. Estimates are desired for domains with small sample sizes, such as counties and small watersheds. Challenges include a complex sample design, skewed distributions containing zeros, and a desire for estimates of quantiles. Finding a single family of parametric models to adequately describe the distributions for the wide range of variables and geographic domains is difficult.  We present methodology for small area estimation for quantiles for CEAP.

**An Objective Prior for Hyperparameters in Normal Hierarchical Model**

*James O. Berger,*

*Duke University and East China Normal University*

Hierarchical models are the workhorse of much of Bayesian analysis, yet there is uncertainty as to which objective priors to use for hyperparmeters. Formal approaches to objective Bayesian analysis, such as the Jeffreys-rule approach or reference prior approach, are only implementable in simple hierarchical settings. It is thus common to use less formal approaches, such as utilizing formal priors from non-hierarchical models in hierarchical settings. This can be fraught with danger, however. For instance, non-hierarchical Jeffreys-rule priors for variances or covariance matrices result in improper posterior distributions if they are used at higher levels of a hierarchical model. Berger et al. (2005) approached the question of choice of hyperpriors in normal hierarchical models by looking at the frequentist notion of admissibility of resulting estimators. The motivation was that hyperpriors that are too diffuse result in inadmissible estimators, while hyperpriors that are concentrated enough result in admissible estimators. Hyperpriors that are ‘on the boundary of admissibility’ are sensible choices for objective priors, being as diffuse as possible without resulting in inadmissible procedures. The admissibility (and propriety) properties of a number of priors were considered in the paper, but no overall conclusion was reached as to a specific prior. In this paper, we complete the story and propose a particular objective prior for use in all normal hierarchical models, based on considerations of admissibility, ease of implementation (including computational considerations), and performance.

**Kernel Estimation for a Superpopulation Probability Density Function under Informative Selection**

*Daniel Bonnéry, University of Maryland*

Kernel density estimation of the probability density function (pdf) of a response variable is considered under informative selection from a finite population. The informative selection implies that the conditional pdf of a response, given that it was selected for observation, is not the same as the inferential target, which is the unconditional pdf of the response in the superpopulation. Instead, the pdf of the observations (sample pdf) is a weighted version of the superpopulation pdf of interest. Properties of the standard kernel density estimator are described under an asymptotic framework that covers a wide range of informative selection mechanisms. The theory allows for the possibility that the selection mechanism has a parametric structure. A variety of adjustments (parametric or nonparametric) to account for the informative selection are proposed, and investigated via simulation. Joint work with Jay Breidt, Colorado State University and François Coquet, IRMAR and ENSAI

**Hierarchical Bayesian multivariate Fay-Herriot model for estimating domain discontinuities in the Dutch Crime Victimization Survey**

*Jan van den Brakel*

*Statistics Netherlands and Maastricht University*

The design of long-standing repeated surveys are kept unchanged as long as possible with the purpose to construct consistent series. In a proper redesign of this kind of surveys, the old and the new design are conducted in parallel for some period of time to quantify the discontinuities which are initiated by the modifications in the survey process. Often there is only limited budget or field work capacity available to conduct the new approach in parallel with the regular survey.

In a recent redesign of the Dutch Crime Victimization Survey (CVS), the new approach was conducted in parallel with the regular approach were the sample size of the new approach equals 1/3 of the sample size of the regular approach. This hampers the application of direct estimators for population parameters under the new approach and thus also for discontinuities, even for the planned domains. In this paper the direct estimates obtained under the regular and new approach are modelled in a bivariate hierarchical Bayesian Fay Herriot model to obtain more precise predictions for domain discontinuities. This method is compared with tow alternative methods. The first alternative is a univariate Fay Herriot model where the direct estimates under the regular approach are used as covariates in a Fay Herriot model for the new approach conducted on a reduced sample size. The second alternative is to model the direct estimates for the discontinuities in a univariate model Fay Herriot model.

**Unsupervised clustering of spectra of galaxies**

*Didier Fraix-Burnet*

*Institut de Planétologie et d'Astrophysique de Grenoble (IPAG)*

Dealing with large databases of galaxy spectra is a good example of a new problematic task in astrophysics. Current and forthcoming big surveys provide millions of spectra each containing thousands of wavelengths. These spectra must be confronted with physical and chemical models. This requires a classification which is a dimensionality reduction in both the number of observations and parameters. In this talk, we present a few approaches we are experimenting on 700 000 spectra from the SDSS using parsimony models (HDDC) and discriminant clustering (Fisher-EM).

**Bayesian Spatial Clustering Method and Its Application in Radiology**

*Jing Cao, Southern Methodist University*

Kidney cancer is among the ten most common cancers in human. The dynamic contrast-enhanced MRI (DCE-MRI) takes advantage of the interaction between a contrast agent and adjacent water protons which generates brighter signals in the scan image. In this study, we propose a novel Bayesian spatial clustering method based on a mixture of multivariate normal distribution. A latent conditional regression (CAR) process is employed to account for the spatial correlation among clustering indexes. The proposed method is demonstrated to provide smoother and more accurate clustering of pixels. A simulation study and a real application example are presented.

**Analysis of longitudinal data with omitted asynchronous longitudinal covariate**

*Hongyuan Cao*

*University of Missouri*

Long term follow-up with longitudinal data is common in many medical investigations. In such studies, some longitudinal covariates can be omitted for various reasons. In cross sectional studies, coefficient estimation of a covariate is unbiased if the covariate is orthogonal to the omitted covariate. This is not true in longitudinal data analysis, where omission of time dependent covariate can lead to biased coefficient estimate even if the corresponding covariate is orthogonal to the omitted longitudinal covariate. In this article, we propose a new unbiased estimation method to accommodate omitted longitudinal covariate. In addition, if the omitted longitudinal covariate is asynchronous with the longitudinal response, we propose a two stage approach for valid statistical inference. Asymptotic properties of the proposed parameter estimates are established. Extensive simulation studies provide numerical support for the theoretical findings. We illustrate the performance of our method on a dataset from an HIV study.

**Domain Estimation Under Informative Linkage**

*Ray* *Chambers*

*University of Wollongong*

Suppose we have a finite population that can be partitioned into a set of domains with domain membership indicated by a categorical variable X. We are interested in estimating the average value of a target variable Y in each category of X (i.e. in each domain). Our available auxiliary information corresponds to these average values, but at a higher level of aggregation, corresponding to a smaller set of strata which cross the domains of interest. Both domain and stratum affiliation are available on a sampling frame that we denote by RX. The values of Y, along with stratum affiliations, are kept on a separate unavailable register RY. An independent sample is taken from each stratum in RX. These sampled records are then probabilistically linked to records on RY (e.g. via Trusted Third Party Linkage), and we use the linked values Y\* for these sampled records, along with their domain affiliations, to estimate the small area means of Y. It is known that the expected value of Y varies from domain to domain. We assume an Exchangeable Linkage Error Model for the linkage process, where the probability of correct linkage varies substantially from stratum to stratum, but is the same for all records within a stratum. When modelling linked data such as these it is standard to assume that the stochastic properties of the linking process and those of the stochastic process underpinning the population values of the response variable Y are independent of one another given population covariate information ("non-informative linkage"). In this case, these covariates correspond to domain and stratum affiliations. But what happens when this assumption fails? The mechanism that we investigate for this failure is where there is another covariate, denoted by Z, which is an indicator for "ease of linkage" and is strongly correlated with Y. In particular, we assume that Z characterises the strata, and hence is associated (directly or indirectly) with the probabilities of correct linkage. We investigate the situation where the sampling method within each stratum is biased towards more easily linked records, and also where the population values of Z are included in the covariate information on the sampling framework.

**Prediction and inference in small areas using constrained optimization**

*Ansu Chatterjee*

*University of Minnesota*

We consider an approached based on constrained optimization to simultaneously estimate the regression slope and variance components parameters, so that the constraints on the parameters are preserved in the estimates. Since this results in potentially non-smooth and non-regular properties of the small area estimates, methods for obtaining prediction accuracy and inference will be presented. We will present theoretical results and data examples based on the Fay-Herriot model.

**Clustering and classification of astronomical objects: A new paradigm in Statistics**

*Asis Kumar Chattopadhyay*

*Department of Statistics, University of Calcutta*

This collection of works involves the application of statistics to astronomy and the development of statistical methods to solve the problems related to the universe, leading us to discoveries of new astrophysical phenomena. Data collection missions like Galaxy Evolution Explorer, Kepler Space Telescope, Hubble Space Telescope and virtual archives like Sloan Digital Sky Survey, Multi-mission Archive at STSCI, NASA Extragalactic Data base (see, for example, Chattopadhyay and Chattopadhyay, 2014 ) preserve petabytes of data, which can be used for big data analyses. Usually collection of data on celestial bodies is obscured by bad weather conditions, obstruction by another celestial object or instrumental restrictions and it cannot be repeated. Hence we often get data contaminated with noise, aﬀected by outliers or sparsely distributed (see, for example, Feigelson and Babu, 2013 ). In all such situations, the usual statistical methods fail and we need to use their adaption or to introduce new methods as per requirements. To overcome such problems, there are various transformations and denoising techniques available in the literature (e.g. kernel principal component analysis (KPCA), see, Scho¨lkopf and Smola 2002; Modak et al. 2017a). Sometimes there are rare objects (e.g. supernovas, galaxies at high redshift zones), unevenly spaced data of unequal lengths (e.g. light curves on variable stars, see, Modak et al. 2017b) where classical statistical methods are only applicable when the data is interpolated to get into a form suitable for the methods. We have suggested some possible solutions under the above scenario.

**Study on Star Formation History of Nearby Galaxies: A Bayesian Approach**

*Tanuka Chattopadhyay*

*Department of Applied Mathematics, University of Calcutta, Kolkata, India,*

Star formation scenario in galaxies of various morphological types is significant in a sense that it characterizes the structure formation in the Universe. Star formation Rate (SFR) is an important index to study the above phenomenon. But direct measurement of SFRtrue (i.e., True values of SFR) is not at all possible as one has to count stars formed per year in a galaxy accurately. In this paper the star formation is investigated by Gaussian Mixture Model Based Clustering technique (GMMBC) to form the clusters of the galaxies (using a large data set of galaxies in the Local Volume (LV)). After finding the homogeneous groups in an objective way, the star formation rate (SFR) is predicted within each groups using Bayesian LASSO and Bayesian Linear-Regression Analysis techniques. We have found five homogeneous groups of galaxies having different star formation scenario. Two of them are massive , disc dominated, highly rotating, forming stars with highest rates and free from the influence of environment while other two have lower SFR, probably turbulence resisting the process of star formation. The remaining one is a group of dwarf galaxies where star formation is affected by high density environment. The SFRs in the various coherent groups have been cross validated and also compared with other model (physical) based SFR values ()SFR, say) assumed to be true representative of true SFRs derived by various authors and discussed

**A Conditional Empirical Likelihood Based Bayesian Method for Complex Survey Data**

*Sanjay Chaudhuri.*

*Department of Statistics and Applied Probability, National University of Singapore.*

In many applications, data can only be collected through a complex design. If such a design is informative, the joint distribution of variables of interest in the sample differs drastically from their joint distribution in the population. Modelling and analysing such datasets have been major a interest in statistics for several years. In frequentist paradigm, estimating equations are often used which are are not based on any likelihood and thus cannot be directly used in Bayesian paradigm. In this presentation we discuss an alternative conditional empirical likelihood based Bayesian approach to model and analyse such complex datasets. Our method is based on Pfefferman's sample likelihood. We show that the empirical sample likelihood provides an easy way to include prior information in such analysis. Furthermore, because of the use of empirical likelihood, the procedure is semiparametric and enjoys many advantages over the traditional parametric likelihoods. We shall consider various special cases where the analysis can be further simplified. This is a joint work with

Malay Ghosh and Yin Teng.

**Genomic mediation analysis across many human tissues**

*Lin Chen*

*[The University of Chicago](https://www.uchicago.edu/)*

The impact of inherited genetic variation on gene expression in humans is well-established. The majority of known expression quantitative trait loci (eQTLs) impact expression of local genes (cis-eQTLs). More research is needed to identify effects of genetic variation on distant genes (trans-eQTLs) and understand their biological mechanisms. One common trans-eQTLs mechanism is "mediation" by a local (cis) transcript. Thus, mediation analysis can be applied to genome-wide SNP and expression data in order to identify transcripts that are "cis-mediators" of trans-eQTLs, including those "cis-hubs" involved in regulation of many trans-genes. Identifying such mediators helps us understand regulatory networks and suggests biological mechanisms underlying trans-eQTLs, both of which are relevant for understanding susceptibility to complex diseases. The multitissue expression data from the Genotype-Tissue Expression (GTEx) program provides a unique opportunity to study cis-mediation across human tissue types. However, the presence of complex hidden confounding effects in biological systems can make mediation analyses challenging and prone to confounding bias, particularly when conducted among diverse samples. To address this problem, we propose a method: Genomic Mediation analysis with Adaptive Confounding adjustment (GMAC). It enables the search of a very large pool of variables, and adaptively selects potential confounding variables for each mediation test. Recently, we expand the analysis to jointly analyze multiple human tissues and the study of mediation would also help to identify trans-associations across tissue types.

**Improving timeliness and accuracy of estimates from the UK Labour Force Survey space models**

*Duncan Elliott*

*UK Office for National Statistics*

Unemployment, employment and inactivity are estimated from data collected in the Labour Force Survey (LFS). While data is collected continuously the survey design is structured in such a way as to provide quarterly estimates. These quarterly estimates are published each month as rolling quarterly estimates and have been assessed to be of sufficient quality to be designated as National Statistics. ONS also publish monthly estimates of unemployment, employment and inactivity, but these are designated Experimental Statistics due to concerns over the quality of these data. Due to the sample design of the LFS, monthly estimates of change are volatile as there is no sample overlap. A state space model can be used to develop improved estimates of monthly change, accounting for aspects of the survey design, and so providing increased accuracy and an improvement in timeliness. Additional sources of information related to these variables, such as administrative data sources could be used within such a framework to help improve timeliness further.

**Robust Small Area Estimation with Linked Data**

*Enrico Fabrizi (Universit`a Cattolica del Sacro Cuore di Piacenza)*

Data integration is becoming increasingly important in Official Statistics, due to the increasing accessibility of administrative data from different sources. This has brought new opportunities for statisticians to develop innovative methods that can cut down costs and improve the quality of estimates. Data linkage is a key data integration methodology, and many national statistics institutes now plan on using data linkage to replace survey data collection where possible. However, when probability-based methods are used to link or match records from two or more distinct data sets corresponding to the same target population linkage errors can occur , and these errors can lead to biased estimation when ignored.

In small area estimation context, data linkage can result in the values of the survey variable of interest being drawn from a different source compared with the values of the population covariates used in the small area model. In particular, records from a sample survey source may be matched to records contained in one (or even multiple) population registers. Ignoring the possible biases induced by record linkage errors can mean that small area estimates derived from unit level models fitted to these linked data may no longer have their theoretical optimality properties.

Evidence for data linkage errors can be the presence of outliers in the linked data, and so it is natural to consider outlier robust small area estimation methods when using probability-linked data. One such robust methodology is based on the use of M-quantile ensemble models for small area estimation. These models represent a semi-parametric alternative to linear mixed models. However, the mixture type outlier generation process that underpins both M-quantile models as well as more standard mixed models is not necessarily appropriate when the outlying data values are largely due to linkage errors.

In this paper we investigate how to adapt robust small area estimators based on linear mixed models and linear M-quantiles to a scenario where survey data are probabilistically linked to auxiliary information from administrative archives and linkage errors are present. We adopt a secondary analyst point view, that is we assume that not all the information on the probabilistic linkage process is available to the researcher tasked with producing the small area estimates. Our main contribution to the literature is then to develop extensions of robust small area estimators that also correct for linkage errors.

An extensive simulation exercise shows that standard outlier robust methods that overlook this correction for linkage errors can perform poorly, while those based on the corrected methods that we propose are effective in both reducing bias and down-weighting the effect of outliers. This work is joint with Ray，Chambers(University of Wollongong) and Nicola Salvati (Universit`a di Pisa)

**Objective Bayesian Analysis for Gaussian Hierarchical Models with Intrinsic Conditional Autoregressive Priors**

*Marco A. R. Ferreira*

*[Virginia Tech](http://www.vt.edu/)*

Bayesian hierarchical models are commonly used for modeling spatially correlated areal data. However, choosing appropriate prior distributions for the parameters in these models is necessary and sometimes challenging. In particular, an intrinsic conditional autoregressive (CAR) hierarchical component is often used to account for spatial association. Vague proper prior distributions have frequently been used for this type of model, but this requires the careful selection of suitable hyperparameters. In this paper, we derive several objective priors for the Gaussian hierarchical model with an intrinsic CAR component and discuss their properties. We show that the independence Jeffreys and Jeffreys-rule priors result in improper posterior distributions, while the reference prior results in a proper posterior distribution. We present results from a simulation study that compares frequentist properties of Bayesian procedures that use several competing priors, including the derived reference prior. We demonstrate that using the reference prior results in favorable coverage, interval length, and mean squared error. Finally, we illustrate our methodology with an application to 2012 housing foreclosure rates in the 88 counties of Ohio.

**Internal and External Consistency of Robust Estimates in Surveys**

*David Haziza, University of Montreal*

In surveys, it is not unusual to collect variables for which the distribution is highly skewed. This is the case of business surveys that collect economic variables whose distribution is usually highly skewed. Skewed distributions tend to generate influential values, which are those that have a drastic impact on the estimates if they were to be excluded from the sample. A number of estimation procedures has been developed in the literature to reduce the impact of influential values, including winsorisation and procedures based on the concept of conditional bias of a unit. Regardless of the procedure used to treat the influential values, one is virtually certain to face the issues of internal and external consistency. The former pertains to the relationship between variables, whereas the latter pertains to the consistency between domain-level robust estimates and the population-level robust estimate. In this presentation, we will discuss a method for ensuring both internal and external consistency. The results of a simulation study will be presented. This is the joint work with Cyril Favre Martinoz (INSEE) and Jean-François Beaumont (Statistics Canada).

**Objective Bayesian Analysis for Gaussian Hierarchical Models with Intrinsic Conditional Autoregressive Priors**

*Masayo Y. Hirose*

*Institute of Statistical Mathematics*

The two level Normal Hierarchical model has played an important role in statistical theory and application for not only small area inference but also many research fields. Recently, under the model, Hirose and Lahiri (2017) achieved multiple desirable goals simultaneously via adjusted maximum likelihood method for a model variance parameter of random effect. Their method was developed from a perspective of frequentist approach, but some users may also prefer to use Bayesian approach achieving such desirable goals simultaneously. We thus seek a relationship between their adjusted maximum likelihood method and Bayesian prior. In this paper, we obtain a unique Bayesian matching prior for achieving multiple goals simultaneously in an asymptotical sense. This is the joint work with Prof. Partha Lahiri at University of Maryland.

**Estimation of MSPE for SAE under A Spatial Linear Mixed Model**

*Jiming Jiang, University of California, Davis, USA*  
 Policy decisions regarding allocation of resources to subgroups in a population, called small areas, are based on reliable predictors of characteristics of interest. In view of this, there has been growing demand for reliable small area predictors by borrowing information from other related sources. For the latter purpose, mixed effects models have been commonly used in small area estimation (SAE) assuming independent small areas. There are many situations, however, where the small area parameters are related to the locations of the small areas. We propose spatial linear mixed models for  SAE and also obtain the corresponding mean squared prediction error (MSPE) estimators. We prove second-order unbiasedness of the proposed MSPE estimator. In our simulations, we show that our predictors and MSPE estimators perform very well in prediction accuracy as well as measure of uncertainty. An application to physician visits for Total Respiratory Morbidity conditions in Manitoba, Canada is discussed. This work is joint with Mahmoud Torabi of the University of Manitoba, Canada.

**Approximate Bayesian Inference under informative sampling**

*Jae Kwang Kim (Iowa State University)*

Statistical inference with complex survey data is challenging because the sampling design can be informative, and ignoring it can produce misleading results. Current methods of Bayesian inference under complex sampling assume that the sampling design is noninformative for the specified model. In this paper, we propose a Bayesian approach which uses the sampling distribution of a summary statistic to derive the posterior distribution of the parameters of interest. Asymptotic properties of the method are investigated. It is directly applicable to combining information from two independent surveys and to calibration estimation in survey sampling. A simulation study confirms that it can provide valid estimation under informative sampling. We apply it to a measurement error problem using data from the Korean Longitudinal Study of Aging.

This is a joint work with Zhonglei Wang and Shu Yang.

**Pseudo-population Bootstrap Methods for Imputed Survey Data**

*Christian Leger ( University of Montreal)*

Item nonresponse in surveys is usually dealt with through single imputation. Treating the imputed values as if they were observed values may lead to serious underestimation of the variance of point estimators. Two pseudo-population bootstrap approaches are used for deriving bootstrap variance estimators: the nonresponse model approach and the imputation model approach. We establish the asymptotic properties of the resulting bootstrap variance estimators for population totals and population quantiles. Results from a simulation study suggest that the proposed methods perform well in terms of relative bias and coverage probability. This is a joint work with Sixia Chen, David Haziza, and Zeinab Mashreghi.

**Multivariate Conditional Autoregressive Modelling: A Graphical Model Perspective**

*Ye Liang, Aklahoma State University*

The conditional autoregressive (CAR) model is one of the routine techniques for spatial small area estimation problems. However, it is not a trivial work to extend the univariate CAR model to multivariate models, as suggested by many authors. This paper approaches the problem from the graphical model perspective and builds joint adjacency structures with a fixed spatial graph and a random variable graph. We show that our framework can contain the Kronecker specification, the multifold specification and the coregionalized specification as its special cases. These existing specifications can be further extended to graphical versions under our framework. Algorithms are developed for both parameter estimation and graph learning.

**Objective Bayesian Analysis for Gaussian Hierarchical Models with Intrinsic Conditional Autoregressive Priors**

*Yan Lu*

*University of New Mexico*

In this research, we proposed two test statistics for testing that a predictor variable has no effect on the response variable in nonparametric regression with survey data. This was accomplished by incorporating consistent estimators under the survey design, and by adopting a data-driven nonparametric order selection method. We also investigated the large sample properties of the test statistics under the null hypothesis, and performed simulation studies to examine the statistical power and type I errors of the tests.

**Small Area Models for Over-dispersed Poisson Counts**

*Jerry Maples*

*US Census Bureau*

Recent work in small area estimates at sub-county level has focused on estimating the rate of poverty in school-aged children (e.g. Franco 2015). To obtain counts of children in poverty require having known population counts, often assumed without error, for these small areas. In general, this is not always the case and the uncertainty due to not knowing the true population counts is often not reflected in the predictions. The sampling distributions from simulating the American Community Survey design appear more similar to over-dispersed Poisson distributions than standard Poisson. We propose a small area model for over-dispersed Poisson count data to model the number of children in poverty at the census tract level. Modeling assumptions will be tested using the simulated samples from design-based simulation.

**Best Look-Alike Prediction with Application to Small Area Estimation**

*Thuan Nguyen*

*Oregon Health & Science University*

A criterion of optimality in prediction is proposed that requires the predictor to assume the same type of values as the random variable it is predicting. In the case of categorical responses, the method leads to the Bayesian classifier with a uniform prior. However, the method extends to other cases, such as zero-inflated observations, as well. The method, called best look-alike prediction (BLAP), justifies an ``usual practice'' from a theoretical standpoint. We focus on application of BLAP to small area estimation with zero-inflated area-specific random effects. The work is joint with Hanmei Sun, Yihui Luan of Shandong University, China and Jiming Jiang of University of California, Davis, USA.

**Estimation and Prediction in the Presence of Spatial Confounding**

*Garritt L. Page, Brigham Young University*

  In studies that produce data with spatial structure it is common that covariates of interest vary spatially in addition to the error. Because of this, the error and covariate are often correlated. When this occurs it is difficult to distinguish the covariate effect from residual spatial variation. In an iid normal error setting, it is well known that this type of correlation produces biased coefficient estimates but predictions remain unbiased. In a spatial setting recent studies have shown that coefficient estimates remain biased, but spatial prediction has not been addressed. The purpose of this paper is to provide a more detailed study of coefficient estimation from spatial models when covariate and error are correlated and then begin a formal study regarding spatial prediction. This is carried out by investigating properties of the generalized least squares estimator and the best linear unbiased predictor when a spatial random effect and a covariate are jointly modeled. Under this setup we demonstrate that the mean squared prediction error is possibly reduced when covariate and error are correlated. This work is joint with Yajun Liu, Chong Z. He and Dongchu Sun

**Conditional moment restriction models with incomplete data**

*Valentin Patilea*

*ENSAI*

A general statistical model is considered which is defined by moment restrictions when a subvector of data are missing. The main incomplete data situations we have in mind are missing at random and endogenous selection. Using the inverse probability weighting, we show that such a model is equivalent to a model for the observed variables only, augmented by a moment condition defined by the incompleteness mechanism. In particular, our framework covers parametric and semiparametric mean regressions and quantile regressions. We allow for missing responses, missing covariates and any combination of them. We present a general equivalence result, obtained under minimal technical conditions, that sheds new light on various aspects of interest in the missing data and econometric literature. It also provides guidelines for building (efficient) estimators. This talk is based on a joint work with Marian Hristache.

**A study of the effect of measurement error on EB and HB predictors**

*Silvia Polettini, Serena Arima*

Sapienza University of Rome

Model-based small area estimation relies on mixed effects regression models that link the small areas and borrow strength from similar domains. To describe the population data, we consider a nested error linear regression model with auxiliary variables denoted by X and response variable y. The model can be specified in a hierarchical way as described in detail e.g. in Rao and Molina (2015); we denote the model’s parameters by . Assuming the validity of the model for the population as well as for the sample data, and denoting by y(1),X(1) the sampled portion of the data, the model can be exploited to predict the small area means given the available information. Two approaches can be pursued in this framework: the first is the Empirical Bayesian (EB) approach, under which is evaluated and plug-in of suitable parameter estimates produce the final estimator . The alternative is to pursue a fully Bayesian approach, integrating the distribution of with respect to the posterior distribution of the unknown parameter φ. This leads to the so called Hierarchical Bayesian (HB) predictor. Full account of the different approaches is presented e.g. in Rao and Molina (2015).

In this contribution we consider nested error linear regression models under measurement error in auxiliary variables. Working on the EB predictor proposed in a seminal paper by Ghosh et al. (2006), Torabi et al. (2009) derive the Bayes predictor of the small area means, that they define fully efficient in the sense that it exploits all the information conveyed by the sample, including the values of the observed covariates, which is not included in the predictor of Ghosh et al. (2006). These proposals rely on the conditional distribution of the unobserved portion of the data given the sample and the model’s parameters and have only been investigated for a single continuous auxiliary variable.

Assuming a nested error linear regression model with error in covariates, we analyze the behaviour of the above mentioned EB predictors when the measurement error increases, contrasting it with the direct estimator and with the HB predictor. We also test the performance of the above mentioned estimators through a simulation study exploring a variety of scenarios and study their sensitivity to model assumption violations.

**Estimation of the parameters of the Asymmetric Laplace Distribution under informative sampling.**

*M. Giovanna Ranalli, University of Perugia*

We will use a full-information maximum likelihood approach for the estimation of the parameters of the Asymmetric Laplace Distribution (ALD) in the presence of informative sampling. The interest is not only on the parameters of the ALD per se. The focus on the ALD is motivated by the fact that, in a classical inferential setting, the solution to the MLE estimating equation for the location parameter of the ALD for a given value of the skewness parameter $\tau$, is also the solution to the L1 estimating equation for the $\tau$-th quantile of interest. We will investigate how this property can be used in the presence of informative sampling to estimate the quantiles of the population distribution of a scalar variable and the parameters of a quantile regression. Joint work with Ray Chambers and Nicola Salvati.

**Some Current Developments and Future Directions**

*J. N. K. Rao Carleton University*

Some current developments of methodological interest and practical importance in small area estimation will be presented. Future directions related to the use of big data and non-probability sampling and multiple data sources will also be discussed.

**Small Domain Estimation for discontinuities in the National Survey for Wales**

*Timo Schmid*

*Free University of Berlin*

Redesigning a survey generally affects the non-sampling errors and therefore has a systematic effect on the survey estimates. These kinds of systematic differences are called discontinuities. Separating real changes from discontinuities due to the redesign in a survey transition is important to maintain uninterrupted time series of estimates. Part of the process for transition to the new surveys is to provide users with information on the likely effect of the change to the new survey on the existing estimates, and to help them to use this information to interpret the changes where they are of substantive importance for policy purposes. To provide early information on these effects, a large-scale pilot survey was implemented in 2015 alongside (for some surveys) or shortly after (for others) the last instances of the original surveys. This paper discusses work that was carried out to assess discontinuities, and how users should interpret the discontinuities.

The discontinuities may be assessed in three main ways – at a national level using a direct estimator; at a domain level using a direct estimator; and at domain level using indirect estimation to deal with problems due to small sample sizes. All three strategies were used in this instance. The strategies will be described, and some examples of each will be shown to demonstrate the differences between them in particular cases. We will formulate an outline best practice for general use, although specific analysis may be needed in any particular case to help to make a decision on what is the best approach.

**Real-time estimation of monthly unemployment with dynamic factor and time-varying state space models**

*Caterina Schiavoni*

*Statistics Netherlands and Heerlen and Maastricht University School of Business and Economics*

Long term follow-up with longitudinal data is common in many medical investigations. In such studies, some longitudinal covariates can be omitted for various reasons. In cross sectional studies, coefficient estimation of a covariate is unbiased if the covariate is orthogonal to the omitted covariate. This is not true in longitudinal data analysis, where omission of time dependent covariate can lead to biased coefficient estimate even if the corresponding covariate is orthogonal to the omitted longitudinal covariate. In this article, we propose a new unbiased estimation method to accommodate omitted longitudinal covariate. In addition, if the omitted longitudinal covariate is asynchronous with the longitudinal response, we propose a two stage approach for valid statistical inference. Asymptotic properties of the proposed parameter estimates are established. Extensive simulation studies provide numerical support for the theoretical findings. We illustrate the performance of our method on a dataset from an HIV study.

**Hybrid BRR and Parametric-Bootstrap Variance Estimates for Small Domains in Large Surveys**

*Eric Slud*

*University of Maryland/US Census Bureau*

Following its justification in widely cited papers (McCarthy 1969, Krewski & Rao 1981, Fay 1984, 1989), Balanced Repeated Replication (BRR) has become a standard method for variance estimation in large complex surveys, especially in the US. However, it is also known that BRR variance estimates for very small domains are unreliable. Survey point estimates for small domains are often based on empirical-Bayes small area estimation models (Rao and Molina 2015), with variances estimated through parametric-bootstrap methods. This talk describes theory and practice of a novel hybrid approach, in which variances are estimated via parametric-bootstrap replications nested within BRR weight-replications. The theoretical underpinnings of the method are discussed within the Dirichlet-multinomial hierarchical model describing small outcome proportions for categories within larger (but sometimes still small) domains, as applied within the recent analysis of 2010-2014 American Community Survey data to estimate language-minority and proficiency characteristics in arriving at alternative-language ballot assistance determinations under the terms of Section 203 of the Voting Rights Act.

**Entity Resolution with Societal Impacts in Statistical Machine Learning**

*Rebecca C. Steorts, Duke University*

Very often information about social entities is scattered across multiple databases. Combining that information into one database can result in enormous benefits for analysis, resulting in richer and more reliable conclusions. Among the types of questions that have been, and can be, addressed by combining information include: How accurate are census enumerations for minority groups? How many of the elderly are at high risk for sepsis in different parts of the country? How many people were victims of war crimes in recent conflicts in Syria? In most practical applications, however, analysts cannot simply link records across databases based on unique identifiers, such as social security numbers, either because they are not a part of some databases or are not available due to privacy concerns. In such cases, analysts need to use methods from statistical and computational science known as entity resolution (record linkage or de-duplication) to proceed with analysis. Entity resolution is not only a crucial task for social science and industrial applications, but is a challenging statistical and computational problem itself. In this talk, we describe the past and present challenges with entity resolution, with applications to the Syrian conflict but also official statistics, and the food and music industry. This work, which is a joint collaboration with researchers at Rice University and the Human Rights Data Analysis Group (HRDAG) touches on the interdisciplinary research that is crucial to problems with societal impacts that are at the forefront of both national and international news.

**Multivariate small area estimation under informative sampling and nonresponse（old）**

*Michael Sverchkov (Bureau of Labor Statistics)*

We consider multivariate small area estimation under informative sampling and 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 are missing), and estimate the response probabilities by application of the Missing Information Principle. By this principle, we first define the likelihood score equations as if the missing outcomes were 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. Finally, the integrated score equations are solved with respect to the unknown parameters indexing the response model. See Sverchkov (2008), Riddles et al. (2016) and Sverchkov and Pfeffermann (2018) for application of this approach in the univariate case. Once the response probabilities are estimated, we impute the missing outcomes and then apply the approach of Pfeffermann and Sverchkov (2007) to the complete data set (observed and imputed values), to obtain the small area predictors. This is a joint work with Danny Pfeffermann.

**On Small Area Estimation under Informative Sampling and Response Not Missing at Random: New Developments**

*Michael Sverchkov, US Bureau of Labor Statistics*

In this paper we extend the general approach presented at ISI Satellite SAE conference (2015) by Sverchkov and Pfeffermann. The approach is based on Pfeffermann and Sverchkov (JASA 2007) paper that considered Small Area Estimation when the selection of the sampled areas is with probabilities that are related to the true (unknown) area means, and the sampling of units within the selected areas is with probabilities that are related to the study variable values, both after conditioning on the model covariates, and Sverchkov (JSM 2008) paper that suggest a way of estimating response probabilities under response not missing at random. We illustrate our approach by a real data example.

**Fixed-Effect Log-Linear Models for Small Area Estimation in presence of relatively Sparse Contingency Tables**

*Yves Thibaudeau, US Census Bureau*

In situations where statistical estimation over many small areas is requested and all the measurements are discrete and nominal, “fixed effect” log-linear models can be a simple and easily interpretable avenue to achieve that goal. While random effects are interpreted as realizations of correlateld random variables, fixed-effects are defined as interactions parameters between the response variables and the “small areas”. This approach puts front and center issues of parameter estimability and statistical validity of the model. Log-linear models in various set-up (random and fixed effect models) applied to small area estimation are discussed by Zhang and Chambers (2004), Hasslet and Arnold (2002) and Marker (1999). In the context of fixed effect, Schafer (1997) notes there is an equivalence between conditional logit models and a specific class of fixed-effect log-linear models. This equivalence can be exploited to compute finite population estimators of the type proposed by Pfeffermann, Skinner and Humphreys (1998) in a way that reduces the number of parameters that need to be estimated relative to a strictly log-linear modeling approach (Thibaudeau, Slud and Gottschalck 2017). We study fixed-effect log-linear models and we center the attention on relatively sparse longitudinal cross-classifications over many small areas, as extracted from the Current Population Survey. We experiment with methods such as conditional logistic regression and cell collapsing to deal with sparseness and guarantee estimability of model parameters. . This is a joint work with ric Slud and Yang Cheng.

**A Unified Approach to Goodness-of-fit Tests with Application to Small Area Estimation**

*Mahmoud Torabi*

*University of Manitoba*

We develop a method originally proposed by R. A. Fisher into a general procedure, called tailoring, for deriving goodness-of-fit tests that are guaranteed to have a chi-squared asymptotic null distribution. We apply the method to small area estimation for detecting potential model misspecification. Two tests are proposed using the tailoring method. We evaluate performance of the tests both theoretically and empirically. Our empirical results suggest that the proposed tests are more accurate in size and have higher power than existing tests. This is a joint work with Jiming Jiang (University of California, Davis).

**New developments on the estimation of discontinuities in sample surveys**

*N. Tzavidis, Southampton University*

Many Official statistics rely on survey data collection, and much of the value in such surveys lies in their continuity, enabling developments in society and the economy to be monitored, and policy actions decided. This is often an argument to keep survey processes of repeated surveys unchanged as long as possible. From time to time changes in surveys are needed to improve the efficiency, reduce the survey related costs, or meet new requirements, and this is seen strongly in the use of mixed mode surveys including web-based questionnaires in official statistics. These changes affect the survey outputs. To avoid the implementation of a new survey process disturbing the comparability of estimates over time, it is important to quantify the impact on the estimates of a repeated survey. In this session three papers present recent developments of statistical methods aimed to quantify such discontinuities at domain level.

**A No Calculation When Observation Can Be Made**

*Tommy Wright*

*U.S. Bureau of the Census, Georgetown University*

For use in connection with the general and complete observations that would be known from a full census, Kiaer (1895, 1897) presents a purposive “Representative Method” for sampling from a finite population.  Many credit this method with laying seeds for current sampling methods used in producing official social and economic statistics.  At a time when just about all official statistics were produced by censuses, Kiaer had much opposition, especially from statistician von Mayr, who said (a translation), “…no calculations when observations can be made.”

Neyman (1934) brings probability to this Representative Method using stratified random sampling and presents details for the well-known and widely used optimal allocation of the fixed sample size among the various strata to minimize sampling error.  When sample sizes are rounded to integers from Neyman’s allocation, minimum sampling error is not guaranteed.  Wright (2012,2014,2016,2017) improves Neyman’s result with a simple decomposition obtaining exact results that always yield integer sample size allocations while minimizing sampling error. We demonstrate a decrease in needed sample size with exact optimal allocation for the same precision using 2007 Economic Census data in the sample design for part of the subsequent Service Annual Survey.

We conclude by calling on the phrase “…no calculation when observation can be made” to muse about current world-wide considerations to make greater use of data from additional sources (e.g., administrative records, commercial data, big data…) to produce official statistics.

**Regroup model in blinded sample size re-estimation for cluster randomized trials**

*Chien-Hua Wu*

*Chung-Yuan Christian University*

Cluster randomization design plays an important role in clinical trials in recent years. The cluster randomized controlled trial considers the cluster as the unit of randomization, rather than allocating each individual patient to treatment groups. The clusters are nested within the treatments. The treatment responses are obtained from subjects in various clusters within selected treatments. Because of uncertainty in the ingredients of sample size calculation, it is desirable to reevaluate the sample size in the middle of the trial. Without breaking the randomization codes, we provide a guideline to evaluate the minimum total sample size and the number of clusters for interim analysis. This is a joint work with Shu-Mei Wan ([Lunghwa University of Science and Technology](http://english.lhu.edu.tw/)).

**Statistical Inference with Non-probability Survey Samples**

*Changbao Wu*

*University of Waterloo*

We establish a general framework for statistical inferences with non-probability survey samples. We develop a rigorous procedure for estimating the propensity scores for units in the non-probability sample, and construct robust estimators for finite population means. Variance estimation is discussed under the proposed framework. Results from simulation studies and real data analysis show the robustness and the efficiency of our proposed estimators as compared to existing methods. This is a joint work with Yilin Chen and and Pengfei Li (University of Waterloo).

**Generalized Variance Functions for Longitudinal Data**

*Guoyi Zhang*

*University of New Mexico*

This is a joint work with Yang Cheng from the U.S. Census Bureau and Yan Lu from the University of New Mexico.

Generalized variance functions (GVFs) are used to produce convenient published estimates of variances for a number of large surveys such as the Current Population Survey (CPS). The GVF pooled together one year or some certain time period of data and treated them cross sectionally. However, as the size of population changes over time, the standard error of totals can actually change substantially. In this research, we propose longitudinal generalized variance functions (LGVFs) by incorporating the time effect into modeling. Asymptotic properties for estimators that are linear combinations of sample cluster means from stratified, two-stage cluster samples are investigated. Implementation of the methods to CPS are illustrated.

**Bayesian Estimation of Demographic Systems**

*Junni Zhang*

*Beijing University*

Many problems in applied demography consist of estimating demographic systems. We present a general approach to estimating demographic systems based on the idea of a demographic account.  A demographic account is a set of linked tabulations of demographic series, such as fertility, mortality, migration, and population. The estimation methods are fully Bayesian and allow for multiple noisy data sources. We simultaneously estimate (i) the true finite-population counts of events and populations, (ii) the demographic rates underlying these counts, and (iii) indicators of data quality. The methods also generate uncertainty measures for all estimated quantities. We illustrate the methods using examples from New Zealand and China.

**On the validity of descriptive inference from non-probability samples**

*Li-Chun Zhang*

*University of Southampton and Statistisk sentralbyrå*

The uptake of non-probability samples is often motivated on account of data collection costs and extended data scope and availability. Conditions for valid inference have previously been established for inference from non-probability samples, under the super-population modelling approach. However, the formulation does not apply to the so-called quasi-randomisation approach, under which a model of the inclusion probability of the population units is postulated but the outcome values of interest are treated as constants. Moreover, it seems unrealistic, as one often does in the literature, to assume that the inclusion probability is a function of whichever available covariates that fully capture the underlying observation mechanism, e.g. when the non-probability sample suffers from under-coverage because some units have zero inclusion probability. In this talk I outline and discuss the conditions for valid inference from non-probability samples under the quasi-randomisation approach. Special attention is given to descriptive inference of a finite population. The difficult challenge of validating the validation conditions themselves is emphasised.

**Using state space models for an efficient parallel run design and measuring statistical impacts of Australian Labour Force Survey Redesign**

*Xichuan (Mark) Zhang*

*Australian Bureau of Statistics*

The implementation of a major business transformation program in an official statistical agency is designed to achieve, among other things, improvements in data collection efficiency, data processing methodology and data quality. However, achieving such improvements can, in itself, result in transitional statistical impacts which could be misinterpreted as real world change if they are not measured and handled appropriately.

This paper describes a range of statistical methods to measure the statistical impacts which may be encountered in future ABS Labour Force Survey (LFS) redesigns, including designing and conducting parallel collection activities such that the outgoing and the incoming surveys are run in parallel for a period of time to measure the impact of any collection changes, and refining the precision of impact measurement while implementing a new survey design.

State space modelling techniques are proposed as the main approach to incorporate sampling error structure and time series intervention, and to take advantage of multiple data sources related to the LFS to improve impact measurement efficiency and accuracy. The models and methods developed can be extended to other surveys.

**The Use of Machine Learning Methods to improve the US National Resources Inventory Survey**

*Zhengyuan Zhu*

*Iowa State University*

The National Resources Inventory (NRI) Survey is one of the largest annual longitudinal survey of soil, water, and related environmental resources in the US designed to assess conditions and trends on non-federal US lands. The data collected for NRI is based on a probability-based area sample, which is designed to provide accurate national and state estimates. One particular challenge we have in NRI is that there is a 3-year lag in publishing the NRI data due to resource constraints on data collection. It is of great interests to federal and state stakeholders to have more timely preliminary NRI estimates based on partially collected data and auxiliary information under the current budget constraints. We also receive strong requests from local stakeholders to provide data at county and small watershed level. In order to provide more timely estimates at smaller spatial scales, it is necessary to integrate alternative big data sources such as administrative data and satellite data with the survey data in our estimation. In this talk we give a brief introduction to the NRI, and share our experience using satellite data and machine learning methods to improve NRI estimation. New statistical method for satellite data gap-filling and machine learning methods for satellite data based land-cover classification will be introduced, which are useful for the NRI small area estimation and forecasting.

**Poster Abstracts**

**Challenges Confronted and Insights Revealed in Synthesizing State-level Integrated Data**

*Daniel Bonnery (University of Maryland and Maryland)*

The Maryland Longitudinal Data System (MLDS) is a central repository of student and workforce data, including data provided by the Maryland State Department of Education, the Maryland Higher Education Commission and the Maryland Department of Labor, Licensing and Regulation. The Institute of Educational Sciences is funding a project to produce and release synthetic versions of selected longitudinal state-level datasets. The use of synthetic data is of increasing interest to many state longitudinal integrated data systems that also seek to balance researcher access with data privacy concerns. Practical tools that implement generic synthesization methods exist. Nevertheless, synthesizing large integrated data presents specific methodological and practical challenges. Longitudinal integrated data involves: a lot of variables, redundancy (and inconsistency) of information, specific often non-random missing data patterns, and different levels or dimensions. We propose to detail the nature and implications of these challenges and describe the solutions we are applying in our ongoing MLDS Synthetic Data Project. This is a joint work with Laura Stapleton (College of Education, University of Maryland and Maryland Longitudinal Data System Center) and Michael Woolley(University of Maryland School of Social Work and Maryland Longitudinal Data System Center)

**Hierarchical Bayesian Models for Sub-Areas with Binary Data**

*Lu Chen*

*Worcester Polytechnic Institute*

Many population-based surveys have binary responses from a large number of individuals in each household (sub-area) within small areas. An example is the Nepal Living Standards Survey (NLSS II), in which health status binary data for each individual from sampled household are available in sampled wards (small areas). To make inference for the finite population proportion of individuals in each household, we develop hierarchical Bayesian models with sub-area random effects under two models. The first model is the sub-area Beta-Binomial model without covariates. We use an approximation method with random sampling to fit the model efficiently. We applied our model to NLSS II data to show that approximation method can provide good estimates as the exact method. The second model is the two-fold logistic regression model with reliable auxiliary information. The contribution of this model is twofold. First, we extend an integrated nested normal approximation (INNA) area level model to sub-area level model. Second, because there are numerous sub-areas, standard Markov chain Monte Carlo (MCMC) methods to find the joint posterior density are very time consuming. Therefore, we provide a sampling-based method, which permits fast computation. Our main goal is to describe this twofold hierarchical Bayesian logistic regression model and to show that the computation is much faster than the exact MCMC method and also reasonably accurate. The performance of our method is studied by using NLSS II data. We further compare these models with one-fold logistic regression model using NLSS II data and it shows that the twofold model is preferred over one-fold model that ignores the sub-areas within areas. Our models can borrow strength from both areas and sub-areas to obtain more efficient and precise estimates. The hierarchical structure of our model captures the variation in the binary data measurably well.

**Estimation of Contingency Tables for Small Areas Using Multi-Level Loglinear Models**

*Robert Clark (Australian National University)*

*Mossamet Kamrun Nesa (University of Wollongong)*

*Carole Birrell (University of Wollongong)*

It is sometimes of interest to estimate population counts cross-classified by multiple survey variables, for example Smoking Status by High Blood Pressure by the health districts of a country. Methods are available for estimating counts by a single categorization (for example labour force status) using random effects multinomial models. We apply these methods to the estimation of 2 by 2 tables for small areas, using saturated and unsaturated loglinear models with random effects. Mean squared errors are estimated by parametric bootstrapping. The methods are evaluated by simulation and applied to the New Zealand Health Survey. Small area estimators using the new approach have much lower mean squared errors than direct estimators.

**Estimation of Contingency Tables for Small Areas Using Multi-Level Loglinear Models**

*Isabela Bertolini Coelho*

*Regional Center for Studies on the Development of the Information Society*

The Regional Center for Studies on the Development of the Information Society (Cetic.br), a department of the Brazilian Network Information Center - NIC.br, collects data about access, use and appropriation of Information and Communication Technologies (ICT) in several segments of society. The Survey on the use of information and communication technologies in Brazilian households: ICT Households (NIC.br, 2016) is designed to produce reliable estimates for five Brazilian regions. In order to estimate some ICT Indicators (ITU, 2014) disaggregated at the state level, three different small area estimation methods are compared: simple mean of consecutive years (Thomas & Wannell, 2009); pooling samples of consecutive years (Thomas & Wannell, 2009); and the composite estimator (RAO, 2003). The first method computes the average between the estimates for 2014 and 2015 for each state. The second, combines data from the 2014 and 2015 samples into one dataset and is analyzed as sample from a combined population. The last method, which considers data just from 2015, adopts the regions estimate as the synthetic estimate, and combines this with the direct estimate for the state in a composite estimator, using varying weights empirically chosen according to the direct estimate’s coefficient of variation (CV) and the representativeness of the state’s available sample (Schouten et.al., 2009). Results are compared based on mean square error estimates obtained using a bootstrap estimation procedure proposed in (Mecatti, 2000) and adapted to our survey situation. The best results are observed for the composite estimator. For these estimates, a comparison with the IBGE National Household Sample Survey was made to provide insights that contribute to the interpretation of the estimates obtained. This is a joint work with Marcelo Trindade Pitta and Pedro Lu ́ıs do Nascimento Silva.

**Measurement Error in Small Area Estimation: Functional Versus Structural Versus Naive Models**

Gauri S. Datta

University of Georgia and U.S. Census Bureau

Small area estimation using area-level models can sometimes benefit from covariates that are observed subject to random errors, such as covariates that are themselves estimates drawn from another survey. Given estimates of the variances of these measurement (sampling) errors for each small area, one can account for the uncertainty in such covariates using measurement error models (e.g., Ybarra and Lohr, 2008, Biometrika). Two types of area-level measurement error models have been introduced in the small area estimation literature. The functional measurement error model assumes that the underlying true values of the covariates with measurement error are fixed but unknown quantities. The structural measurement error model assumes that these true values follow a model, leading to a multivariate model for the covariates observed with error and the original dependent variable. We compare and contrast these two models with the alternative of simply ignoring measurement error when it is present (naive model), exploring the consequences for prediction mean squared errors of use of an incorrect model under different underlying assumptions about the true model. Comparisons done both analytically and via simulations yield some surprising results. We also illustrate results when the models are applied to data from the U.S. Census Bureau's Small Area Income and Poverty Estimates Program.This is a joint work with William R. Bell, Hee-Cheol Chung and Carolina Franco.

**Comparative Study of Confidence Intervals for Proportions in Complex Sample Surveys**

*Carolina Franco (USA Census Bureau)*

The most widespread method of computing confidence intervals (CIs) in complex surveys is to add and subtract the margin of error (MOE) from the point estimate, where the MOE is the estimated standard error multiplied by the suitable Gaussian quantile. This Wald-type interval is used by the American Community Survey (ACS), the largest US household sample survey. For inferences on small proportions with moderate sample sizes, this method often results in marked under-coverage and lower CI endpoint less than 0. We assess via simulation the coverage and width, in complex sample surveys, of seven alternatives to the Wald interval for a binomial proportion with sample size replaced by the ‘effective sample size,’ that is, the sample size divided by the design effect. Building on work of Franco et al. (2014), our simulations address the impact of clustering, stratification, different stratum sampling fractions, and stratum-specific proportions. We show that all intervals undercover when there is clustering and design effects are computed from a simple design-based estimator of sampling variance. Coverage can be better calibrated for the alternatives to Wald by improving estimation of the effective sample size through superpopulation modeling. This approach is more effective in our simulations than modifications of effective sample size proposed by Korn and Graubard (1998) and Dean and Pagano (2015). We recommend intervals of the Wilson or Bayes uniform-prior form, with the Jeffreys-prior interval not far behind. This is a joint work with Roderick J. A. Little (University of Michigan), Thomas A. Louis (Johns Hopkins Bloomberg School of Public Health) and Eric V. Slud (US Census Bureau and University of Maryland, College Park).

**Bayesian Hierarchical Spatial Regression Model for Missing Covariates with Application to Chinese Health and Nutrition Survey Data**

*Guanyu Hu (University of Connecticut)*

Compared with existing models, we introduce spatial random effects on missing covariates model. It will improve the quality of the imputation of missing co- variates. We modify the deviance information criterion (DIC) and the logarithm of the pseudo-marginal likelihood (LPML) based on conditional predictive ordinate (CPO) to evaluate model performance. The empirical results suggest our model has competitive estimation results with all records estimation. And our model performs better than other models basing on DIC and LPML. Additionally, we analyze a Chinese Health and Nutrition Survey (CHNS) 2011 dataset as an illustration of the effectiveness of our approach. This is a joint work with Zhihua Ma (Jinan University) and Ming-Hui Chen (University of Connecticut).

**Use of SPREE estimation to estimate the number of unemployed**

*Tomasz Józefowski (Statistical Office in Poznań)*

*Tomasz Klimanek (Statistical Office in Poznań)*

*Marcin Szymkowiak (Poznań University of Economics and Business, Statistical Office in Poznań)*

Unemployment is one of the most serious socio-economic phenomena and remains a challenge, which is particularly difficult to solve. It can lead to poverty, social pathology and exclusion of entire social groups and areas which it affects. This in turn can cause social tensions. This is why, in order to make appropriate decisions and counteract the negative effects of unemployment, national and local authorities require reliable and detailed information about the number of unemployed.

At the moment the Labour Force Survey is the main source of quarterly data about the level of unemployment as defined by the International Labour Organisation. The LFS is a representative survey, whose methodology is periodically updated to comply with Eurostat guidelines. LFS-based statistics provide information about the number of unemployed at province level broken down by sex or place of residence (urban/rural). Information for more detailed domains is not available because of insufficient sample size, which is related to the unacceptable level of estimation precision in the case of the classical Horvitz-Thompson estimator.

This problem can be solved by applying methods of small area estimation (SAE). The main advantage of SAE estimators over direct estimators used in the classical representative method is the fact that they provide reliable estimates even if the sample size in a given domain is small or equal to zero [Rao and Molina 2015]. However, the problem is that when small area estimates for lower levels of spatial aggregation or smaller subpopulations are summed up, they often tend to differ from totals estimated using direct estimators for higher level domains, for which the sample size is sufficient.

Such inconsistencies go against the policy of the Central Statistical Office, which strives to ensure coherence between different statistical outputs. This difficulty can be overcome by applying SPREE estimation [Zhang and Chambers 2004]. The main goal of this article is to present how the SPREE estimator and its generalized versions (GLSM, GLSMM) can be used to obtain quarterly estimates for domains that have so far not been published, i.e. the number of unemployed by province by sex by place of residence, using data from the unemployment register and the LFS for 2012-2015. In addition, the obtained results will be assessed in terms of quality.

**Weighted Matching on-the-fly: Improved Sequential Allocation with Higher Power and Efficiency**

*Adam Kapelner*

*City University of New York*

We propose a covariate-adjusted response-adaptive randomization design that increases power and efficiency when measuring an average treatment effect in a two-arm sequential randomized trial. As subjects arrive, they are either randomized or matched to a previously randomized subject and administered the alternate treatment. The match criterion is a distance function between covariates updated iteratively. The distance components are weighted by the strength of the relationship of each covariate to the response. These weights are also robust to collinearity between the covariates. The threshold that determines closeness of the match is also updated iteratively via a bootstrapped estimate of its quantile. We develop estimators for the average treatment effect that combine information from both the matched pairs and unmatched subjects and recommend a permutation procedure for inference. Simulations illustrate our designs higher efficiency (and power) over several competing allocation procedures in simulations and in data from a clinical trial.

**Estimating a finite population mean under random non response in two stage cluster sampling with replacement**

*Nelson Kiprono*

*Strathmore University*

Non-response is a regular occurrence in sample surveys. Developing estimators when non-response exists may result in large biases when estimating population parameters. In this paper, a finite population mean is estimated when non-response exists randomly under two stage cluster sampling with replacement. It is assumed that non-response arises in the survey variable in the second stage of cluster sampling. Weighting method of compensating for non-response is applied. Asymptotic properties of the proposed estimator of the population mean are derived. Under mild assumptions, the estimator is shown to be asymptotically consistent.

**Marginal and Conditional Simultaneous Inference in Small Area Estimation**

*Peter Kramlinger Tatyana Krivobokova Stefan Sperlich*

*Australian Bureau of Statistics*

In small area estimation, the empirical best linear unbiased predictor (EBLUP) is used if the interest lies in both characterizing the realized random effect as well as in predicting future realizations. However, inference in the former case is plausible to be performed conditional on the random effects. Then, marginal individual con- fidence intervals for small area parameters are inaccurate due to the inherent bias of the estimators. On the other hand, direct conditional confidence intervals are unacceptably wide due to the small sample size in each area. These effects can be avoided when performing inference simultaneously to all areas.

Simultaneous inference ist best applicable if the estimators are largely correlated. As the EBLUPs are relatively stronger correlated for smaller sample sizes, small area estimation is a natural candidate for simultaneous inference.

We provide and discuss marginal simultaneous confidence sets suitable for testing for all areas as well as multiple comparisons and conditional simultaneous confi- dence sets. In particular, we show that marginal simultaneous confidence sets are approximately accurate even when performing inference conditional on the random effects.

**Estimation of wealth on spatially disaggregated levels in Germany based on the Household Finance and Consumption Survey (HFCS)**

*Ann-Kristin Kreutzmann, Freie Universität Berlin*

The Deutsche Bundesbank and other central banks are supposed to report the wealth distribution. For instance, the European Household Finance and Consumption Network reports median values of net wealth and its components by countries and different demographics (Household Finance and Consumption Network 2016). However, when the sample size is below 25 estimates are not published. The Deutsche Bundesbank (2016) reports on a regional level only indicators for four regions (north, west, south and east). One option to tackle the problem of small sample sizes when the interest lies on subpopulations is to use small area estimation methods. These methods combine the existing survey data with administrative or census data.

In this presentation we use Fay-Herriot (FH) type models (Fay and Herriot 1979) in order to get reliable estimates for the mean of household net wealth in the German federal states and 96 planning regions using the HFCS data. As the HFCS faces item non-response, the missing values are imputed by multiple imputation. Therefore, we propose a FH estimator that additionally accounts for the variability due to the multiple imputation. The estimators are based on a logarithmic transformation and meet a benchmarking constraint for internal (national) consistency. Finally, we give an outlook to other indicators like the wealth head count ratio. This is a joint work with Philipp Marek, Timo Schmid and Nicola Salvati.

**A Kernel Weighting Approach to Improve Population Representativeness of Epidemiological Cohort in the Analysis**

*Yan Li*

*University of Maryland*

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. In this paper, 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 proposed JK variance estimators are accurate for variance of IPS-weighed estimators. The Taylor Linearization estimators, ignoring variability due to estimating propensity scores, substantially underestimate variances of IPS-weighted estimators, but are good alternative to the JK for variance estimation of kernel-weighted estimators. 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

**Uniform prediction intervals for small area means in linear mixed models**

*María José Lombardía ∗1, Katarzyna Reluga †2 and Stefan Sperlich ‡2*

*1 Department of Mathematics, Universidade da Coruña, A Coruña, Spain*

*2 Research Center for Statistics and Geneva School of Economics and Management, University of Geneva,*

*Switzerland*

An increasing demand for reliable statistics regarding the socio-demographic groups and geo- graphical regions contributed to the development of Small Area Estimation (SAE). To the best of our knowledge, little has been proposed to address the multiple comparison (or testing) problem, which boils down to the simultaneous comparison of all or several areas. Such a comparison can be of huge interest, for example, in statistical offices providing reports to policy makers, or in public health research centres carrying out analyses on demographic groups. We fill an important gap in the field of SAE inference and present a general methodology to construct uniform prediction intervals (UPI) for the linear combination of fixed and random effects in linear mixed models. We consider three frameworks: an analytical derivation based on the volume of tube formula of Weyl (1939), a numerical estimation which approximates the aforementioned analytical expression, and a new approach based on bootstrapping that we develop. As far as the two former methods are taken into consideration, we follow closely the reasoning of Sun and Loader (1994) and Krivobokova et al. (2010), adapting it to the SAE framework. When it comes to the bootstrap based method, it is more flexible as it does not require normality assumptions on the distribution of random effects and errors. Furthermore, we developed simultaneous intervals, which are not only accurate but also not excessively wide. Our proposal is accompanied by a simulation experiment and a data example.

**Use of Drone Camera on Tracking Impacts of Human Activities on Forestry Development.**

*Elieza Paul Milonga*

Development of methodology, data collection, capacity development, and coordination in the fields of environmental statistics and indicators depend much on the technology applied in the whole process of Environment statistics production, with an attachment that equipment eg.satellites,remote sensing tools are very expensive for developing countries to afford. Tanzania like other developing countries is resource constrained because it relies on government subventions and international donors as sources of revenue to finance its expenditure budget both recurrent and development that leads to poorly implemented and monitored environment protection and climatic changes. This has not only translated in poor agricultural production but also a weak forest conservation. However, the Tanzania has a number of opportunities that can be tapped to enhance production of environment statistics and changes tracking system. This paper shall discuss the use of affordable Drone Cameras that can be applied routinely in collection of data to support decision makers in policy making so as to improve, encourage and ensure that people occupy and utilize land in a proper manner for socio-economic developments without causing degradation to the environment.

**Small area estimation under Post Randomization (PRAM)**

*Silvia Polettini, Serena Arima*

Sapienza University of Rome

It is crucial for Statistical Offices to be able to meet the growing demand for microdata without disclosing confidential information about individual records; in this respect, the availability of methods that enable to release data at the record level is one of the main options that Statistical Offices must consider as an alternative to remote access or research data centers.

In this paper we focus on a perturbation techniques, called Post Randomization Method (Gouweleeuw et al. 1998). PRAM is a disclosure limitation technique that protects categorical data by applying a known random perturbation scheme. The authors show that, when the associated transition probability matrix is released, valid statistical inferences are allowed based on the perturbed data. In practice however, Statistical Offices prefer not to release the perturbation matrices for confidentiality concerns, which may limit data utility to a large extent.

We focus on post-release analyses and consider the case when PRAM is applied to the categorical covariates of a small area model; we assume that users adapt their inferences to account for perturbation. We define a Bayesian mixed effects model that allows for measurement errors in categorical variables, by explicitly introducing the error distributions induced by the data protection. We show that the transition probabilities among categories in the original and perturbed file need not to be released as they can be estimated as additional model parameters.

Based on a simulation study, we investigate inferential validity and disclosure protection of PRAM for both known as well as unknown PRAM matrices. We study the role of the proposed measurement error model in obtaining valid inferences and contrast it with the naive model that simply neglects the measurement error. Data utility is assessed by evaluating bias, MSE and coverage of the estimators of model parameters.

We also introduce a predictive measure of disclosure risk that takes into account the whole model to infer the original data categories and show that if the model fit is good, knowledge of the PRAM matrix actually increases the presumed protection level.

**CAUSAL INFERENCES IN SMALL AREA ESTIMATION**

*Setareh Ranjbar (University of Geneva, GSEM)*

Nicola Salvati (Università Di Pisa, DEM)

When doing impact evaluation and making causal inferences in many cases, it is important to acknowledge the heterogeneity of the treatment effects for different domains. Where certain geographic, socio- demographic, or socio-economic unplanned domains may benefit from a program/policy intervention, others may be worse off. If the domain for which we are interested in the impact, is small with regards to its sample size (or even zero in some cases), then the evaluator has entered the small area estimation dilemma. We propose new methods that allows one to estimate the area specific average treatment effects for such unplanned domains. The techniques are based on the modification of Inverse Propensity Weighting and the (robust) small area estimators. The Mean Squared Error (MSE) of the proposed predictors are analytically approximated for the situations that propensity scores are taken as known. In case of using the robust techniques a bias calibration method is also provided. These methods enable us to estimate the impact even for those small areas where all units belong to treated or control group. This is through borrowing strength from other domains that can form the control or treated group for this specific small domain. The results of our simulations, based on realistic scenarios, show a clear gain of the proposed techniques to the existing methods. This is the case where the average treatment effects are homogenous or heterogenous across all domains. By means of these methods we can also provide a map of policy impacts, that can help to better understand the outcome of an intervention and also can help to better target the treatment group(s).

**Application of small area estimation in clinical trials**

*Angela Tang*

*Amgen Inc*

In multi-regional clinical trials, consistency of a specific subgroup to overall estimation is one of interested topics (eg, countries, biomarkers). Different methods are used to evaluate the consistency in efficacy. Due to the limitation of sample size, subjects in the interested group are small and one of the application options to evaluate it.

**Small Area Estimations and Outlier Treatments in Governments Surveys**

*Bac Tran*

*U.S. Census Bureau*

The Annual Survey of Public Employment & Payroll (ASPEP) and the Annual Survey of State and Local Government (ASLGF), conducted by the U.S. Census Bureau. ASPEP provides statistics on the number of federal, state, local government civilian employees and their gross payrolls. ASLGF 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.

**Estimating the autocorrelations of sampling errors and State-Space models**

*Yingfu Xie\*, Suad Elezović, Frida Videll, Fredrik Olsson, Mickael Pettersson*

*Department for Process and Method Development, Statistics Sweden*

It is a known difficult problem to study the correlation of sampling errors in a rotating sample survey since sampling errors are non-observable. In a seminal work from 1998, Prof. Pfeffermann and co-authors proposed an interesting approach to utilize the availability of data in rotation panels in such surveys. By integrating this approach into a State-Space model framework, this and following applications have shown significant gains in areas such as small area estimation and seasonal adjustment. In our study this approach is compared with the traditional survey method to estimate the autocorrelations in the Swedish Labour Force Survey. State-Space models are then used to estimate and separate the sampling errors. Potential benefits are shown to decrease the variances of seasonally adjusted data, in particular for volatile time series.

**Hierarchical Bayes Small Area Estimation with Unknown Link Function**

*Shonosuke Sugasawa, The University of Tokyo*

Area level unmatched sampling and linking models (You and Rao, 2002) have been widely used as a model-based method for producing reliable estimates of area means. However, one practical difficulty is the specification of a linking function. In this talk, we relax the assumption of known linking function by not specifying its form and estimating it from the data. A penalized spline method is adopted for estimating the linking function and a hierarchical Bayes method of estimating area means is developed using a Markov chain Monte Carlo method for posterior computation. Results of simulation studies comparing the proposed method with a conventional approach with known linking function are presented. The proposed method is applied to data from a survey of family income and expenditure in Japan and poverty rates in Spanish provinces. This is a joint work with Tatsuya Kubokawa and J. N. K. Rao.

**Extended Rao-Yu model for small area estimation with area dependent seasonality.**

*Xichuan (Mark) Zhang*

*Australian Bureau of Statistics*

Using both time-series and cross-sectional data, Rao-Yu model incorporates autocorrelated random effects and sampling errors for small area estimation. However, seasonality is not explicitly addressed when the target and auxiliary series contain seasonality. In order to handle the seasonality, extra terms of area dependent seasonal dummies can be added to reflect seasonal variations in the applications. This complicates the estimation problem because a large number of parameters has to be estimated. Ghosh et al. (1996) solved the problem by assuming seasonal parameters were not area specific. Datta et al. (1999) assumed the seaonal parameters are independent realisation from a common probability distribution.  
 In this talk we show that both assumptions are not applicable to Australian Labour Force statistical area data. We propose to extend the Ray-Yu model to a more complex ARMA modeling rather than just an AR(1) for the area-time random effects. The model is formulated in state space form in order to use the Kalman filter to fit the model. An alternative to the extended Rao-Yu model, namely, a structural time series model based on the SUTSE model (Harvey & Chung, 2000) will also be discussed. We will show that this model is more flexible than the extended Rao-Yu model.

This work is joint with Siu-Ming Tam(Australian Bureau of Statistics) and Jan Van Den Brakel, (Statistics Netherlands and Maastricht University School of Business and Economics).