Small Area Estimation: Its Evolution in Five Decades

Malay Ghosh
University of Florida

Hansen Lecture
October 30, 2019

Outline

- Introduction
- Synthetic Estimation
- Model-Based Estimation: Area Level Models
- Model-Based Estimation: Unit Level Models
- Benchmarking
- Fixed Versus Random Effects Models
- Variable Transformation
- Miscellaneous Remarks

Introduction

- What is “small area estimation” ?
- Small area estimation is any of several statistical techniques involving estimation of parameters in small ‘sub-populations’ of interest included in a larger ‘survey’.
- The term ‘small area’ in this context generally refers to a small geographical area such as a county, census tract or a school district.
- It can also refer to a ‘small domain’ cross-classified by several demographic characteristics, such as age, sex, ethnicity etc.
- I want to emphasize that it is not just the area, but the ‘smallness’ of the targeted population within an area which constitutes the basis of small area estimation.
- For example, if a survey is targeted towards a population of interest with prescribed accuracy, the sample size in a particular subpopulation may not be adequate to generate similar accuracy.

- A domain (area) estimator is ‘direct’ if it is based only on the domain-specific sample data.
- A domain is regarded as ‘small’ if domain-specific sample is not large enough to produce estimates of desired precision.
- Domain sample size often increases with population size of the domain, but that need not always be the case.
- This requires use of ‘additional’ data, be it other administrative data not used in the original survey, or data from other related areas.
- The resulting estimates are called ‘indirect’ estimates that ‘borrow strength’ for the variable of interest from related areas and/or time periods to increase the ‘effective’ sample size.
- This is usually done through the use of models, mostly ‘explicit’, or at least ‘implicit’ that links the related areas and/or time periods.

- Historically, small area statistics have long been used.
- For example, such statistics existed in eleventh century England and seventeenth century Canada based on either census or on administrative records.
- Demographers have long been using a variety of indirect methods for small area estimation of population and other characteristics of interest in postcensal years.
- In recent years, the demand for small area statistics has greatly increased worldwide.
- The need is felt for formulating policies and programs, in the allocation of government funds and in regional planning.
- Legislative acts by national governments have created a need for small area statistics.
- A good example is SAIPE (Small area Income and Poverty Estimation) mandated by the US Legislature.

- Demand from the private sector has also increased because business decisions, particularly those related to small businesses, rely heavily on local socio-economic conditions.
- Small area estimation is of particular interest for the economics in transition in central and eastern European countries and the former Soviet Union countries.
- In the 1990's these countries have moved away from centralized decision making.
- As a result, sample surveys are now used to produce estimates for large areas as well as small areas.

- Some Examples.
- Hierarchical Bayes estimates of overweight prevalnce of adults by states using data from NHANES III. (Malec, Davis and Cao, 1999);
- Model-based county estimates of crop acreage using remote sensing satellite data as auxiliary information (Battese, Harter and Fuller, 1988).
- Income for small places (Fay and Herriott, 1979).
- Model-based county estimates of the proportion of K-12 children under poverty.
- Estimation of Median Household Income.
- Empirical and Hierarchical Bayes methods for different small area poverty measures (Molina and Rao, 2010).

Synthetic Estimation

- An estimator is called 'Synthetic' if a direct estimator for a large area covering a small area is used as an indirect estimator for that area.
- The terminology was first used by the U.S. National Center for Health Statistics.
- A strong underlying assumption is that the small area bears the same characteristic for the large area.
- For example, if $y_1, \cdots, y_m$ are the direct estimates of average income for m areas with population sizes $N_1, \cdots, N_m$, we may use the overall estimate $\bar{y}_s = \sum_{j=1}^m N_j y_j / N$ for a particular area, say, i ,where $N = \sum_{j=1}^m N_j$.
- The idea is that this synthetic estimator has less mean squared error (MSE) compared to the direct estimator y_i if the bias $\bar{y}_s - y_i$ is not too strong.
- On the other hand, a heavily biased estimator can affect the MSE as well.

- Hansen, Hurwitz and Madow (1953, pp 483-486) applied synthetic regression estimation in the context of radio listening. (Rao and Molina. *Small Area Estimation*, p. 37).
- Synthetic regression estimation of the median mumber of radio stations heard during the day in each of more than 500 counties in the US.
- The direct estimate y_i of the true (unknown) median M_i was obtained from a radio listening survey based on personal interviews.
- The estimate x_i of M_i, obtained from a mail survey was used as a single covariate in the linear regression of y_i on x_i.
- The mail survey was first conducted by sampling 1,000 families from each county area and mailing questionnaires.
- The x_i were biased due to nonresponse (about 20% response rate) and incomplete coverage, but were anticipated to have high correlation witn the M_i.

- Direct estimates y_i for a sample of 85 county areas were obtained through an intensive interview survey.
- The seletion was made by first stratifying the population county areas into 85 strata based on geographical region and available radio service type.
- Then one county was selected from each startum with probability proportional to the estimated number of families in the counties.
- A subsample of area segments was selected from each of the sampled county areas .
- Families within the selected area segments were interviewed.
- $\mathrm{Corr}(y_i, x_i) = .70$.
- For nonsampled counties, regression synthetic estimates were $\hat{M}_i = .52 + .74x_i$.

- Another example of Synthetic Estimation is due to Maria Gonzalez and Christine Hoza (JASA, 1973, pp 7-15).
- Their objective was to develop intercensal estimates of various population characteristics for small areas.
- They discussed syenthetic estimates of unemployment where the larger area is a geographic division and the small area is a county.
- Let p_{ij} denote the proprtion of labor force in county i that corresponds to cell j ($j = 1, \cdots, G$).
- Let u_j denote the corresponding unemployment rate for cell j based on the geographic division where county i belongs.
- Then the synthetic estimate of the unemployment rate for county i is given by $u_i^* = \sum_{j=1}^{G} p_{ij} u_j$.
- These authors also suggested synthetic regression estimate for unemployment rate.

- Composite Estimators: These estimators are weighted averages of Direct Estimators and Synthetic Estimators.
- The motivation is to balance the design bias of synthetic estimators and the large variability of direct estimators in a small area.
- y_{ij}: characteristic of interest for the jth unit in the ith area; $j = 1, \cdots, N_i$; $i = 1, \cdots, m$.
- $\boldsymbol{x}_{ij}$: vector of auxiliary characteristics for the jth unit in the ith local area.
- For simplicity, take x_{ij} as a scalar.
- Population means: $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i$; $\bar{X}_i = \sum_{j=1}^{N_i} x_{ij}/N_i$.
- Sampled observations: y_{ij}, $j = 1, \cdots, n_i$.
- $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$; $\bar{x}_i = \sum_{j=1}^{n_i} x_{ij}/n_i$.

- Direct Estimator (Ratio Estimator) of $\bar{Y}_i$ is $\bar{y}_i^R = (\bar{y}_i/\bar{x}_i)\bar{X}_i$.
- Ratio Synthetic Estimator of $\bar{Y}_i$ is $(\bar{y}_s/\bar{x}_s)\bar{X}_i$, where $\bar{y}_s = \sum_{i=1}^m N_i\bar{y}_i / \sum_{i=1}^m N_i$ and $\bar{x}_s = \sum_{i=1}^m N_i\bar{x}_i / \sum_{i=1}^m N_i$.
- A Composite Estimator of $\bar{Y}_i$ is $(n_i/N_i)\bar{y}_i + (1 - n_i/N_i)(\bar{y}_s/\bar{x}_s)\bar{X}_i'$, where $\bar{X}_i' = (N_i - n_i)^{-1}\sum_{j=n_i+1}^{N_i} x_{ij}/(N_i - n_i)$.
- $N_i\bar{X}_i = n_i\bar{x}_i + (N_i - n_i)\bar{X}_i'$.
- The Composite Estimator can be given a model-based justification as well. (Holt, Smith and Tomberlin, JASA, 1979, 405-410)

- A model-based justification of $(n_i/N_i)\bar{y}_i + (1 - n_i/N_i)(\bar{y}_s/\bar{x}_s)\bar{X}_i'$.
- Consider the model $y_{ij} \overset{\text{ind}}{\sim} (bx_{ij}, \sigma^2 x_{ij})$.
- Best linear unbised estimator of b is obtained by minimizing $\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - bx_{ij})^2$.
- The solution is $\hat{b} = \bar{y}_s/\bar{x}_s$.
- Now estimate $\bar{Y}_i = (\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} y_{ij})/N_i$ by $\sum_{j=1}^{n_i} y_{ij}/N_i + \hat{b}\sum_{j=n_i+1}^{N_i} x_{ij}/N_i$.
- This simplifies to the expression in the top.

Model-Based Small Area Estimation: Area Level Models

- Small area models link explicitly the sampling model with random area specific effects.
- The latter accounts for between area variation beyond that is explained by auxiliary variables.
- We classify small area models into two broad types.
- First the "area level" models that relate small area direct estimators to area-specific covariates.
- Such models are necessary if unit (or element) level data are not available.
- Second the "unit level" models that relate the unit values of a study variable to unit-specific covariates.
- Indirect estimators based on small area models will be called "model-based estimators".

- The model-based approach to small area estimation offers several advantages.
- First "optimal" estimators can be derived under the assumed model.
- Second area specific measures of variability can be associated with each estimator unlike global measures (averaged over small areas) often used with traditional indirect estimators.
- Third models can be validated from the sample data.
- Fourth, one can enetertain a variety of models depending on the nature of the response variables and the complexity of data structures.
- One of the key ongoing application of model-based estimation is the Small Area Income and Poverty Estimation (SAIPE) project of the US Bureau of the Census.

- The classic small area model is due to Fay and Herriott (JASA, 1979).
- Sampling Model: $y_i = \theta_i + e_i$, $i = 1, \ldots, m$.
  Linking Model: $\theta_i = \boldsymbol{x}_i^T \boldsymbol{b} + u_i$, $i = 1, \ldots, m$.
- Target is estimation of the θ_i, $i = 1, \ldots, m$.
- It is assumed that e_i are independent $(0, D_i)$, where the D_i are known and the u_i are iid $(0, A)$, where A is unknown.
- The asumption of known D_i can be put to question because they are, in fact, sample estimates.
- But the assumption is needed to avoid nonidentifiablity in the absence of microdata which can be used for modeling the D_i as well.
- This is evident when one writes $y_i = \boldsymbol{x}_i^T \boldsymbol{b} + u_i + e_i$.

- Some notations: $\boldsymbol{y} = (y_1, \cdots, y_m)^T$; $\boldsymbol{e} = (e_1, \cdots, e_m)^T$; $\boldsymbol{u} = (u_1, \cdots, u_m)^T$; $\boldsymbol{X} = (\boldsymbol{x}_1^T, \cdots, \boldsymbol{x}_m^T)$; $\boldsymbol{b} = (b_1, \cdots, b_p)^T$.
- We assume $\boldsymbol{X}$ has rank $p(< m)$.
- In vector notations, we write $\boldsymbol{y} = \boldsymbol{\theta} + \boldsymbol{e}$ and $\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{u}$.
- For known A, the best linear unbiased predictor (BLUP) of $\boldsymbol{\theta}$ is $(1 - B_i)y_i + B_i\boldsymbol{x}_i^T\tilde{\boldsymbol{b}}$ where $\tilde{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$, where $\boldsymbol{V} = \text{Diag}(D_1 + A, \cdots, D_m + A)$ and $B_i = D_i/(A + D_i)$.
- The BLUP is the best unbiased predictor under assumed normality.

- An alternative Bayesian formulation.
- $y_i|\theta_i \overset{\text{ind}}{\sim} \mathrm{N}(\theta_i, D_i)$; $\theta_i|\boldsymbol{b} \overset{ind}{\sim} N(\boldsymbol{x}_i^T\boldsymbol{b}, A)$.
- Then the Bayes estimator of θ_i is $(1 - B_i)y_i + B_i\boldsymbol{x}_i^T\boldsymbol{b}$, where $B_i = D_i/(A + D_i)$.
- If instead we put a uniform(R^p) prior for $\boldsymbol{b}$, the Bayes estimator of θ_i is the same as its BLUP.
- But A is unknown.
- A hierarchical Bayesian will put a prior on A as well.
- $\pi(\boldsymbol{b}, A) = 1$. (Morris, 1983, JASA).
- Otherwise, estimate A to get the resulting empirical Bayes or empirical BLUP.
- Fay and Herriott: Solve iteratively the two equations
  (i) $\tilde{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$;
  (ii) $\sum_{i=1}^m (y_i - \boldsymbol{x}_i^T\tilde{b})^2 = m - p$.

- FH iterative method may not be too convenient for analytical studies.
- Prasad and Rao (1990, JASA) suggested instead a unweighted least squares approach to estimate A.
- $\hat{\boldsymbol{b}}_L = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y}$.
- $E||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}_L||^2 = (m-p)A + \sum_{i=1}^m D_i(1-r_i)$, $r_i = \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i$, $i = 1, \cdots, m$.
- $\hat{A}_L = \max(0, \frac{||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}_L||^2 - \sum_{i=1}^m D_i(1-r_i)}{m-p})$.
- $\hat{B}_i^L = D_i / (\hat{A}_L + D_i)$.
- $\hat{\theta}^{PR} = (1 - \hat{B}_i^L) y_i + \hat{B}_i^L \tilde{b}(\hat{A}_L)$.
- $\tilde{b}(\hat{A}_L) = [\boldsymbol{X}^T \boldsymbol{V}^{-1}(\hat{A}_L)\boldsymbol{X}]^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1}(\hat{A}_L)\boldsymbol{y}$.

- Prasad and Rao also found an approximation to the mean squared eror (Bayes risk) of their EBLUP or EB estimators.
- Under the subjective prior $\theta_i \overset{\text{ind}}{\sim} \text{N}(\boldsymbol{x}_i^T \boldsymbol{b}, A)$, the Bayes estimator of θ_i is $\hat{\theta}_i^B = (1 - B_i)y_i + B_i \boldsymbol{x}_i^T \boldsymbol{b}$, $B_i = D_i/(A + D_i)$.
- Also, write $\tilde{\theta}_i^{EB}(A) = (1 - B_i)y_i + B_i \boldsymbol{x}_i^T \tilde{\boldsymbol{b}}(A)$
- $\hat{\theta}_i^{EB} \equiv \tilde{\theta}_i^{EB}(\hat{A}_L) = (1 - \hat{B}_i^L)y_i + \hat{B}_i^L \boldsymbol{x}_i^T \tilde{\boldsymbol{b}}(\hat{A}_L)$.
- $E(\hat{\theta}_i^{EB} - \theta_i)^2 =$
  $E(\hat{\theta}_i^B - \theta_i)^2 + E(\tilde{\theta}_i^{EB}(A) - \hat{\theta}_i{}^B)^2 + E(\hat{\theta}_i^{EB} - \tilde{\theta}_i^{EB}(A))^2$.
- The first term is the Bayes risk if boh $\boldsymbol{b}$ and A were known
- The second term is the additional uncertainty due to estimation of $\boldsymbol{b}$ when A is known.
- The third term is extra uncertainty due to estimation of A.

- $E(\theta_i - \hat{\theta}_i^B)^2 = D_i(1 - B_i) = g_{1i}(A)$, say;
- $E(\hat{\theta}_i^{EB}(A) - \hat{\theta}_i^B)^2 = B_i^2 \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{x}_i = g_{2i}(A)$, say;
- $E(\hat{\theta}_i^{EB} - \hat{\theta}_i^{EB}(A))^2 \doteq 2B_i^2(D_i + A)^{-1}A^2 \sum_{i=1}^m (1 - B_i)^2/m^2 = g_{3i}(A)$, say.
- This MSE approximation (or Bayes risk) is correct up to $O(m^{-1})$.
- Prasad and Rao: An estimator of this MSE correct up to $O(m^{-1})$ is $g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2g_{3i}(\hat{A})$.
- $E[g_{1i}(\hat{A})] = g_{1i}(A) - g_{3i}(A) + o(m^{-1})$.
- A further refinement to this approximation is due to Datta, Rao and Smith (Biometrika, 2005).

- An Example: Estimation of Median Income of Four Person Families.
- The U.S. Dept. of Health and Human Services provides energy assistance to low-income families.
- Eligibility for the program is determined by a formula where the most important variable is an estimate of the current median income of four-person families by states (the 50 states and the District of Columbia).
- The Bureau of the Census, by an informal agreement, provided such estimates to the HHS through a linear regression methodology since the latter part of the 1970's.
- Sample estimates of the state medians for the current year (c) as obtained from the Current Population Survey (CPS) were used as dependent variables.
- Adjusted census median (c) defined as the base year (the recentmost decennial census) census median (b) times the ratio of the BEA PCI in year (c) to year (b) was used as the dependent variable.

- Following the suggestion of Fay (1987), we used the census median for the base year (b) as a second independent variable.
- We compared the EB estimates, CPS estimates, and the Bureau of the Census estimates against the 1979 census estimates.
- The comparison was based on four criteria recommended by the panel on small area estimates of population and income set up by the committee on National Statistics.
- Average Relative Bias $= (51)^{-1}\sum_{i=1}^{51}|e_i - e_{i,TR}|/e_{i,TR}$.
- Average Squared Relative Bias $= (51)^{-1}\sum_{i=1}^{51}(e_i - e_{i,TR})^2/e_{i,TR}^2$.
- Average Absolute Bias $= (51)^{-1}\sum_{i=1}^{51}|e_i - e_{i,TR}|$.
- Average Squared Deviation $= (51)^{-1}\sum_{i=1}^{51}(e_i - e_{,TR})^2$.

Table 1. Average Relative Bias, Average Squared Relative Bias, Average Absolute Bias and Average Squared Deviations (in 100,000) of the Estimates.

	Bureau Estimate	Sample Median	EB
Aver. rel. bias	0.325	0.498	0.204
Aver. sq. rel bias	0.002	0.003	0.001
Aver. abs. bias	722.8	1090.4	450.6
Aver. sq. dev.	8.36	16.31	3.34

- Lahiri and Rao (JASA, 1995): Avoid normality assumption of the random effects, and assume instead its 8th moment in the Fay-Herriott model.
- Datta and Lahiri (Statistica Sinica, 2000): ML and REML estimation of variance components in linear mixed models.
- Das, Jiang and Rao (Annals of Statistics, 2004): Same goal for more general mixed models.
- Jiang, Lahiri and Wan: Annals of Statistics, 2002): Second order correct MSE estimation of EBLUP
- Chen and Lahiri (2002): Weighted version of Jiang-Lahiri-Wan jackknife.
- Butar and Lahiri (JSPI, 2003), Pfeffermann and Tiller (2002): Botstrap estimation of the variance components.
- Yoshimori and Lahiri (2014; Journal of Multivariate Analysis): Adjusted Maximum Likelihood.
- Fuller (Contemporary Mathematics, 1990), Booth and Hobert (JASA, 1998): conditional approach for estimating the MSE.

- General Exponential Family Model: $y_i|\theta_i$ are independent with $f(y_i|\theta_i) = \exp[y_i\theta_i - \psi(\theta_i) + h(y_i), \, i = 1, \ldots, m$.
- Bernoulli (p_i): $\theta_i = \text{logit}(p_i) = \log(p_i/(1-p_i))$.
- Poisson(λ_i): $\theta_i = \log(\lambda_i)$.
- Model the θ_i as independent $\text{N}(\boldsymbol{x}_i^T\boldsymbol{b}, A)$ and proceed.
- Alternately use beta priors for the p_i and gamma priors for the λ_i.
- Estimate the prior parameters in an empirical Bayes approach or put a prior distribution on the prior parameters in a hierarchical Bayes approach.
- Malec et al. (JASA; 1997): An example of small area estimation with binary data in National Health Interview Survey.
- Jiang and Lahiri (2001; Annals of the Institute of Statistical Mathematics) : A jackknife Method for MSE estimation.

- Jiang,Nguyen and Sunil Rao (JASA; 2011): Evaluate the performance of a BLUP or EBLUP using only the sampling model $y_i \overset{\text{ind}}{\sim} (\theta_i, D_i)$.
- Recall $B_i = D_i/(A + D_i)$.
- $E[\{(1-B_i)y_i+B_i\boldsymbol{x}_i^T\boldsymbol{b}-\theta_i\}^2|\theta_i] = (1-B_i)^2D_i+B_i^2(\theta_i-\boldsymbol{x}_i^T\boldsymbol{b})^2$.
- $E(y_i - \boldsymbol{x}_i^T\boldsymbol{b})^2 = D_i + (\theta_i - \boldsymbol{x}_i^T\boldsymbol{b})^2$.
- Unbiased estimator of the above MSE is $(1 - B_i)^2D_i - B_i^2D_i + B_i^2(y_i - \boldsymbol{x}_i^T\boldsymbol{b})^2$.
- Minimize the above wrt $\boldsymbol{b}$ and A. The resulting quantities are referred to as observed best predictive estimators of $\boldsymbol{b}$ and A.
- They refer to the resulting estimators of the θ_i as "observed best predictors".
- Use Fay-Herriott or Prasad-Rao method for estimation of $\boldsymbol{b}$ and A.

Model Based Small Area Estimation: Unit Specific Models

- Unit Specific Models: observations are available for the sampled units in the local areas.
- In addition, unit-specific auxiliary information is available for these sampled units, and possibly for the non-sampled units as well.
- m local areas. The ith local area has N_i units with a sample of size n_i.
- Sampled observations: $y_{i1}, \ldots, y_{in_i}$, $i = 1, \ldots, m$
- Model; $y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{b} + u_i + e_{ij}$, $j = 1, \ldots .N_i$, $i = 1, \ldots, m$.
- u_i's and e_{ij}'s are mutually independent with the u_i iid $(0, \sigma_u^2)$, and the e_{ij} independent $(0, \sigma^2 \psi_{ij})$.

- Nested Error Regression Model (Battese, Harter and Fuller, JASA, 1988).
- y_{ij}: area devoted to corn or soybean for the jth segment in the ith county.
- $\boldsymbol{x}_{ij} = (1, x_{ij1}, x_{ij2})^T$, where x_{ij1} denotes the no. of pixels classified as corn for the jth segment in the ith county and x_{ij2} denotes the no. of pixels classified as soybean for the jth segment in the ith county.
- $\boldsymbol{b} = (b_0, b_1, b_2)^T$ is the vector of regression coefficients.
- They took $\psi_{ij} = 1$.
- A second example (Ghosh and Rao, 1994; Statistical Science):
- y_{ij}: wages and salaries paid by the jth business firm in the ith census division in Canada.
- $\boldsymbol{x}_{ij} = (1, x_{ij})^T$, where x_{ij} denotes the gross business income of the jth business firm in the ith census division.
- Here $\psi_{ij} = x_{ij}$ was found more appropriate than the usual model involving homoscedasticity.

- Consider the Battese, Harter and Fuller (1988) model.
- In matrix notations, we write $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})^T$, $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{in_i})^T$, $i = 1, \ldots, m$.
- The model is $\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{b} + u_i \mathbf{1}_{n_i} + \boldsymbol{e}_i$, $i = 1, \ldots, m$.
- $E(\boldsymbol{y}_i) = \boldsymbol{X}_i \boldsymbol{b}$ and $\boldsymbol{V}_i = V(y_i) = \sigma_e^2 \boldsymbol{I}_{n_i} + \sigma_u^2 \boldsymbol{J}_{n_i}$.
- Also let $\bar{\boldsymbol{x}}_i = \sum_{j=1}^{n_i} \boldsymbol{x}_{ij}/n_i$ and $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$.
- The target is estimation of $\bar{\boldsymbol{X}}_i^T \boldsymbol{b} + u_i \mathbf{1}_{n_i}$, where $\bar{\boldsymbol{X}}_i = N_i^{-1} \sum_{j=1}^{N_i} \boldsymbol{x}_{ij}$, $i = 1, \ldots, m$
- For known σ_u^2 and σ_e^2, the BLUP of $\bar{\boldsymbol{x}}_i^T \boldsymbol{b} + u_i \mathbf{1}_{n_i}$ is $(1 - B_i)\boldsymbol{y}_i + B_i \bar{\boldsymbol{x}}_i^T \tilde{\boldsymbol{b}}$, where $B_i = (\sigma_e^2/n_i)/(\sigma_e^2/n_i + \sigma_u^2)$ and $\tilde{\boldsymbol{b}} = (\sum_{i=1}^m \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i)^{-1} (\sum_{i=1}^m \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{y}_i)$.
- Hence, BLUP of $\bar{\boldsymbol{X}}_i^T \boldsymbol{b} + u_i \mathbf{1}_{n_i}$ is $[(1 - B)[\bar{y}_i + (\bar{\boldsymbol{X}}_i - \bar{\boldsymbol{x}}_i)^T \tilde{\boldsymbol{b}}] + B_i \bar{\boldsymbol{X}}_i^T \tilde{\boldsymbol{b}}$.

- Method of moment estimation to get unbiased estimators of unknown σ_u^2 and σ_e^2.
- The EBLUP of $\bar{\boldsymbol{X}}_i^T \boldsymbol{b} + u_i$ is now found by substituting these estimates of σ_u^2 and σ_e^2 in the BLUP formula.
- This involves two ordinary least squares regression.
- Estimation of σ_e^2 involves moment estimation based on the sum of squares $\sum_{i=1}^m \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$.
- The next equation involves residual sum of squares by regressing $y_{ij} - \bar{y}_i$ on the $x_{ij} - \bar{x}_i$ involving those areas with sample size exceeding 1.
- A full hierarchical Bayesian approach appears in Datta and Ghosh (1991, Annals of Statistics).

- Predict areas under corn and soybeans for 12 counties in North Central Iowa based on 1978 June Enumerative Survey data as well as LANDSAT satellite data.
- The USDA Statistical Reporting Service field staff determined the area of corn and soybeans in 37 sample segments (each segment was about 250 hectares) of 12 counties in North Central Iowa by interviewing farm operators.
- Based on LANDSAT readings obtained during August and September 1978, USDA procedures were used to classify the crop cover for all pixels (a term for picture element about 0.45 hectares) in the 12 counties.
- The next table gives HB predictors (e_{HB}), the EB predictors (e_{EB}), the Battese, Harter and Fuller predictors (e_{BHF}), and the associated standard errors s_{HB}, s_{EB}, and s_{BHF} respectively for mean areas under soybeans in the 12 counties.

Table 2. The predicted hectares of soybeans and standard errors

County	e_{HB}	e_{EB}	e_{BHF}	s_{HB}	s_{EB}	s_{BHF}
Cerro Gordo	78.8	78.2	77.5	11.7	11.6	12.7
Franklin	67.1	65.9	64.8	8.2	7.5	7.8
Hamilton	94.4	94.6	95.0	11.2	11.4	12.4
Hancock	100.4	100.8	101.1	6.2	6.1	6.3
Hardin	75.4	75.1	74.9	6.5	6.4	6.6
Humboldt	81.9	80.6	79.2	10.4	9.3	10.0
Kossuth	118.2	119.2	120.2	6.6	6.0	6.2
Pocahontas	113.9	113.7	113.8	7.5	7.5	7.9
Webster	110.0	109.7	109.6	6.6	6.6	6.8
Winnebago	97.3	98.0	98.7	7.7	7.5	7.9
Worth	87.8	87.2	86.6	11.1	11.1	12.1
Wright	111.9	112.4	112.9	7.7	7.6	8.0

Benchmarking

- The model-based small area estimates, when aggregated may not equal the corresponding estiamte for the larger area.
- On the other hand the direct estimate for a larger area, for example, a national level estimate, is quite reliable.
- Moreover, matching the latter may be a good idea, for instance to protect against model failure and very often for political reasons as well.
- Suppose θ_i is the ith area mean and $\theta_T = \sum_{i=1}^{m} w_i \theta_i$ is the overall mean, where w_j may be the known proportion of units in the jth area.
- The direct estimate for θ_T is $\sum_{i=1}^{m} w_j \hat{\theta}_i$ which is usually not equal to a model based estimator.
- In order to address this, people have suggested (i) ratio adjusted estimators $\hat{\theta}_i^{RA} = \hat{\theta}_i^{EB} (\sum_{j=1}^{m} w_j \hat{\theta}_j)/(\sum_{j=1}^{m} w_j \hat{\theta}_j^{EB})$ and (ii) difference adjusted estimator $\hat{\theta}_i^{DA} = \hat{\theta}_i^{EB} + \sum_{j=1}^{m} w_j \hat{\theta}_j - \sum_{j=1}^{m} w_j \hat{\theta}_j^{EB}$.

- One criticism against such adjustments is that a common adjustment is used for all small areas regardless of their precision.
- Wang, Fuller and Qu (2008: Survey Methodology) proposed instead minimizing $\sum_{j=1}^{m} \phi_j E(e_j - \theta_j)^2$ for specified weights $\phi_j (> 0)$ subject to the constraint $\sum_{j=1}^{m} w_j e_j = \hat{\theta}_T$.
- The resulting estimate of θ_i is $\hat{\theta}_i^{WFQ} = \hat{\theta}_i^{BLUP} + \lambda_i (\sum_{j=1}^{m} w_j \hat{\theta}_j - \sum_{j=1}^{m} w_j \hat{\theta}_j^{BLUP})$, where $\lambda_i = w_i \phi_i^{-1} / (\sum_{j=1}^{m} w_j^2 \phi_j^{-1})$.
- Datta, Ghosh, Steorts and Maples (2011) took instead a general Bayesian approach and minimized instead $\sum_{j=1}^{m} \phi_j [E(e_j - \theta_j)^2 | \boldsymbol{\theta}]$ subject to $\sum_{j=1}^{m} w_j e_j = \hat{\theta}_T$ and obtained the estimator $\hat{\theta}_i^{AB} = \hat{\theta}_i^{B} + \lambda_i (\sum_{j=1}^{m} w_j \hat{\theta}_j - \sum_{j=1}^{m} w_j \hat{\theta}_j^{B})$, with the same λ_i.

- The approach of Datta, Ghosh, Steorts and Maples extends readily to multiple benchmarking constraints.
- In a frequentist context. Bell, Datta and Ghosh (Biometrika, 2013) extended the work of Wang, Fuller and Qu to multiple benchmarking constraints.
- There are situations also when one needs two-stage benchmarking.
- An example is the cash rent estimates of the Natural Agricultural Statistics Service (NASS) where one needs the dual control of matching the aggregate of county level cash rent estimates to the corresponding agricultural district (comprising of several counties) level estimates, and the aggregate of the agricultural district level estimates to the final state level estimate.
- Berg, Cecere, Erciulescu and Ghosh (2019; Survey Methodology) adopted an approach of Ghosh and Steorts (2013; Test) to address the NASS problem.

- Second order unbiased MSE estimators are not typically available for ratio adjusted benchmaked estimators.
- In contrast, second order unbiased MSE estimators are available for difference adjusted benchmaked estimators, namely, $\hat{\theta}_i^{DB} = \hat{\theta}_i^{EB} + (\sum_{j=1}^m w_j \hat{\theta}_j - \sum_{j=1}^m w_j \hat{\theta}_j^{EB})$.
- Steorts and Ghosh (2013; Statistica Sinica) have shown that $\text{MSE}(\hat{\theta}_i^{DB}) = \text{MSE}(\hat{\theta}_i^{EB}) + g_4(A) + o(m^{-1})$, where $\text{MSE}(\hat{\theta}_i^{EB})$ is the same as the one given in Prasad and Rao (1990; JASA).
- A second order unbiased estimator of $\text{MSE}(\hat{\theta}_i^{DB})$ is obtained by adding $g_4(\hat{A})$ to the Prasad-Rao second order unbiased estimator of $\text{MSE}(\hat{\theta}_i^{EB})$.

- There are two available approaches for self benchmarking which do not require any adjustment to the EBLUP estimators.
- The first, proposed in You and Rao (2002: Canadian Journal of Statistics), replaces the estimator $\hat{\boldsymbol{b}}$ in EBLUP by an estimator which depends both on $\hat{\boldsymbol{b}}$ as well as the weights w_i.
- This changes the MSE calculation.
- Recall the Prasad-Rao MSE of EBLUP given by $\text{MSE}(\hat{\theta}_i^{EB}) = g_{1i} + g_{2i} + g_{3i}$, where $g_{1i} = D_i(1 - B_i)$, $g_{2i} = B_i^2 \boldsymbol{x}_i^T (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{x}_i$ and $g_{3i} = 2D_i^2 (A + D_i)^{-3} m^{-2} \{\sum_{j=1}^m (A + D_j)^2\}$.
- For the Benchmarked EBLUP, g_{2i} changes.
- The second approach is by Wang, Fuller and Qu (2008; Survey Methodology) which uses an augmented model with new covariates $(\boldsymbol{x}_i, w_i D_i)$.
- This second approach was extended by Bell, Datta and Ghosh (2013; Biometrika) to accommodate multiple benchmarking constraints.

Fixed Versus Random Effects

- A different but equally pertinent issue has recently surfaced in the small area literature.
- This concerns the need for random effects in all areas, or whether even fixed effects models would be adequate for certain areas ?
- Datta, Hall and Mandal (2011; JASA) were the first to address this problem.
- They suggested essentially a preliminary test-based approach, testing the null hypothesis that the common random effect variance was zero.
- Used a fixed or a random effects model for small area estimation based on acceptance or rejection of the null hypothesis.
- This amounted to use of synthetic or regression estimates of small area means upon acceptance of the null hypothesis, and composite estimates which were weighted averages of direct and regression estimators otherwise.

- The DHM procedure works well when the number of small areas is moderately large, but not necessarily when the number of small areas is very large.
- In such situations, the null hypothesis of no random effects is very likely to be rejected.
- This is primarily due to a few large residuals causing significant departure of direct estimates from the regression estimates.
- This was realized by Datta and Mandal (2015;JASA) who proposed instead a mixture model for random effects with "spike and slab priors".
- These priors put a positive mass at zero resulting in a spike at zero, while for the slab part, they used a normal distribution with zero means and a common unknown variance across all small areas.
- Their approach amounts to taking $\delta_i u_i$ instead of u_i for random effects where the δ_i and the u_i are independent with δ_i iid Bernoulli(γ) and u_i iid $\mathrm{N}(0, \sigma_u^2)$.

- In contrast to the spike and slab priors of Datta and Mandal, Tang, Ghosh, Ha and Sedransk (2018; JASA) considered a different class of priors which meets the same objective. as spike and slab priors, but uses instead absolutely continuous priors.
- Moreover, these priors allow different variance components for different small areas, and intend to capture local small area effects better than the priors of Datta and Mandal who considered prior variances to be either zero or else common across all small areas.
- This seems to be particularly useful when the number of small areas is very large, for example, when one considers more than 3000 counties of the US, where one expects a wide variation in the county effects.
- The proposed class of priors, is usually referred to as "global-local shrinkage priors".

- References: Carvalho, Polson and Scott (2010; Biometrika) Polson and Scott (2010; Bayesian Statistics).
- These priors are essentially scale mixtures of normals.
- Goal: capture potential "sparsity", which means lack of significant contribution by many of the random effects, by assigning large probabilities to random effects close to zero.
- But also assign non-trivial probabilities to random effects which differ significantly from zero.
- This is achieved by employing two levels of parameters to express prior variances of random effects.
- The first, the "local shrinkage parameters", act at individual levels, while the other, the "global shrinkage parameter" is common for all random effects.
- Fay and Herriott: only one global parameter; Datta and Mandal: the variance parameter of random effects is either zero or common across all small areas.

- Specifically, the radom effects u_i have independent $\mathrm{N}(0, \lambda_i^2 A)$ priors.
- While the global parameter A tries to cause an overall shrinking effect, the local shrinkage parameters λ_i^2 are useful in controlling the degree of shrinkage at the local level.
- If the mixing density corresponding to local shrinkage parameters is appropriately heavy-tailed, the large random effects are almost left unshrunk.
- The class of "global-local" shrinkage priors includes the three parameter beta (TPBN) priors (Armagon, Clyde and Dunson, 2011; Advances in Neural Information Processing Systems), Generalized Double Pareto priors (Armagon, Dunson and Lee, 2012; Statistica Sinica).
- TPBN includes the now famous horseshoe (HS) priors (Scott and Berger, 2010; Annals of Statistics) and the normal-exponential-gamma priors (Griffin and Brown, 2005)

- Goal: Estimate 5-year (2007–2011) county-level overall poverty ratios.
- There are 3,141 counties in the data set.
- Covariates: foodstamp participation rates (0.81).
- Estimated poverty ratios are between 3.3% (Borden County, TX) and 47.9% (Shannon County, SD). The median is 14.7%.
- In Mississippi, Georgia, Alabama and New Mexico, 55%+ counties have poverty rates > the third quartile (18.9%).
- In New Hampshire, Connecticut, Rhode Island, Wyoming, Hawaii and New Jersey, 70%+ counties have poverty rates < the first quartile (11.1%).

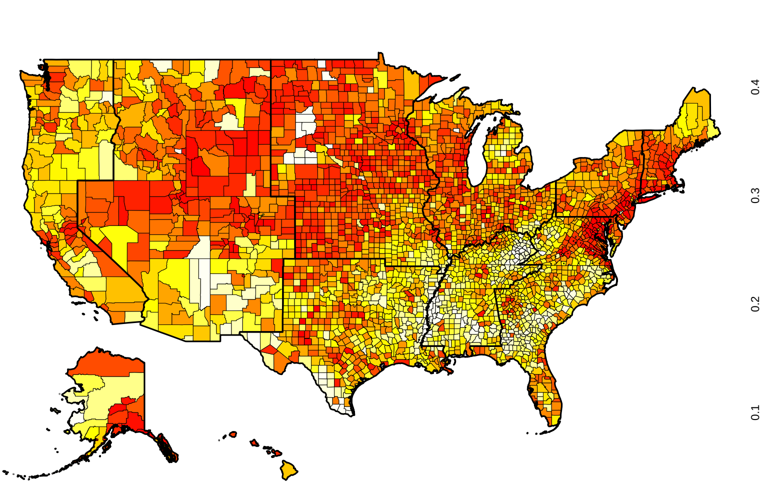

Figure: Map of posterior means of θ's.

Variable Transformation

- Often the normality assumption can be justified only after transformation of the original data.
- Then one performs the analysis based on the transformed data, but transform back properly to the original scale to arrive at the final conclusion.
- One common example is transformation of skewed positive data, for example, income data where log transformation gets a closer normal approximation.
- Slud and Maiti (2006; JRSS B) and Ghosh and Kubokawa (2015; Biometrika) took this approach, providing final results for the back-transformed original data.

- For example, consider a multiplicative model $y_i = \phi_i \eta_i$ with $z_i = \log(y_i)$, $\theta_i = \log(\phi_i)$ and $e_i = \log(\eta_i)$.
- Fay-Herriott (1979; JASA) model (i) $z_i|\theta_i \overset{\text{ind}}{\sim} \mathrm{N}(\theta_i, D_i)$ and (ii) $\theta_i \overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{x}_i^T \boldsymbol{\beta}, A)$.
- θ_i has the $\mathrm{N}(\hat{\theta}_i^B, D_i(1-B_i))$ posterior, $\hat{\theta}_i^B = (1-B_i)z_i + B_i \boldsymbol{x}_i^T \boldsymbol{\beta}$, $B_i = D_i/(A+D_i)$.
- Now $E(\phi_i|z_i) = E[\exp(\theta_i)|z_i] = \exp[\hat{\theta}_i^B + (1/2)D_i(1-B_i)]$.

- Another interesting example will be variance stabilizing transformation.
- For example suppose $y_i \overset{\text{ind}}{\sim} \text{Bin}(n_i, p_i)$.
- Arc sin transformation $z_i = \sin^{-1}((2y_i/n_i) - 1)$.
- One can start with $z_i \overset{\text{ind}}{\sim} \text{N}(\theta_i, 1/n_i)$, where $\theta_i = \sin^{-1}(2p_i - 1)$.
- Back transformation: $p_i = (1/2)[1 + \sin(\theta_i)]$.
- Another is the Poisson model for count data.
- $y_i \overset{\text{ind}}{\sim} \text{Poisson}(\lambda_i)$.
- Then one models $z_i = y_i^{1/2}$ as independent $\text{N}(\theta_i, 1/4)$ where where $\theta_i = \lambda_i^{1/2}$.
- An added advantage here is that the assumption of known D_i which is really untrue, can be avoided.

MIscellaneous Other Topics

- Design consistency of small area estimators.
- Spatial and Space-Time Models.
- Measurement errors in the covariates.
- Poverty counts for small areas.
- Empirical Bayes confidence intervals.
- Robust small area estimation.
- Misspecification of linking models.
- Informative sampling.
- Constrained small area estimation.
- Disease Mapping
- Etc, Etc., Etc.