Big Data For Small Area Estimation

Partha Lahiri

Joint Program in Survey Methodology
University of Maryland, College Park, USA

UPM, Malaysia

August 13-14, 2018

Introduction

A Historical Note

Brackstone (1987)

- 11th century England and 17th century Canada
  - Based on census or administrative records
- Recent three decades
  - Increasing demand for small area statistics, due to growing use in formulating policies and programs in the allocation of government funds and in regional planning

What is a Small Area or Domain?

A subpopulation of interest with meager or no survey data.

Examples:

- In a nationwide survey, cells obtained by finer classification of age-group, race, gender even at the national level (small domains).
- In NHANSE III, a majority of US states do not have sample (small area).
- Even for a very large scale sample survey (e.g., American Community Survey), we can easily cite examples of small domains or areas (e.g., small counties or school districts).
- Number of job vacancies by industry x state

Direct Estimation

- A direct small area estimator uses y, the variable of interest, only from the sampled units in the small area using the primary source of information.
- The estimator may or may not use auxiliary variable(s).
- If the estimator uses auxiliary variable(s), it may or may not use auxiliary information from other domains.
- Estimators are typically p-unbiased or approximately p-unbiased with respect to the randomization that generates survey data.
- Direct estimators are usually design-consistent for large domain sample size. In small area estimation domain sample sizes are typically small and thus design-consistency property does not have much appeal.
- Ref: Cochran (1977), Lohr (1999), Särndal et al. (1992).

Two Simple SAE Settings: Planned Domains

Planned domains are domains for which samples have been planned. Thus we can take such domains as strata.

- U: a finite population with m strata U_i $(i = 1, \cdots, m)$.
- y_{ij}: value of the jth unit in the ith stratum $(i = 1, \cdots, m;\ j = 1, \cdots, N_i)$.
- Parameter of interest: $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$, $(i = 1, \cdots, m)$, where N_i's are known and $N_T = \sum_{i=1}^{m} N_i$.
- n_T: total sample size allocated to these strata using an allocation scheme.
- n_i: fixed sample size for the ith area $(i = 1, \cdots, m)$. Thus, $n_T = \sum_{j=1}^{m} n_i$.
- Although the total sample size n_T is typically large in a sample survey, n_i could be small for some or all of the areas.

A Planned Domain Example: SRS within Each Domain

- A_i: sample of units from the ith domain (stratum) $(i = 1, \cdots, m)$.
- The usual Horvitz-Thompson (HT) estimator of $\bar{Y}_i$:

$$\bar{y}_i = \frac{1}{n_i} \sum_{j \in A_i} y_{ij}.$$

- True design-based variance: $V_p(\bar{y}_i) = (1 - f_i)\frac{S_i^2}{n_i}$, where $S_i^2 = (N_i - 1)^{-1} \sum_{j=1}^{N_i} (y_{ij} - \bar{Y}_i)^2$ and $f_i = \frac{n_i}{N_i}$.
- The magnitude of the variance depends on three factors: f_i, S_i^2, and n_i.
- We have a small area situation in the area i if $V_p(\bar{y}_i)$ is larger than the specified requirement. When can we have a small area situation?
- If $n_i > 1$, we can estimate $V_p(\bar{y}_i)$ by $v_i = (1 - f_i)\frac{s_i^2}{n_i}$, where $s_i^2 = (n_i - 1)^{-1} \sum_{j \in A_i} (y_{ij} - \bar{y}_i)^2$. Is v_i design-unbiased? What can you say about $V_p(v_i)$?
- Write down the formulae for binary data.

Two Simple SAE Settings: Unplanned Domains

Unplanned domains are domains that were not identified at the design stage so sample sizes cannot be controlled. Consider a SRS of size n_T from U.

- Are n_i are fixed or random?
- Is $\bar{y}_i$ an unbiased estimator of $\bar{Y}_i$? What can be said for a general sample design?
- A variance estimator:

$$\tilde{v}_i \approx (1-f)\frac{s_i^2}{n_{i;exp}},$$

  where $f = n_T/N_T$, and $n_{i;exp} = n_T\frac{N_i}{N_T}$, the expected sample size for area i. What can be said about this variance estimator?
- An alternative variance estimator: v_i. What can be said about this variance estimator?

An Implicit Working Superpopulation Model

$$\mathrm{E}[y_{ij}] = \theta_i, \qquad V[y_{ij}] = \sigma_i^2, \qquad \mathrm{Cov}[y_{ij},\ y_{i'j'}] = 0,$$

for $(i, i' = 1, \cdots, m;\ j, j' = 1, \cdots, N_i,\ j \neq j')$. Under the above superpopulation model, we can show that

- $\bar{y}_i$ is model-unbiased with prediction variance $V(\bar{y}_i - \bar{Y}_i) = (1 - f_i)\frac{\sigma_i^2}{n_i}$.
- A model-unbiased estimator of the prediction variance is v_i.
- Under normality of y_{ij}, we have $\frac{(n_i - 1)s_i^2}{\sigma_i^2} \sim \chi^2_{n_i - 1}$ and thus

$$V(v_i) = (1 - f_i)^2 \frac{2\sigma_i^4}{n_i^2(n_i - 1)}.$$

- Thus, n_i being small, we expect $V(v_i)$ to be large unless f_i is close to 1 and/or σ_i^4 is small. Find the design variance of v_i.

A Frequently Asked Question

Question:

- Standard survey-weighted estimators are commonly used by survey organizations.
- When do we decide to switch to SAE?

How do we respond to such an apparently simple question?

Two Possible Natural Answers:

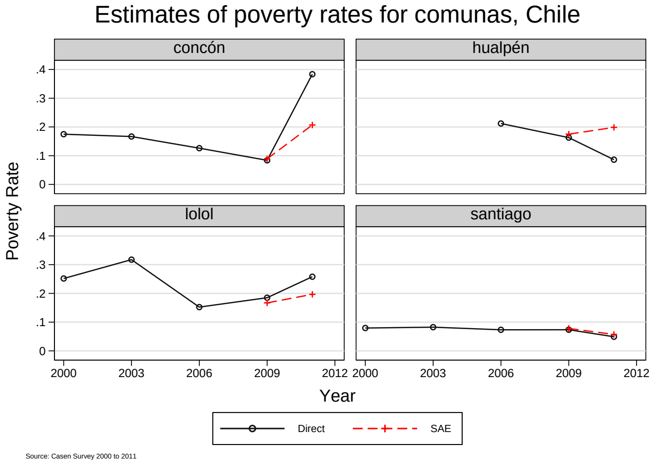
- Go for SAE methods if estimates of CVs or standard errors of standard survey-weighted estimates are high.
- Go for design-consistent model-based estimates for all situations.

One can argue against each of the above answers

Poverty mapping: the Chilean Case

- High poverty rates can work favorably to a Chilean municipality in terms of securing more funds from the Chilean central government.
- Consider the following situation. For a given small municipality, poverty rate for the current year turns out to be high by standard design-based method.
- How do we convince the mayor of that municipality to go for a statistically efficient SAE method that yields lower poverty rate?
- Can repeated survey data help?

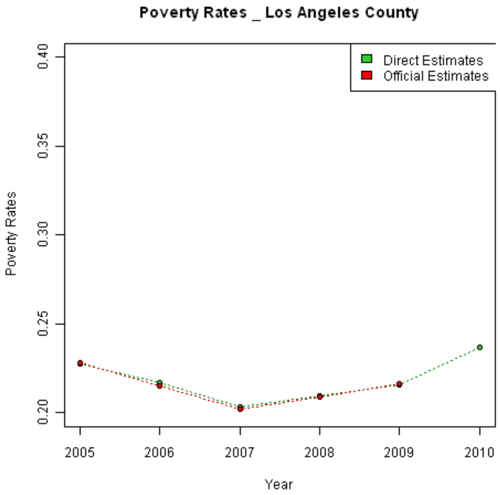
Plots of Survey-Weighted Poverty Rates and SAE for Selected Comunas (drawn by Carolina Casas-Cordero)



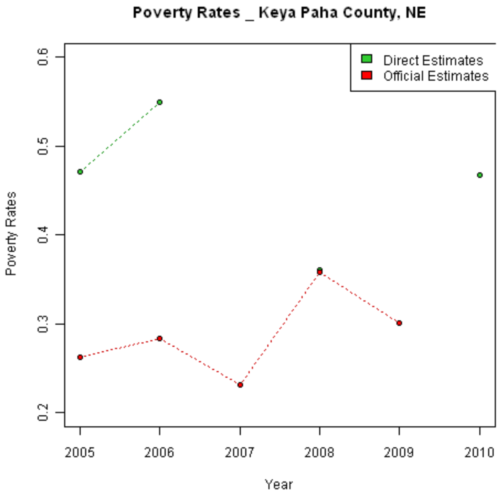
Example: Small Area Income and Poverty Estimates (SAIPE)

- The primary source of the data for this problem is the American Community Survey (ACS).
- The direct survey estimate of poverty rate is a weighted average of poverty status of the sampled respondents for the group and year of interest.
- The weight for a sampled respondent can be viewed as the number of population units the sampled respondent represents.
- The official Small Area Income and Poverty Estimates (SAIPE) that the U.S. Census Bureau routinely produces uses model-based method that combine ACS with various administrative data.
- Next few figures compare direct survey estimates and their standard errors with the official estimates over different years for one big county (Los Angeles county, CA) and two small counties (Keya Paha county, NE and Lincoln county, SD).

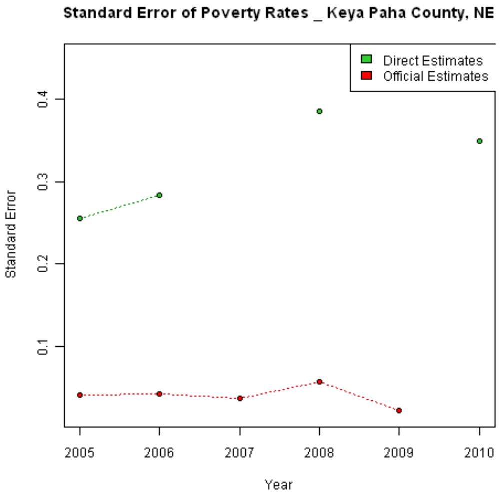
Plots of Survey-Weighted Poverty Rates and SAE for a Small County (drawn by Sam Hawala)



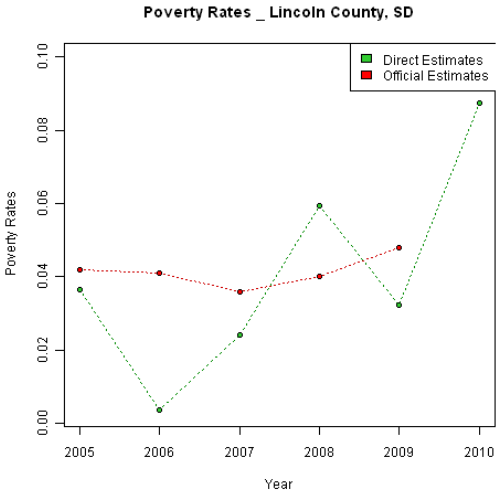
Plots of Survey-Weighted Poverty Rates and SAE for a Small County (drawn by Sam Hawala)



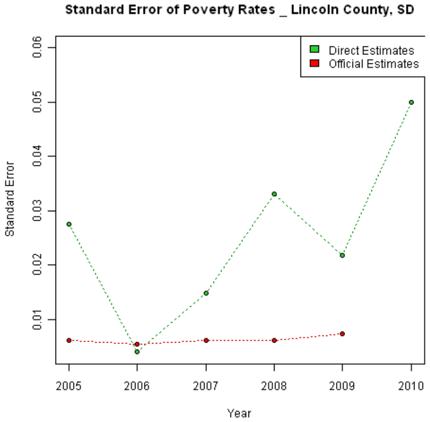
Plots of Survey-Weighted Poverty Rates and SAE for a Small County (drawn by Sam Hawala)



Plots of Survey-Weighted Poverty Rates and SAE for a Small County (drawn by Sam Hawala)



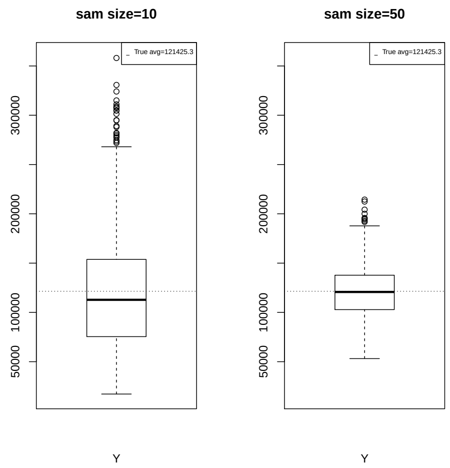
Plots of Estimated SE Survey-Weighted Poverty Rates and SAE for a Small County (drawn by Sam Hawala)



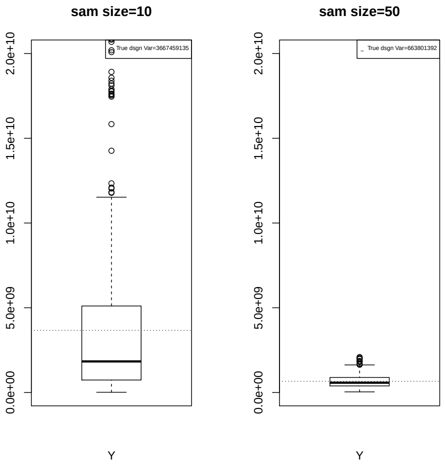
Simulation from the Australian Beef Farm Data

- Finite population: $N = 431$ farms
- Variable: income from beef
- Simulate several samples of size n from the finite population.
- For a given variable, sample means from several simulated samples are displayed in the box plots and compared with the corresponding true value for $n = 10,\ 50$.
- Sample means and the associated variance estimates, though unbiased, exhibit high variability for $n = 10$. Variability decreases as we increase n.

Box Plots of Estimates: Income from Beef

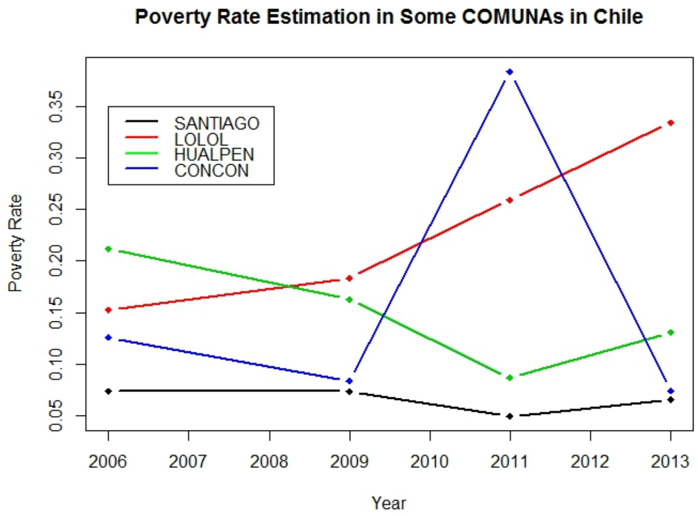


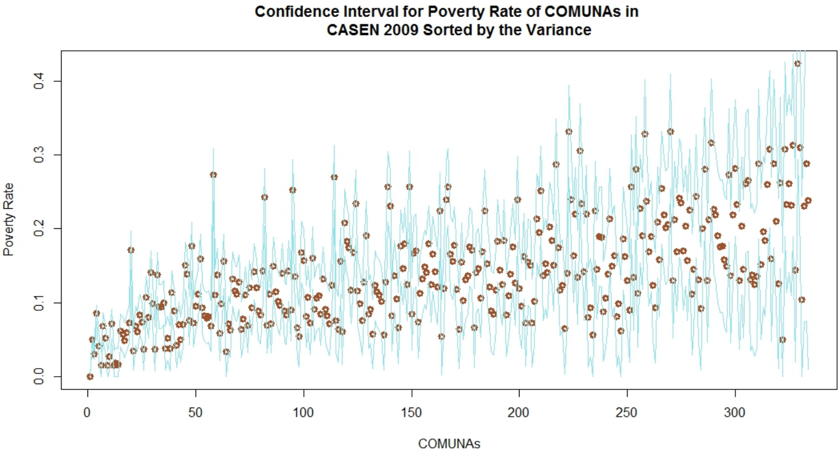
Box Plots of Variance Estimates: Income from Beef



Poverty Mapping in Chile

- The poverty rate (also known as head count Index) is the proportion of households with income below the poverty threshold or poverty line.
- The per-capita income is the ratio of the total household income and the household size. National and regional estimates of per-capita income are produced using the CASEN survey and standard design-based methods.
- The official national-level poverty rate estimates are published every two or three years following the release of each CASEN data.


Poverty Rate Estimation in Some COMUNAs in Chile
SANTIAGO
LOLOL
HUALPEN
CONCON
Poverty Rate
0.05
0.10
0.15
0.20
0.25
0.30
0.35
2006
2007
2008
2009
2010
2011
2012
2013
Year


Confidence Interval for Poverty Rate of COMUNAs in
CASEN 2009 Sorted by the Variance
Poverty Rate
0.0
0.1
0.2
0.3
0.4
0
50
100
150
200
250
300
COMUNAs

A few illustrative examples

Example: Estimation of proportion

Consider the problem of estimating finite population proportion of some attribute using a SRS.

- Suppose we have just one sample and value is 1. The standard direct unbiased estimate is then 1 with a direct standard error estimate 0.
- Suppose you now have a sample of size 2 and both observations are 1. In this case also the direct unbiased estimate is 1 with a direct standard error estimate 0.
- Suppose we have a sample of size 2 – one observation is 1 and the other is 0. Then the direct unbiased estimate is 0.5. In this case the direct standard error estimate, margin of error and confidence intervals are approximately 0.35, 0.7, and $(-0.2, 1.2)$, respectively.

Example: Paper submission

Ref: Carlin, B. and Louis, T.A. (2009), *Bayesian Methods for Data Analysis*, A Chapman & Hall Book.

Your first paper submitted to a journal with a historical acceptance rate of 30% is accepted.

What is the chance that your second paper of similar quality will be accepted in the same journal?

Example: Missing data

89	92	99
100	*	110
109	105	108

- The above table provides an array of death rates per 10,000 persons, perhaps arranged geographically or cross-tabulated by clinic and age-group.
- Without any direct information on the missing value *, does an estimate 200 seem reasonable?
- How do we incorporate the following information?
    - We collect data in the missing cell and we get 2 deaths in a population of 100 so that a direct estimate is $200 = 10000 \times 2/100$.
    - We collect more data and we have 20 deaths in a population of 1000.

Addressing SAE Issues at the Design Stage

In general, it helps small area estimation if this is considered as one of the several factors before collecting data.

References: Singh, Gambino and Mantel (1994), Marker (2001).

Design Issues

- Large surveys usually do not consider desired precision at small domain levels at the design stage.
- "...handling of this growing challenge...at the estimation stage should be viewed as a last resort." Singh et al. (1994)
- Need to meet SAE needs in planning, sample design and estimation stages.
- Planning depends on how well the small areas are identified in advance so that they can be treated as planned domains. But, *The client will always require more than is specified at the design stage.* (Fuller, 1999, p. 344).

Design Options: Stratification

- We can control sample sizes for planned domains by treating them design strata.
- If there are a large number of planned domains, it may not be possible to consider all planned domains as strata. One may apply some grouping idea in such cases.
- Use a large number of smaller strata. But, this increases the costs and so one needs to have some balance between costs and efficiency of the estimators.
- Given a fixed budget, a large number of strata will reduce sample sizes per stratum. But this strategy should help unplanned domain estimation since the number of unplanned areas with some samples is likely to increase.

Design Options: Degree of Clustering

- Minimize clustering whenever possible.
- Large surveys often use multi-stage design and are often highly clustered.
- Unplanned small domains may not have been sampled.
- Important factors: choice of frame, size of strata.

Design Options: Sample Allocation

- Compromise Allocation (Singh et al., 1994)
  - Reallocate sample from larger planned domains to smaller planned domains.
  - Small reduction in sample size for large domains usually has little effect.
  - Small increases in small domains may have a large effect on reliability.
  - **Canadian Labor Force Survey** Two-Step Allocation: 42,000 Households for national and province level estimates, 17,000 for Unemployment Insurance (UI) region level estimates.
    Effects of Reallocation on Areas:
    UI region (worst case): CV decreased from 17.7 to 9.4
    Provincial Level (Ontario): CV increased from 2.8 to 3.4

Design Options: Sample Allocation

- Minimize a weighted sum of sampling variances of direct small area estimators subject to fixed overall sample size. Ref: Longford (2006)
- Costa et al. (2004): a convex combination of proportional allocation and equal allocation.
- Choudhry, Rao and Hidiroglou (2010) used a non-linear programming (NLP) method to derive the "optimal" sample size allocation that minimizes the total sample size subject to specified tolerances on the coefficients of variation of the domain estimators and the associated aggregate estimator.

Other Design Options

- Integration of surveys [e.g, European Community Household Panel Survey (ECHP)]
- Multiple frame surveys Hartley (1974) [e.g., Canadian Community Health Survey (CCHS)]
- Repeated surveys [e.g., American Community Survey (ACS)],

Use of Auxiliary Variables

Two uses:

- Survey design
- Estimation

An Example:

- SRS within each small area and one auxiliary variable x for which we know both the sample mean and population mean for every area.
- Ratio estimator:

$$\hat{\bar{Y}}_{i;R} = \frac{\bar{y}_i}{\bar{x}_i}\bar{X}_i,\ i = 1, \cdots, m,$$

  where $\bar{y}_i$, $\bar{x}_i$ and $\bar{X}_i$ are the samples means of y and x and population mean of x for area i, respectively.
- For large n_i, $\hat{\bar{Y}}_{i;R}$ is approximately design-unbiased.
- The order of bias is $O(n_i^{-1})$.

Use of Auxiliary Variables

- The approximate true design-variance is given by:

$$V_p(\hat{\bar{Y}}_{i;R}) \approx (1 - f_i)\frac{S^2_{i;E}}{n_i},$$

  where $S^2_{i;E} = (N_i - 1)^{-1}\sum_{j=1}^{N_i}(E_{ij} - \bar{E}_i)^2$, the finite population variance of the residuals:
  $E_{ij} = y_{ij} - R_i x_{ij}$, $(i = 1, \cdots, m;\ j = 1, \cdots, N_i)$, and
  $R_i = \bar{Y}_i/\bar{X}_i$, $\bar{Y}_i$ and $\bar{X}_i$ being the finite population means of y and x, respectively.
- For a biased estimator, design-based mean squared error (MSE) could be a reasonable uncertainty measure since it incorporates both variance and bias:
  $\text{MSE} = \text{Variance} + (\text{Bias})^2$.
- Note that variance contributes more to MSE than the bias does (why?).

Use of Auxiliary Variables

- For large n_i, we can reduce the variance at the expense of slight increase of bias if a line passing through the origin fits the entire finite population well. However, for small n_i, both bias and variance could be substantial. HW: Device a simulation study using the beef data.
- An design-based estimator of $V_p(\hat{\bar{Y}}_{i;R})$ is given by

$$v_p(\hat{\bar{Y}}_{i;R}) = (1 - f_i)\frac{s_{i;e}^2}{n_i},$$

where $s_{i;e}^2 = (n_i - 1)^{-1}\sum_{j=1}^{n_i}(e_{ij} - \bar{e}_i)^2$, the sample variance of the observed residuals:
$e_{ij} = y_{ij} - \hat{R}_i x_{ij}, \ (i = 1, \cdots, m; \ j = 1, \cdots, n_i)$, and
$\hat{R}_i = \bar{y}_i/\bar{x}_i$, respectively.
- An implicit model that could justify the above ratio estimator is:

$$y_{ij} = \beta_i x_{ij} + \epsilon_{ij}, \ \ (i = 1, \cdots, m; \ j = 1, \cdots, N_i),$$

where β_i is fixed area specific slope and $\epsilon_{ij} \sim (0, \sigma_i^2)$.

Borrowing Strength

- Relevant Source of Information
  - Census data
  - Administrative records
  - Related surveys
- Method of Combining Information
  - Choices of good small area models
  - Use of a good statistical methodology

Big Data

Problem 1: BIGDATA from Administrative Records

Estimation of income and poverty statistics for the administration of federal programs and the allocation of federal funds to local jurisdictions.

- Internal Revenue Service Data
- Supplemental Nutrition Assistance Program (SNAP) data

Problem 2: Remote Sensing BIGDATA

Estimation of crop acreage, crop production, crop yield for the purpose of local agricultural decision making, payments to farmers if crop yields are below certain levels.

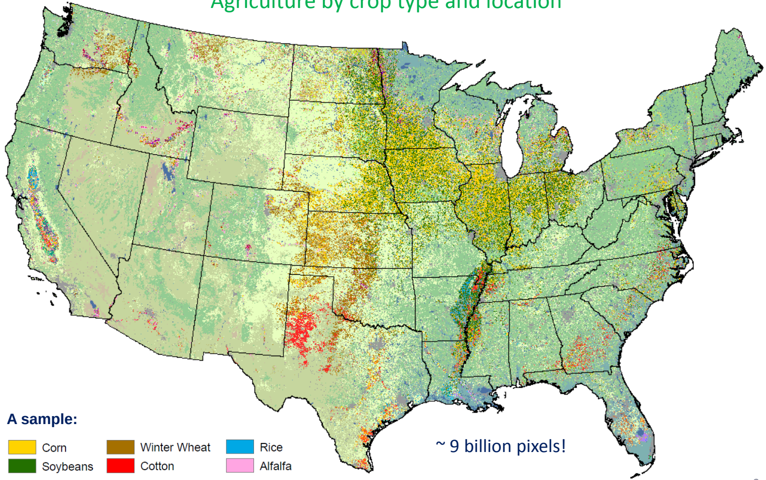
- Can earth resources satellite data provide useful ancillary data source for county estimates of crop acreage?
- Satellite information is recorded for *pixels* (a term for *picture elements*). A pixel is about .45 hectares;
- Based on satellite readings in early Fall, it is possible to classify the crop cover all pixels. This generates big data.

A Quote from Bellow et al.

The polar-orbiting Landsat satellites contain a multi-spectral scanner (MSS) that measures reflected energy in four bands of the electromagnetic spectrum for an area of just under one acre. The spectral bands were selected to be responsive to vegetation characteristics. In addition to the MSS sensor, Landsats IV and V have a Thematic Mapper (TM) sensor which measures seven energy bands and has increased spatial resolution. The large area (185 by 170 km) and repeat (16 day per satellite) coverage of these satellites opened new areas of remote sensing research: large area crop inventories, crop yields, land cover mapping, area frame stratification, and small area crop cover estimation.

Cropland Data Layer

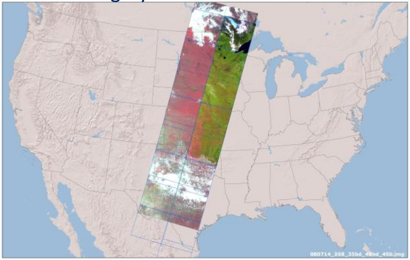
Agriculture by crop type and location



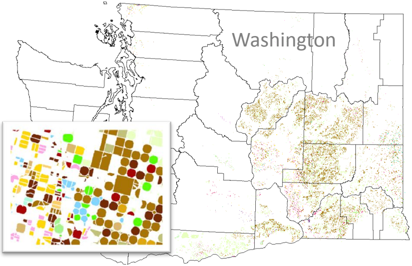
2014 Cropland Data Layer Inputs

Satellite Imagery – Deimos & UK2

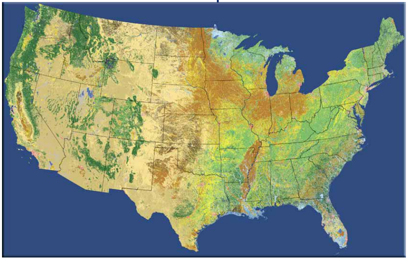
Satellite Imagery – Landsat 8



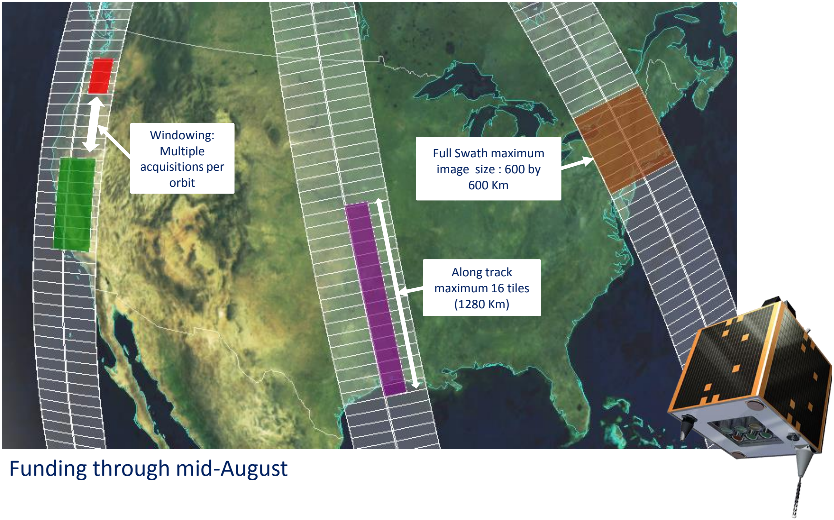
Farm Service Agency: Common Land Unit



2011 NLCD & Derivative products



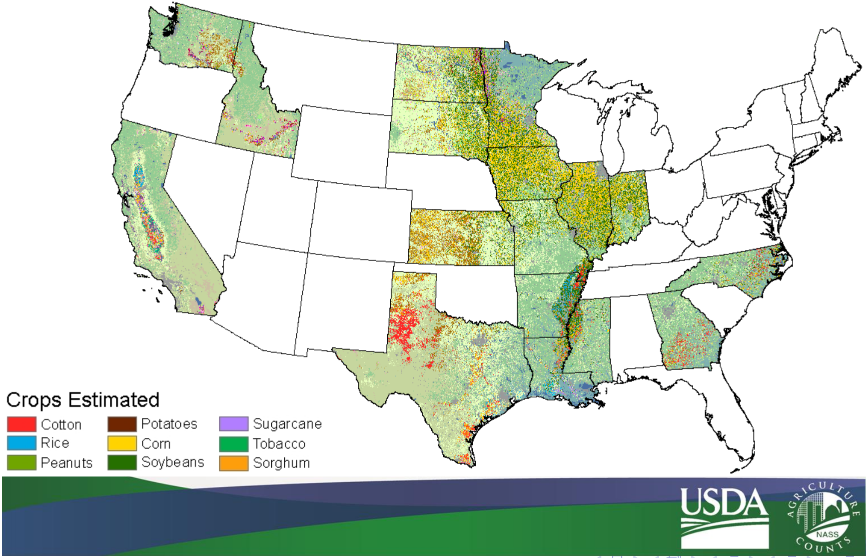
2014 Deimos-1/UK2 Satellite Tasking



Funding through mid-August

September

17 States Classified
9 Crops Estimated
Imagery from April - August

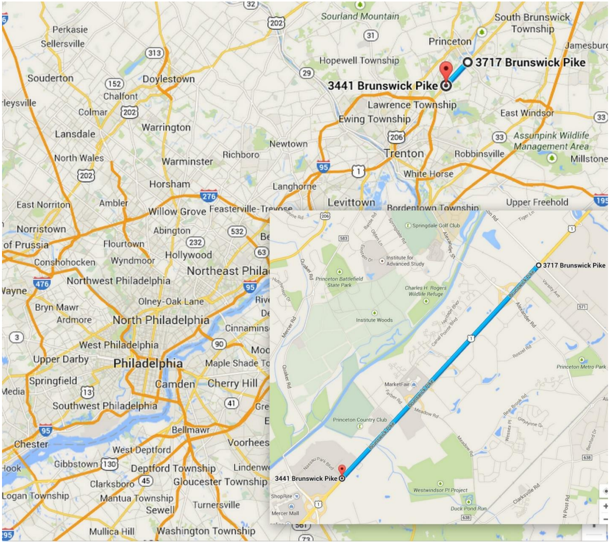


Problem 3: Vehicle Probe Project (VPP) BIGDATA

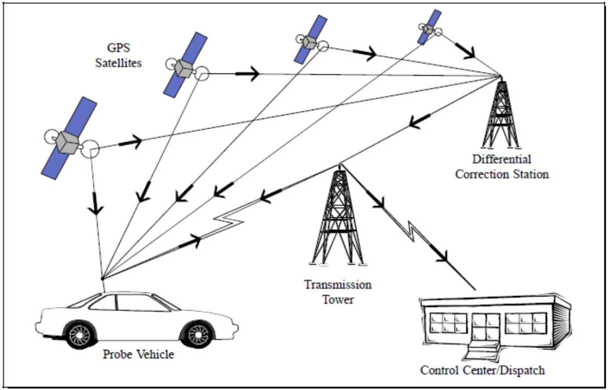
Estimation of transportation related variables such as purpose of the trip (work, shopping, social, etc.), means of transportation (car, walk, bus, subway, etc.), travel time of trip to assist transportation planners and policy makers who need comprehensive data on travel and transportation patterns.

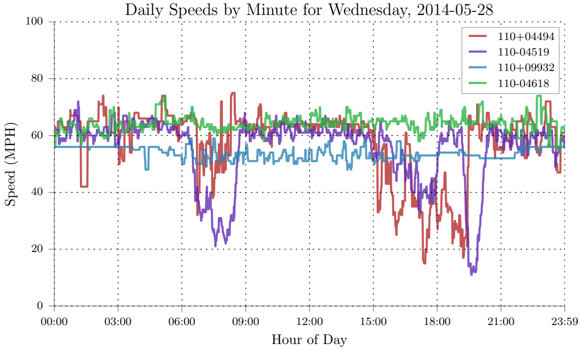
- Currently, the VPP contractually reports traffic conditions on over 7,000 miles of freeways and 32,000 miles of arterials.
- Original goal: to enable a wide-variety of transportation operations and planning applications that require a high-quality data source.
- Data contains travel time, speed, historic speed, etc. for different road segments called Traffic Message Channels (TMC).
- Applications include congestion management systems, traveler information systems, travel-time on changeable message signs.
- If data for a whole year, for all 12,295 TMC segments in Maryland were to be downloaded, the estimated number of records is 6.46 billion. The physical disk size of this data is estimated to be 375GB.

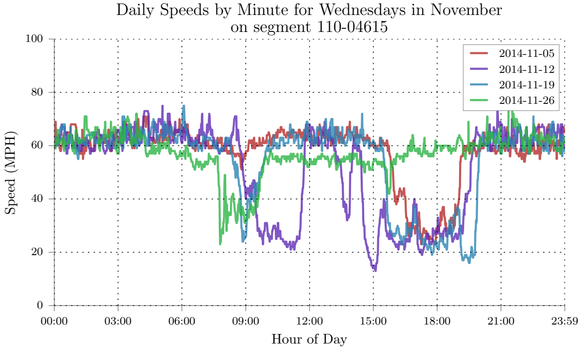
FIGURE: Location of NJ11-0009 segment in New Jersey, near Philadelphia.

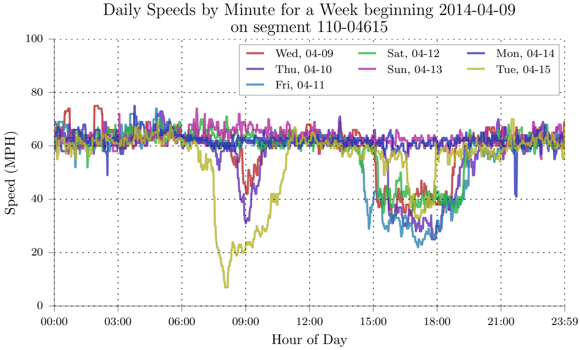


Communication from GPS (FHWA, 1998) [Ref: Kartika, C.S.D (2015)




Daily Speeds by Minute for Wednesday, 2014-05-28
Speed (MPH)
Hour of Day
100
80
60
40
20
0
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
23:59
110+04494
110-04519
110+09932
110-04618


Daily Speeds by Minute for Wednesdays in November
on segment 110-04615
Speed (MPH)
100
80
60
40
20
0
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
23:59
Hour of Day
2014-11-05
2014-11-12
2014-11-19
2014-11-26



Daily Speeds by Minute for a Week beginning 2014-04-09
on segment 110-04615
Wed, 04-09
Thu, 04-10
Fri, 04-11
Sat, 04-12
Sun, 04-13
Mon, 04-14
Tue, 04-15
Speed (MPH)
0
20
40
60
80
100
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
23:59
Hour of Day

Table 3: County-wise Number of TMC Segments

County	Number of TMC Segments
ALLEGANY	114
ANNE ARUNDEL	1,128
BALTIMORE	3,666
BALTIMORE CITY	8
BALTIMORE COUNTY	64
CALVERT	52
CAROLINE	120
CARROLL	305
CECIL	299
CHARLES	263
DORCHESTER	78
FREDERICK	617
GARRETT	86
HARFORD	491
HOWARD	634
KENT	22
MONTGOMERY	1,905
PRINCE GEORGE'S	1,694
QUEEN ANNE'S	148
SOMERSET	30
ST. MARY'S	66
TALBOT	30
WASHINGTON	261
WICOMICO	107
WORCESTER	107
Total	**12,295**

Some features of BIGDATA

- May not contain the variable(s) of interest
- Missing-data
- Errors due to measurement, classification, self selection, etc.
- Massive complex data for local area
- Computational issue

How do we correct Big Data?

Look for existing sample survey data or conduct a new survey

Some features of sample surveys

- Finite populations
- Representativeness
- Large samples for large areas, but small or no sample for small areas
- Variable(s) of interest can be included
- Chance selection: equal/epsem
- Stratification to improve precision and administrative control

Ref: Cochran (1977); Kalton (1983); Lohr (2010)

Sample Survey Data

- **Problem 1:** ACS
- **Problem 2:** June Enumerative Survey
- **Problem 3:** National Household Travel Survey (NHTS) and American Community Survey (ACS)

How do we combine Big Data with Sample Survey Data?

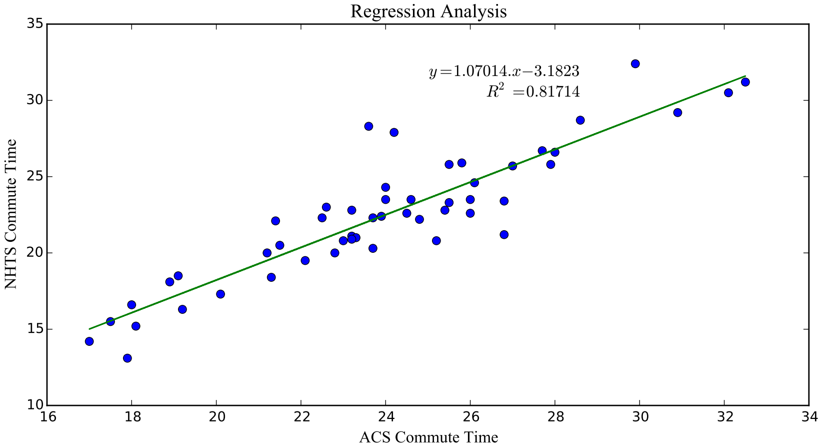
Data Fusion

- **Sample Survey Data**
  - National Household Travel Survey (NHTS)
  - American Community Survey (ACS)

- **Aggregated Administrative Data**
  - Supplemental Nutrition Assistance Program (SNAP) data (county level)
  - Internal Revenue Service Aggregate data (state level)

- **BIGDATA**
  - Vehicle Probe Project (VPP)
  - National Performance Management Research Data Set (NPMRDS)

A Proof of Data Fusion Concept



Synthetic Estimation

Introduction

- Small areas have the same characteristics as the large area (e.g., unemployment rate for a given demographic group remains the same across different states)
- implicit or fixed effects explicit modeling
- Simple and intuitive.
- Applies to any sampling design.
- Provides estimates for areas with no sample from the sample survey.

Explicit Modeling: Example 1

Ref: Hansen et al. (1953)

Estimate the median number of radio stations heard during the day for over 500 counties of the USA (small areas).

Two different survey data used:

- Mail Survey
  - large sample (1000 families/county) from an incomplete list frame
  - response rate was low (about 20%)
  - estimates x_i are biased due to non-response and incomplete coverage

Explicit Modeling: Example 1

- Personal Interview Survey: stratified multi-stage area frame
  - Nonresponse and coverage error properties were better than the mail survey
  - reliable estimates y_i for the 85 sampled counties were available, but no estimate can be produced for the remaining 415 counties
- Using (y_i, x_i) for the 85 sampled counties, the following fitted line was obtained:

$$\hat{y}_i = 0.52 + 0.74x_i$$

- Use y_i for the 85 sampled counties and $\hat{y}_i$ for the rest.

Explicit Modeling: Example 1

Table 5. Comparison of specified results from the interview survey with corresponding unadjusted figures from the mail survey

Number of stations heard	Per cent of households reporting hearing specified number of stations			
	During the day		At night	
	Mail (per cent)	Interview (per cent)	Mail (per cent)	Interview (per cent)
0	0	0	0	0
1	4	6	6	6
2	11	15	15	16
3	12	22	16	22
4	17	24	16	23
5	16	15	15	16
6	12	8	11	9
7	10	5	8	5
8	7	2	5	2
9	5	2	3	1
10 or more	6	1	5	1
Total	100	100	100	100
Median number of stations heard	4.9	3.8	4.3	3.8

Explicit Modeling: Example 1

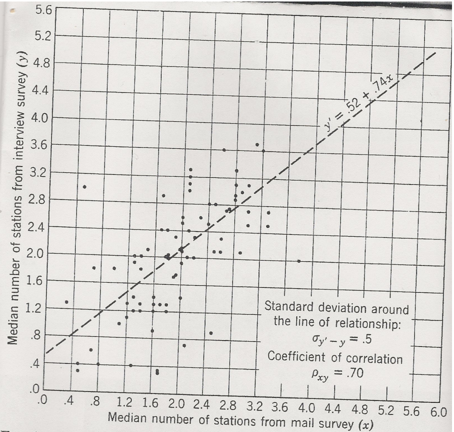


FIG. 2. Comparison of median numbers of stations heard during the day without trouble, as estimated from mail and interview surveys in selected primary sampling units.

Explicit Modeling: Example 2

Stasny et al. (1991) considered the problem of county level farm production in the state of Kansas.

- County estimates of farm production are often used in local decision making and companies selling fertilizers, pesticides, crop insurance and farm equipment.
- Non-probability sample
- y_{ij}: wheat production of the jth farm in the ith county $(i = 1, \cdots, m; j = 1, \cdots, N_i)$.
- x_{ijk}: value of kth auxiliary variable for the jth farm in the ith county $(i = 1, \cdots, m; j = 1, \cdots, N_i; k = 1, \cdots, p)$.
- Auxiliary variables chosen have known area totals $X_{ik} = \sum_{j \in U_i} x_{ijk}$ and include size of farm to reduce selection bias.

Explicit Modeling: Example 2

Estimation: Consider the following multiple linear regression model:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \epsilon_{ij} \\ &= \mathbf{x}_{ij}^T \boldsymbol{\beta} + \epsilon_{ij}, \end{aligned}$$

where $\epsilon_{ij} \overset{iid}{\sim} (0, \sigma^2)$.

- Estimate $\boldsymbol{\beta}$ by the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$.
- Obtain the fitted values

  $$\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_{ij1} + \cdots + \hat{\beta}_p x_{ijp} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}},$$

  for $(i = 1, \cdots, m; j = 1, \cdots, N_i)$.

Explicit Modeling: Example 2

Regression synthetic estimator of the ith county total $Y_i = \sum_{j \in U_i} y_{ij}$ is given by

$$\begin{aligned} \tilde{Y}_{iS} &= \sum_{j \in U_i} \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} \\ &= \sum_{j \in U_i} [\hat{\beta}_0 + \hat{\beta}_1 x_{ij1} + \cdots + \hat{\beta}_p x_{ijp}] \\ &= N_i \hat{\beta}_0 + X_{i1} \hat{\beta}_1 + \cdots + X_{ip} \hat{\beta}_p \\ &= \mathbf{X}_i^T \hat{\boldsymbol{\beta}} \end{aligned}$$

Questions:

- Do you need values of the auxiliary variables for the unobserved units of the population?
- Do their county regression-synthetic estimates add up to the state direct estimate?
- If the sample fractions $f_i = n_i/N_i$, can you propose an alternate estimator? Ref: Holt et al. (1979).

Implicit Modeling

Ref: Heuser et al. (1984)

- N_{ig}= Female population size for the gth race x age-group for the ith state. We consider the state of Pennsylvania and the data are obtained from the hospital registration system.
- $p_{.g}$= national level direct estimate of the proportion of jaundiced infants whose mother is in the gth group. The data is obtained from the 1980 National Natality Survey.

Implicit Model

Subgroup		N_{ig}	$p_{.g}$	$N_{ig}p_{.g}$
White	Under 20	16382	0.216	3539
	20-24	44100	0.214	9437
	25-29	46421	0.222	10305
	30-34	22400	0.224	5018
	35+	5896	0.244	1439
All Other	Under 20 20-24	5493 7657	0.173 0.167	950 1279
	25-29	5063	0.19	962
	30+	3387	0.266	901
		156799		33830

- A synthetic estimate of the percentage of jaundiced infants in Pennsylvania: $p_i^s = \frac{33830}{156799} * 100 = 21.6\%$.
- Estimate of total number of jaundiced infants in Pennsylvania$=N_{i.}p_i^s = 33,830$.

Other Applications

- In 1968, the National Center for Health (NCHS) used synthetic method to estimate state long term and short term disabilities from the National Health Interview (NHIS) survey data.
- US Census Bureau used synthetic method to estimate unemployment rates for counties, Gonzalez and Hoza (1978).
- Reweigting Methods: Schirm and Zaslavsky (1997)

Mean Squared Error (MSE) of Synthetic Estimators

$$
\begin{aligned}
& MSE(\hat{\bar{Y}}_{iS}) \\
= \quad & E(\hat{\bar{Y}}_{iS} - \bar{Y}_i)^2 \\
= \quad & E(\hat{\bar{Y}}_{iS} - \hat{\bar{Y}}_i + \hat{\bar{Y}}_i - \bar{Y}_i)^2 \\
= \quad & E(\hat{\bar{Y}}_{iS} - \hat{\bar{Y}}_i)^2 + E(\hat{\bar{Y}}_i - \bar{Y}_i)^2 + 2E(\hat{\bar{Y}}_{iS} - \hat{\bar{Y}}_i)(\hat{\bar{Y}}_i - \bar{Y}_i),
\end{aligned}
$$

$$
\begin{aligned}
& E(\hat{\bar{Y}}_{iS} - \hat{\bar{Y}}_i)(\hat{\bar{Y}}_i - \bar{Y}_i) \\
= \quad & E[(\hat{\bar{Y}}_{iS} - E(\hat{\bar{Y}}_{iS})) + (E(\hat{\bar{Y}}_{iS}) - \bar{Y}_i) + (\bar{Y}_i - \hat{\bar{Y}}_i)](\hat{\bar{Y}}_i - \bar{Y}_i) \\
= \quad & Cov(\hat{\bar{Y}}_{iS}, \hat{\bar{Y}}_i) + E\{[E(\hat{\bar{Y}}_{iS} - \bar{Y}_i)][\hat{\bar{Y}}_i - \bar{Y}_i]\} - Var(\hat{\bar{Y}}_i) \approx -Var(\hat{\bar{Y}}_i),
\end{aligned}
$$

since $Cov(\hat{\bar{Y}}_{iS}, \hat{\bar{Y}}_i) \approx 0$ and $E(\hat{\bar{Y}}_i) \approx \bar{Y}_i$. Therefore

$$
MSE(\hat{\bar{Y}}_{iS}) \approx E(\hat{\bar{Y}}_{iS} - \hat{\bar{Y}}_i)^2 - V(\hat{\bar{Y}}_i).
$$

Mean Squared Error (MSE) of Synthetic Estimators

I. Estimate MSE($\hat{\hat{Y}}_{iS}$)
$mse(\hat{\hat{Y}}_{iS}) = (\hat{\hat{Y}}_{iS} - \hat{\hat{Y}}_i)^2 - v(\hat{\hat{Y}}_i)$ (unstable), where $v(\hat{\hat{Y}}_i)$ is a design-unbiased estimator of $V(\hat{\hat{Y}}_i)$

II. Estimate average MSE($\hat{\hat{Y}}_{iS}$)(Gonzalez, 1973.)
$\frac{1}{m}\sum_{i=1}^{m}(\hat{\hat{Y}}_{iS} - \hat{\hat{Y}}_i)^2 - \frac{1}{m}\sum_{i=1}^{m} v(\hat{\hat{Y}}_i)$

III. Marker (1995) $MSE(\hat{\hat{Y}}_{iS}) = V(\hat{\hat{Y}}_{iS}) + \text{Bias}_i^2(\hat{\hat{Y}}_{iS})$.

$$mse_M(\hat{\hat{Y}}_{iS}) = v(\hat{\hat{Y}}_{iS}) + \frac{1}{m}\widehat{\sum_{i=1}^{m}\text{Bias}_i^2}$$

$\frac{1}{m}\widehat{\sum_{i=1}^{m}\text{Bias}_i^2} = \frac{1}{m}\sum(\hat{\hat{Y}}_{iS} - \hat{\hat{Y}}_i)^2 - \frac{1}{m}\sum v(\hat{\hat{Y}}_i) - \frac{1}{m}\sum v(\hat{\hat{Y}}_{iS})$

True Percent and Estimated RRMSE for Direct and Synthetic Estimates

Char. State	True Pct.	Dir Est	Dir Est RRMSE	Syn Est	Syn Est RRMSE
Low birth					
Penn	6.5	6.8	22	6.5	0
Tenn	8.0	8.5	23	7.2	10
Mont	5.6	9.2	71	6.3	13
PN Care					
Penn	3.9	4.3	21	4.3	10
Tenn	5.4	4.7	26	5.0	7
Mont	3.7	3.0	62	4.3	16
Apgar					
Penn	7.9	7.7	14	9.4	19
Tenn	9.6	7.3	18	9.7	1
Mont	11.6	12.9	40	9.4	19

Estimation of Sampling Variance of Direct Estimator

- Consider the SRS case. As pointed out earlier, estimation of the sampling variance of the direct estimator $V_p(\bar{y}_i)$ is challenging since this involves estimation of the finite population variance S_i^2. This is indeed another (possibly more difficult) small area estimation problem.
- The direct estimator s_i^2 of S_i^2 is unreliable due to small sample size and does not even exist when area sample size is 1.
- A synthetic variance estimator can be obtained as

$$v_S(\bar{y}_i) = (1 - f_i)\frac{s^2}{n_i},$$

  where $s^2 = (n-1)^{-1}\sum_{j\in A}(y_j - \bar{y})^2$, the pooled sample variance, and $\bar{y} = n^{-1}\sum_{j\in A} y_j$, overall sample mean.
- The variance of this synthetic variance estimator is expected to be small at the expense of increased bias.

Estimation of Sampling Variance of Direct Estimator

- We can also propose the following synthetic estimator of $V_p(\hat{\bar{Y}}_{i;R})$:

$$v_S(\hat{\bar{Y}}_{i;R}) = (1 - f_i)\frac{s_e^2}{n_i},$$

where $s_e^2 = (n-1)^{-1}\sum_{j\in A}(e_j - \bar{e})^2$, the pooled sample variance of the residuals, and $\bar{e} = n^{-1}\sum_{j\in A} e_j$, overall sample mean of the residuals.

- For a complex survey design, a possible synthetic estimator of sampling variance of the direct survey-weighted estimator is given by

$$v_S(\bar{y}_{iw}) = v_S(\bar{y}_i) \times \text{deff}_i,$$

where deff_i is an approximation of design effect. Often time design effect for a large area that covers small area is used for deff_i.

Extension of the Generalized Variance Function (GVF) Method

- Fit a model relating standard variance estimates v_i to the estimates $\bar{y}_{iw}$ and auxiliary variables x_i based on relatively larger area data. Let the fitted model be $g(\bar{y}_{iw}, x_i; \hat{\phi})$, where $\hat{\phi}$ is a vector of model parameters.
- A synthetic estimator of the sampling variance of $\bar{y}_{iw}$ is then given by

$$v_S(\bar{y}_{iw}) = g(\hat{\bar{Y}}_{i;S}, x_i; \hat{\phi}),$$

where $\hat{\bar{Y}}_{i;S}$ is a synthetic estimator of $\bar{Y}_i$.
- Fay and Herriot (1979) used:

$$g(\bar{y}_{iw}, x_i; \hat{\phi}) = \frac{9}{N_i}\bar{y}_{iw}^2,$$

where $x_i = N_i$ is population size in area i and $\hat{\phi} = 9$.

Extension of the GVF Method

Using data from relatively large areas, Liu (2009) fitted the following logistic model:

$$\text{logit}(p_{iw}) = x_i'\beta + \epsilon_i,$$

where p_{iw} is the direct survey-weighted proportion; x_i is a vector of auxiliary variables; β is the unknown vector of regression coefficients; the random errors ϵ_i are assumed to follow a distribution with zero mean and variance σ^2. Then synthetic estimate of all small area proportions are obtained as:

$$\tilde{p}_{i;S} = \frac{\exp(x_i'\hat{\beta})}{1 + \exp(x_i'\hat{\beta})},$$

where $\hat{\beta}$ is an estimator of β. The synthetic estimator of the sampling variance of p_{iw} is then obtained as:

$$v_{i;S}(p_{iw}) = \frac{\tilde{p}_{i;S}(1 - \tilde{p}_{i;S})}{n_i}\text{deff}_i,$$

where n_i is the number of respondents and deff_i is an approximation to the design effect.

Composite Estimation

Introduction

Aim: To balance the potential bias of the synthetic estimator against the instability of the design-based direct estimator.

$$\hat{Y}_{ic} = (1 - B_i)\hat{Y}_i + B_i\hat{Y}_{iS},$$

where

$$\begin{array}{rcl} \hat{Y}_i & : & \text{direct estimate for ith small-area} \\ \hat{Y}_{iS} & : & \text{synthetic estimate for ith small-area} \\ B_i & : & \text{suitably chosen weight}, 0 \leq B_i \leq 1. \end{array}$$

Sample Size Dependent (SD) estimator

$$B_i = \begin{cases} 0 & \text{if } \hat{N}_i \geq \delta N_i \\ 1 - \hat{N}_i/(\delta N_i) & \text{otherwise,} \end{cases}$$

where δ is subjectively chosen.

$\delta \in [2/3, 3/2]$ for most practical situations. $\delta = 2/3$ for Canadian LFS (Ghosh & Rao 1994, Drew, Singh and Choudhry 1982).

Remark:
Consider a SRS of size n from a population of N units and $\delta = 1$. Then, $\hat{N}_i = (N/n)n_i$, where n_i is the sample size for the ith small area.

In this case, $\hat{N}_i \geq N_i \Rightarrow n_i \geq E(n_i) = n(N_i/N)$. The method assigns the same weight no matter what variable we consider.

Optimal B_i (COM)

Minimize $MSE_p(\hat{Y}_{ic})$ w.r.t. B_i assuming

$$\text{Corr}_p(\hat{Y}_i, \hat{Y}_{iS}) \approx 0.$$

$$\begin{aligned}
& MSE_p(\hat{Y}_{ic}) \\
= \;& E_p\{(1 - B_i)\hat{Y}_i + B_i\hat{Y}_{iS} - Y_i\}^2 \\
= \;& E_p\{(1 - B_i)(\hat{Y}_i - Y_i) + B_i(\hat{Y}_{iS} - Y_i)\}^2 \\
\approx \;& (1 - B_i)^2 V_p(\hat{Y}_i) + B_i^2 MSE_p(\hat{Y}_{iS}) \\
= \;& f(B_i), (\text{ say}),
\end{aligned}$$

since

$$\begin{aligned}
& E_p(\hat{Y}_i - Y_i)(\hat{Y}_{iS} - Y_i) \\
= \;& E_p(\hat{Y}_i - Y_i)\{(\hat{Y}_{iS} - E_p\hat{Y}_{iS}) + (E_p\hat{Y}_{iS} - Y_i)\} \\
= \;& Cov_p(\hat{Y}_i, \hat{Y}_{iS}) + (E_p\hat{Y}_{iS} - Y_i)E_p(\hat{Y}_i - Y_i) \\
\approx \;& 0.
\end{aligned}$$

We used

$$E_p\hat{Y}_i \approx Y_i \text{ and } Cov_p(\hat{Y}_i, \hat{Y}_{iS}) \approx 0.$$

Thus,

$$f'(B_i) = -2(1 - B_i)V_p(\hat{Y}_i) + 2B_i MSE_p(\hat{Y}_{iS}).$$

Therefore, the approximately optimal B_i is given by

$$B_i^* = \frac{V_p(\hat{Y}_i)}{MSE_p(\hat{Y}_{iS}) + V_p(\hat{Y}_i)} = \frac{F_i}{1 + F_i},$$

where $F_i = \frac{V_p(\hat{Y}_i)}{MSE_p(\hat{Y}_{iS})}$.

The parameter B_i^* can be estimated by

$$\hat{B}_i^* = \frac{v(\hat{Y}_i)}{(\hat{Y}_{iS} - \hat{Y}_i)^2}.$$

Remarks:

- $\hat{B}_i^*$ is very unstable.
- $\hat{B}_i^*$ could be more than 1.
- There are several choices of $\hat{Y}_i$ and $\hat{Y}_{iS}$.

The Purcell-Kish Estimator

Minimize $m^{-1}\sum_{i=1}^{m} MSE_p(\hat{Y}_{ic})$ w.r.t. a common weight $B_i = B$ $(i = 1, \cdots, m)$. The approximately optimal B is given by

$$B^* = \frac{\sum_i V_p(\hat{Y}_i)}{\sum_i [MSE_p(\hat{Y}_{iS}) + V_p(\hat{Y}_i)]} = \frac{F}{1+F},$$

where $F = \frac{\sum_i V_p(\hat{Y}_i)}{\sum_i MSE_p(\hat{Y}_{iS})}$.

The Purcell-Kish estimator is given by:

$$\hat{Y}_{iPK} = (1 - \hat{B}^*)\hat{Y}_i + \hat{B}^*\hat{Y}_{iS},$$

where

$$\hat{B}^* = \frac{\sum v(\hat{Y}_i)}{\sum_i (\hat{Y}_{iS} - \hat{Y}_i)^2}.$$

A Simulation Experiment

Falorsi, P. D., Falorsi, S., Russo, A. (1994). Empirical Comparison of small area estimation methods for the Italian Labor Force Survey, *Survey Methodology*, **20**, 171-176.

- Parameter: unemployment counts for small areas
- Small areas: 14 Health Service Areas (HSA) of the Friuli Region. The small areas are unplanned areas that cut across design strata.
- Performances of direct post-stratified, sample dependent (SSD) with $\delta = 1$ and optimal composite (ϕ_i determined from the census) small-area estimators were studied by simulating sample from the 1981 Italian General Population Census.

- Samples are drawn following the LFS design (two stages with stratification of the PSUs). PSU: municipalities, SSU: HH. There were 39 PSUs and 2,290 SSUs.

- 400 sample replicates each of identical size (in terms of PSUs and of SSUs) of the LFS sample.

- $\overline{ARB} = \frac{1}{14}\sum_{i=1}^{14} |ARB_i|$, where $ARB_i = 100 \times \frac{1}{400}(\sum_{r=1}^{400} \frac{\hat{Y}_i(r) - Y_i}{Y_i})$.

- $\overline{RRMSE} = \frac{1}{14}\sum_{i=1}^{14} RRMSE_i$, where $RRMSE_i = 100 \times \frac{\sqrt{MSE_i}}{Y_i}$, and $MSE_i = \frac{1}{400}\sum_{r=1}^{400}(\hat{Y}_{i(r)} - Y_i)^2$.

Results

$\overline{\text{ARB}}$ and $\overline{\text{RRMSE}}$ for Unemployed by Estimator

Estimator	$\overline{\text{ARB}}$	$\overline{\text{RRMSE}}$
POS	1.75	42.08
SYN	8.97	23.80
COM	6.00	23.57
SD	2.39	31.08

$\overline{ARB}$

- POS presents the smallest bias.
- Bias of SYN is larger than that of the other estimators.
- Bias of COM is roughly 30% lower than that of SYN.
- The bias of POS is only slightly lower than that of SD.

$\overline{RRMSE}$

- SYN and COM have the smallest $\overline{RRMSE}$.
- POS has largest $\overline{RRMSE}$.
- $\overline{RRMSE}$ of SD is approx. 30% higher than SYN and COM.

Model-Based Methods

Components of Model-Based methods

- Identify good auxiliary information, X
  -area specific
  -element specific
  -over space and time
- Model selection & Model diagnostics
- Choice of model-based method
- Benchmarking.
- Measurement of uncertainty
- Robustness
- Evaluation studies

Area Level Models

The Fay-Herriot Model: Fay and Herriot (1979)

Let $\hat{\bar{Y}}_i$: direct survey estimate of true area mean $\bar{Y}_i$

Level 1: (Sampling Model) $\hat{\bar{Y}}_i \mid \bar{Y}_i \overset{ind}{\sim} N[\bar{Y}_i, \psi_i]$;

Level 2: (Linking Model) $\bar{Y}_i \overset{ind}{\sim} N[\mathbf{x}_i'\boldsymbol{\beta}, A]$.

- The hyper-parameters $\boldsymbol{\beta}$ and A are unknown,
- The sampling variances ψ_i are assumed to be known.
- Linear Mixed Model: $\hat{\bar{Y}}_i = \bar{Y}_i + e_i = \mathbf{x}_i'\boldsymbol{\beta} + v_i + e_i$, where $\{e_i\}$ and $\{v_i\}$ are independent with $e_i \sim N(0, \psi_i)$ and $v_i \sim N(0, A)$.

Small Area Income and Poverty Estimates (SAIPE)

- $\hat{\bar{Y}}_i$: ACS survey-weighted proportion of poor school-age children for the ith state $(i = 1, \cdots, 51)$.
- ψ_i: Fay's successive difference replication sampling variance estimate from ACS.
- Area level Covariates $(\mathbf{x}_i)$
  - The proportion of child exemptions reported by families in poverty on their tax returns.
  - The proportion of people under 65 who did not file income tax returns.
  - The proportion of people receiving food stamps.
  - Residual from a linear regression of the proportion of poor school-age children from the most recent census.

A General Area Level Model

Let $\hat{\bar{Y}}_i$: direct survey estimate of true area mean $\bar{Y}_i$

A Two-Level Model

Level 1: (Sampling Model) $\hat{\theta}_i = g(\hat{\bar{Y}}_i) \mid \theta_i = g(\bar{Y}_i) \overset{ind}{\sim} N[\theta_i, \psi_i]$;

Level 2: (Linking Model) $h(\theta_i) \overset{ind}{\sim} N[\mathbf{x}_i'\boldsymbol{\beta}, A]$.

- $g(\cdot)$ and $h(\cdot)$ are two specified functions.
- The hyper-parameters $\boldsymbol{\beta}$ and A are unknown,
- The sampling variances ψ_i are assumed to be known.

Baseball Data; Efron and Morris (1975)

- The batting average of an extremely good hitter Roberto Clemente was obtained from New York Times dated April 26, 1970 when he had already batted $n = 45$ times.

- The batting average for a player is just the proportion of hits among the number of the times he batted.

- Seventeen other major league baseball players who had also batted 45 times from the April 26 and May 2, 1970 issues of New York times were selected.

- Consider the problem of predicting the batting averages of all the 18 players for the entire 1970 season.

Let $\hat{P}_i$= batting average of player i, $i = 1, \cdots, 18$ $(= m)$. After $n = 45$ at bats,

$$n\hat{P}_i \overset{\text{ind}}{\sim} Bin(n, P_i), \text{ i=1}, \cdots, 18,$$

where

$$\begin{aligned} P_i &: \text{ true season batting average} \\ \hat{\theta}_i &= \sqrt{n}\arcsin(2\hat{P}_i - 1) \\ \theta_i &= \sqrt{n}\arcsin(2P_i - 1) \end{aligned}$$

Table: Batting Average Data

Player	$\hat{P}$	x_1	x_2	P
Clemente	0.400	0.314	8142	0.352
F.Robins	0.378	0.303	7542	0.306
Johnston	0.333	0.255	1139	0.238
Santo	0.244	0.244	1967	0.233
⋮	⋮	⋮	⋮	⋮
Petrocel	0.222	0.234	291	0.225
L.Alvara	0.267	0.118	51	0.224
Alvis	0.156	0.249	3514	0.183

The Efron-Morris Model

For $i = 1, \cdots, m$,

- Level 1: $\hat{\theta}_i | \theta_i \overset{\text{ind}}{\sim} N(\theta_i, 1)$;

- Level 2: $\theta_i | \mu, A \overset{\text{iid}}{\sim} N(\mu, A)$.

Remarks:

- Level 1 is known as the sampling distribution. We are interested in estimating the level 1 or high-dimensional parameters θ_i.

- Level 2 is known as the prior distribution of θ_i's. The level 2 parameters μ and A are often called hyperparameters. The number of hyperparameters are smaller than the number of Level 1 parameters.

Bayes and Empirical Bayes (Empirical Best Predictor)

- The posterior distribution of θ_i's:

$$\theta_i|\hat{\theta}_i; B \overset{\text{ind}}{\sim} N[(1-B)\hat{\theta}_i + B\mu, 1-B],$$

$i = 1, \cdots, m$, where $B = \frac{1}{1+A}$.

- The marginal distribution of $\hat{\theta}_i$'s: $\hat{\theta}_i \overset{iid}{\sim} N(\mu, 1+A)$.
- For the Efron-Morris model, the Bayes estimator of θ_i is given by:

$$\hat{\theta}_i^B = \hat{\theta}_i^B(\boldsymbol{\phi}) = (1-B)\hat{\theta}_i + B\mu,$$

where $\boldsymbol{\phi} = (\mu, B)$.

- An empirical Bayes estimator of θ_i is then given by

$$\hat{\theta}_i^{EB} = \hat{\theta}_i^B(\hat{\boldsymbol{\phi}}) = (1-\hat{B})\hat{\theta}_i + \hat{B}\mu,$$

where $\hat{\boldsymbol{\phi}} = (\hat{\mu}, \hat{B})$ is any reasonable estimator of $\boldsymbol{\phi}$.

- $\hat{B} = \frac{m-3}{\sum_{j=1}^{m}(\hat{\theta}_j - \bar{\hat{\theta}})^2}$ and $\bar{\hat{\theta}} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j$.
- The shrinkage factor $1-B$ is the relative contribution of the level 2 variance (or prior variance) A towards the total variance $1+A$.
- The higher the value of B the higher is our faith on the prior. Thus, B is a useful indicator of the effectiveness of the Bayesian model.
- B is generally unknown and thus one may consider an estimator $\hat{B}$ to understand the utility of the empirical Bayes estimator for a given data set.
- For some data set, $\hat{A}$ may be negative in which case it is usually truncated at 0 yielding an unreasonable estimate of $B = 1$.
- Efron and Morris (1975) suggested $B = \frac{m-3}{m}$ in case estimate of A is negative or zero. For strictly positive consistent estimator of B, see Li and Lahiri (2010).

The Carter-Rolph Model: An Extension of the Efron-Morris model

Carter and Rolph (1974, JASA)

- To estimate the probability that a box-reported alarm signals a structural fire given the alarm box location.
- The data from 1967-69 was used to develop estimates for 1970 box-reported alarms in Bronx, New York, and then the estimates were compared with the actual 1970 data.
- First, 2,500 boxes were grouped into 216 similar (in terms of alarm characteristics) neighborhoods with a number of requirements.

- n_i : the number of box-reported alarms at the ith box;
- π_i : the true probability of structural fires at the ith box.
- $\hat{\pi}_i$: sample proportion of structural fires at the ith box.
- m : the number of boxes in the neighborhood.

Then

$$n_i\hat{\pi}_i|\pi_i \overset{\text{ind}}{\sim} Bin(n_i, \pi_i), \ \text{i=1}, \cdots, m.$$

To stabilize variance, take the following transformation:

$$\hat{\theta}_i = \arcsin(\sqrt{\hat{\pi}_i})$$

Using the Taylor series approximation, we get

$$\begin{aligned} E[\hat{\theta}_i|\theta_i] - \theta_i &\approx 0, \\ V[\hat{\theta}_i|\theta_i] &\overset{def}{=} \psi_i \approx \psi_{i,approx}, \end{aligned}$$

where

$$\begin{aligned} \theta_i &= \arcsin(\sqrt{\pi}_i), \\ \psi_{i,approx} &= \frac{1}{4n_i}. \end{aligned}$$

The Carter-Rolph Model:

For $i = 1, \cdots, m$,

- Level 1 : $\hat{\theta}_i | \theta_i, \psi_i = \psi_{i,approx} \stackrel{\text{ind}}{\sim} N(\theta_i, \psi_i)$;
- Level 2 : $\theta_i | \mu, A \stackrel{\text{iid}}{\sim} N(\mu, A)$.

The Bayes estimator of θ_i is given by:

$$\hat{\theta}_i^B = (1 - B_i)\hat{\theta}_i + B_i \mu,$$

where

$$B_i = \frac{i}{A + \psi_i},$$

$i = 1, \cdots, m$. In the above $\psi_i = \psi_{i,approx}$, $i = 1, \cdots, m$.

Example: Estimation of Per-Capita Income of Small Places

See Fay and Herriot (1979, JASA)

- Estimation of 1969 per-capita income (PCI) for small places ($\approx$ 15,000 are for places with population $<$ 500 in 1970.)
- Income data was collected on the basis of about 20% sample in the 1970 census.
- $\hat{\bar{Y}}_i$= survey-weighted direct estimator
- $\hat{N}_i$=$\sum_{j \in s_i} w_i$=weighted sample count
- CV($\hat{\bar{Y}}_i$)$\approx \frac{3}{\sqrt{\hat{N}_i}}$
- CV: about 13% (population ≈ 500)
  about 30% (population ≈ 100)

Standard deviation increases in direct proportion to the expected value.
Let $\hat{\theta}_i = \ln(\hat{\bar{Y}}_i)$ and $\hat{\psi}_i = 9/\hat{N}_i$

Following supplementary information is available:

(1) PCI for the county
(2) value of housing for the place
(3) value of housing for the county
(4) IRS-adjusted gross income per exemption for the place
(5) IRS-adjusted gross income per exemption for the county

The Fay-Herriot Model:

For $i = 1, \cdots, m$,

$$\text{(i) } \hat{\theta}_i | \theta_i, \psi_i = \hat{D}_i \overset{ind}{\sim} N(\theta_i, \psi_i);$$

$$\text{(ii) } \textit{Apriori, } \theta_i \overset{ind}{\sim} N(\mathbf{x}_i^T \boldsymbol{\beta}, A).$$

Fay and Herriot assumed $\psi_i = \hat{\psi}_i,\ i = 1, \cdots, m$. Under the Bayesian Model, the Bayes estimator is given by:

$$\hat{\theta}_i^B = \hat{\theta}_i^B(\boldsymbol{\phi}) = (1 - B_i)\hat{\theta}_i + B_i \mathbf{x}_i^T \boldsymbol{\beta},$$

where $B_i = \frac{\psi_i}{\psi_i + A}$ and $\boldsymbol{\phi} = (\boldsymbol{\beta}, A)^T$.

If A is known, $\boldsymbol{\beta}$ can be estimated by

$$\begin{aligned} & \hat{\boldsymbol{\beta}}(A) \\ = \; & \left(\sum_{j=1}^{m} \frac{1}{D_j + A} \mathbf{x}_j \mathbf{x}_j^T \right)^{-1} \left(\sum_{j=1}^{m} \frac{1}{D_j + A} \mathbf{x}_j \hat{\theta}_j \right). \end{aligned}$$

Note that when A is known, $\hat{\boldsymbol{\beta}}(A)$ is the maximum likelihood (also weighted least square) estimator of $\boldsymbol{\beta}$. Replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}(A)$ we get the following empirical Bayes estimator of θ_i:

$$\tilde{\theta}_i^{EB} = \hat{\theta}_i^B(A) = (1 - B_i)\hat{\theta}_i + B_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(A).$$

Fay and Herriot (1979) obtained their estimator of A by solving

$$\sum_{j=1}^{m} \frac{[\hat{\theta}_j - \mathbf{x}_j^T \hat{\boldsymbol{\beta}}(A)]^2}{A + D_j} = m - p$$

subject to $A \geq 0$. When no positive solution exists, $\hat{A}$ is set to zero. We estimate B_i by $\hat{B}_i = \psi_i/(\psi_i + \hat{A})$.

When both $\boldsymbol{\beta}$ and A are unknown, one can get the following empirical Bayes estimator of θ_i:

$$\hat{\theta}_i^{EB} = \hat{\theta}_i^{B}(\hat{A}) = (1 - \hat{B}_i)\hat{\theta}_i + \hat{B}_i \mathbf{x}_i^T \hat{\boldsymbol{\beta}}(\hat{A}).$$

Fay and Herriot (1979) used the following steps in obtaining their final estimates of per-capita income for small places.

(a) Obtain $\hat{\theta}_i^{EB}$.

(b) Consider the following Winsorized EB:

$$\begin{aligned}
\hat{\theta}_i^{*EB} &= \hat{\theta}_i^{EB} \text{ if } \hat{\theta}_i - c_i \leq \hat{\theta}_i^{EB} \leq \hat{\theta}_i + c_i \\
&= \hat{\theta}_i - c_i \text{ if } \hat{\theta}_i^{EB} < \hat{\theta}_i - c_i \\
&= \hat{\theta}_i + c_i \text{ if } \hat{\theta}_i^{EB} > \hat{\theta}_i + c_i
\end{aligned}$$

where $c_i = \sqrt{\psi_i}$.

(c) A PCI estimator $e^{\hat{\theta}_i^{*EB}}$ is obtained using a simple back transformation.

(d) Apply a two-way iterative proportional adjustment (raking). We denote the final estimator by $\hat{\hat{Y}}_i^*$

Evaluation:

The U.S. Census Bureau conducted complete censuses of a random sample of places and townships in 1973 and collected income data for 1972 on a 100% basis.

$\#$ of places with population size $< 500 : 17$.

$\#$ of places with population size between 500 and 999: 7.

Estimates for 1972 were obtained by multiplying the estimates by updating factors f_i

Average Percent Difference

N	$\hat{\bar{Y}}_i$	$\hat{\bar{Y}}_i^*$	$\hat{\bar{Y}}_i^C$
< 500	28.6	22.0	31.6
500-999	19.1	15.6	19.3

where $\hat{\bar{Y}}_i^C =$ County estimate.

Residual Analysis: Baseball Data

- Standardized residual:

$$e_i = \frac{\hat{\theta}_i - \bar{\hat{\theta}}}{s},$$

  where $s^2 = \frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}_j - \bar{\hat{\theta}})^2$ is the usual sample variance.

- Marginally $\hat{\theta}_i \overset{iid}{\sim} N(\mu, 1 + A),\ i = 1, \cdots, m.$
  Under this marginal model,

$$\mathrm{E}(e_i) \approx 0, \quad \text{and } \mathrm{V}(e_i) \approx 1 + A, \quad \text{for large } m.$$

- If the model is reasonable, a plot of the standardized residuals versus the players is expected to fluctuate randomly around 0.
- If this does not happen, we might suspect the adequacy of the two-level model.
- However, random fluctuation of the residuals may not reveal certain systematic patterns of the data (Fig 0).

Fig 0: Residuals plot for the Efron-Morris model

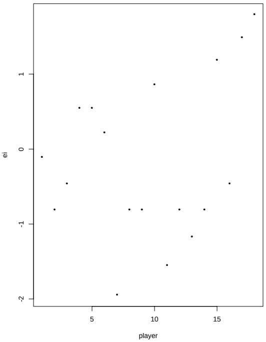
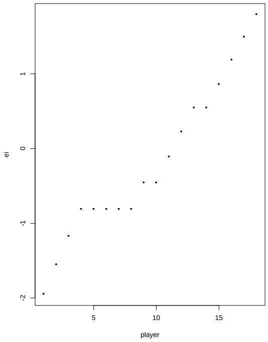


Figure 1. Residual Plot for Efron-Morris Model



- In Figure 1 we note that the residuals, when plotted against players arranged in increasing order of the previous batting averages, does reveal a linear regression pattern, a pattern not apparent when the same residuals were plotted against players arranged in an arbitrary random order. This is probably questioning the exchangeability assumption in the Efron-Morris model, a fact we knew earlier because of the intentional inclusion of a extremely good hitter.
- Let P_{i0} be the batting average of player i through the end of 1969 season. Let $x_{1i} = \sqrt{n}\,\arcsin(2P_{i0} - 1)$, $i = 1, \cdots, m$. We plot $\hat{\theta}_i$ and θ_i vs x_{1i} in Figures 2 and 3 respectively. This probably explains the systematic pattern of the residuals mentioned in the previous paragraph.
- There is striking similarity of the two graphs 2 and 3. While Roberto Clemente seems like an outlier with respect to $\hat{\theta}$, θ, or x_1, player L. Alvarado appears to be an outlier in the sense that his current batting average is much better than his previous batting average.

- Alvarado influences the regression fit quite a bit. For example, the BIC for the two-level model reduced from 55 to 44 when Alvarado was dropped from the model.
- Further investigation reveals that this player is a rookie and batted only 51 times through the end of 1969 season compared to other players in the data set, making his previous batting average information not very useful.
- The BIC for the Fay-Herriot model with and without the auxiliary data are almost the same (54.9 and 55.3 respectively), a fact not expected at the beginning of the data analysis.
- In spite of more or less similar BIC values and a presence of an outlier in the regression, Fig. 4 shows that EMReg did a good job in predicting the batting averages of Clemente and Alvarado, two different types of outliers.

Fig 2: Plot of $\hat{\theta}$ vs x_1

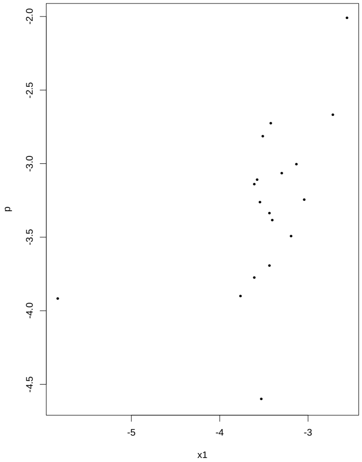


Fig 3: Plot of θ vs x_1

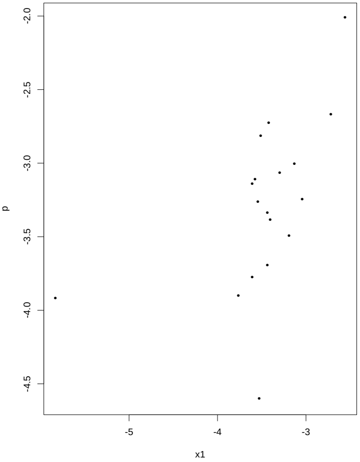
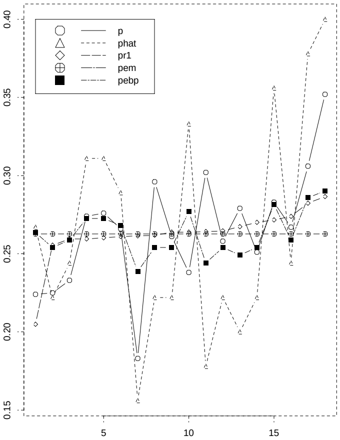


Fig 4: Plot of different estimates and true values for the baseball data



MSE Estimator: Delta Method

Ref: Lahiri and Rao (1995)

$$mse_i^T = g_{1i}(\hat{A}) + g_{2i}(\hat{A}) + 2g_{3i}(\hat{A}) - \frac{{}^{2}_{i}}{(\hat{A}+\psi_i)^2}\mathrm{bias}(\hat{A}),$$

where

$$\begin{aligned}
g_{1i}(A) &= \frac{A\psi_i}{A+\psi_i}, \\
g_{2i}(A) &= \frac{{}^{2}_{i}}{(A+\psi_i)^2} x_i'(X'\Sigma^{-1}X)^{-1}x_i, \\
g_{3i}(A) &= \frac{{}^{2}_{i}}{(A+\psi_i)^3}\frac{2}{\mathrm{tr}(\Sigma^{-2})} \\
\Sigma &= \mathrm{diag}(A+D_1, \cdots, A+D_m).
\end{aligned}$$

Other MSE Estimators

- **Jackknife Method:** Jiang, Lahiri and Wan (JLW, 2002)

$$
\begin{aligned}
mse_i^J &= g_{1i}(\hat{A}) - \frac{m-1}{m}\sum_{j=1}^{m}\{g_{1i}(\hat{A}_{(-j)}) - g_{1i}(\hat{A})\} \\
&+ \frac{m-1}{m}\sum_{j=1}^{m}\{\hat{\theta}_{i;(-j)}^{EB} - \hat{\theta}_i^{EB}\}^2
\end{aligned}
$$

where $\hat{A}_{(-j)}$ and $\hat{\theta}_{i;(-j)}^{EB}$ are obtained after deleting the jth area data

MSE Estimators

- The JLW jackknife method is quite general and applies to a general class of mixed models. Lohr and Rao (2004) discussed a area specific jackknife method to estimate the order $O(1)$ term for a specific small area model.
- **Parametric Bootstrap:** Butar (1997), Butar and Lahiri (2003), Pffermann and Glickmann (2004), Hall and Maiti (2006)
- **Computation:** SAS Proc Mixed can do a few computations.

Ref: Rao (2003) and Jiang and Lahiri (2006)

Some Comments on Modeling

- The model is simple and does not require the knowledge of detailed design Information (e.g., PSU identifiers), which may not be available in a public-use file.
- The rationale behind the transformation may rest on the Taylor series argument and may be used primarily to stabilize the variance. A direct modelling of the direct estimates is possible, but this is likely to lead to a non-linear non-normal mixed model.
- For unspecified non-normality of the sampling and random effects, one can use EBLUP [Lahiri and Rao, 1995] or linear EB [Ghosh and Lahiri, 1987] method.
- A generalized variance function (GVF) type method is generally used to estimate the sampling variances ψ_i. The method usually does not incorporate small area effects and the uncertainty in estimating the sampling variances.

Some Comments on Estimation

- In some situation, standard estimates [REML, ML, ANOVA, etc.] of the model variance A can be zero. When $\hat{A}$ is zero, EB reduces to the regression synthetic estimate. One way to avoid the problem is to use the generalized ML estimates [Morris, 1987; Li and Lahiri, 2007] or mean likelihood estimate (Bell 1999).
- A simple back transformation is often used to obtain the estimate of $\bar{Y}_i$. Good properties of the EB may be lost by such a back transformation.
- Measuring uncertainty and constructing a reliable confidence interval under the EB approach are quite challenging and the theory rests on the higher order asymptotics.
- Hierarchical Bayes implementation of the area level model provides an exact inference at the expense of specification of priors for the hyperparameters.

Interval Estimation

Definition

The $100(1-\alpha)\%$ confidence interval $\mathrm{CI}_i(\hat{\boldsymbol{\theta}})$ satisfies

$$\Pr(\theta_i \in \mathrm{CI}_i(\hat{\boldsymbol{\theta}})) = 1 - \alpha,$$

where the probability is with respect to the joint distribution of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$.

Direct Method

$$CI_i^D = [\hat{\theta}_i - z_{\alpha/2}\sqrt{\quad _i},\ \hat{\theta}_i + z_{\alpha/2}\sqrt{\psi_i}]$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percent point of $N(0,1)$.
We have $Pr(\theta_i \in CI_i^D) = 1 - \alpha$. But the interval is too wide.

Cox Empirical Bayes Confidence Interval

Ref: D.R. Cox (1975)

$$\mathrm{CI}_i^{\mathrm{Cox}} = [\hat{\theta}_i^{\mathrm{EB}} - z_{\alpha/2}\sqrt{g_{1i}(\hat{A})},\ \hat{\theta}_i^{\mathrm{EB}} + z_{\alpha/2}\sqrt{g_{1i}(\hat{A})}].$$

- $\mathrm{P}(\theta_i \in \mathrm{CI}_i^{\mathrm{Cox}}) = 1 - \alpha + O(m^{-1})$.
- The method neglects the additional errors incurred by the estimation of $\boldsymbol{\beta}$ and A.
- Note that the distribution of $\frac{\theta_i - \hat{\theta}_i^{\mathrm{EB}}}{\sqrt{g_{1i}(\hat{A})}}$ is not a standard Normal. It is not appropriate to use the Normal quantile $z_{\alpha/2}$ as the cut-off points.

Parametric Bootstrap Confidence Interval

- Use the distribution of $\frac{\theta_i^* - \hat{\theta}_i^{\text{EB}*}}{\sqrt{g_{1i}(\hat{A}^*)}}$ to approximate the distribution of $\frac{\theta_i - \hat{\theta}_i^{\text{EB}}}{\sqrt{g_{1i}(\hat{A})}}$.
- Compute $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}Y$ and $\hat{A}$, where $\hat{\boldsymbol{\Sigma}} = \text{diag}(\hat{A} + \psi_1, \cdots, \hat{A} + \psi_m)$;
- Draw bootstrap sample from the following bootstrap model:
  (i) $\hat{\theta}_i^* | \theta_i^* \overset{ind}{\sim} N(\theta_i^*, \psi_i)$
  (ii)$\theta_i^* \overset{ind}{\sim} N(\mathbf{x}_i'\hat{\boldsymbol{\beta}}, \hat{A})$
- Compute $\hat{\boldsymbol{\beta}}^*$ and $\hat{A}^*$ from $\hat{\boldsymbol{\theta}}^*$. Then we have $\hat{\theta}_i^{\text{EB}*} = (1 - \hat{B}^*)\hat{\theta}_i^* + \hat{B}^*\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*$, and $g_{1i}(\hat{A}^*) = \frac{\hat{A}^*\psi_i}{\hat{A}^* + \psi_i}$;
- Compute $(\theta_i^* - \hat{\theta}_i^{\text{EB}*})/\sqrt{g_{1i}(\hat{A}^*)}$.

The cut-off points (t_1, t_2) satisfy

$$\mathrm{P}^*[\theta_i^* < \hat{\theta}_i^{\mathrm{EB}*} + t_1\sqrt{g_{1i}(\hat{A}^*)}] = \alpha/2$$

$$\mathrm{P}^*[\theta_i^* > \hat{\theta}_i^{\mathrm{EB}*} + t_2\sqrt{g_{1i}(\hat{A}^*)}] = \alpha/2,$$

Parametric Bootstrap Confidence Interval

$$\mathrm{CI}_i^{\mathrm{PB}} = [\hat{\theta}_i^{\mathrm{EB}} + t_1\sqrt{g_{1i}(\hat{A})},\ \hat{\theta}_i^{\mathrm{EB}} + t_2\sqrt{g_{1i}(\hat{A})}].$$

Theorem

Under regularity conditions $\mathrm{Pr}(\theta_i \in \mathrm{CI}_i^{\mathrm{PB}}) = 1 - \alpha + O(p^3 m^{-3/2})$,

Ref: Chatterjee, Lahiri and Li (2008)

HB estimation

Model:
For $i = 1, \cdots, m$,

$$(i) \qquad \hat{\theta}_i | \theta_i \stackrel{\text{ind}}{\sim} N(\theta_i, \psi_i), \ \psi_i \text{ known},$$

$$(ii) \qquad \theta_i | \boldsymbol{\beta}, A \stackrel{\text{ind}}{\sim} N(\mathbf{x}_i^T \boldsymbol{\beta}, A), i = 1, \cdots, m;$$

$$(iii) \qquad \pi(\boldsymbol{\beta}, A) \propto 1.$$

HB estimation: A known

$$
\begin{aligned}
\tilde{\theta}_i^{HB}(A) &= E(\theta_i|\hat{\boldsymbol{\theta}}, A) = \tilde{\theta}_i^{EB} \\
V(\theta_i|\hat{\boldsymbol{\theta}}, A) &= g_{1i}(A) + g_{2i}(A) \\
&= \text{MSE}(\tilde{\theta}_i^{EB}) \ ,
\end{aligned}
$$

where

$$
\begin{aligned}
g_{1i}(A) &= (1 - B_i)\psi_i, \\
g_{2i}(A) &= B_i^2 V\left[\mathbf{x}_i^T \hat{\boldsymbol{\beta}}(A)\right] = B_i^2 \mathbf{x}_i^T \left(\sum_{j=1}^{m} \frac{1}{A + \psi_j} \mathbf{x}_j \mathbf{x}_j^T\right)^{-1} \mathbf{x}_i
\end{aligned}
$$

HB estimation: A unknown

$$
\begin{aligned}
\hat{\theta}_i^{HB} &= E\{E(\theta_i|\hat{\boldsymbol{\theta}}, A)|\hat{\boldsymbol{\theta}}\} \\
&= \int E(\theta_i|\hat{\boldsymbol{\theta}}, A) f(A|\hat{\boldsymbol{\theta}}) dA.
\end{aligned}
$$

The measure of uncertainty of the HB estimator $\hat{\theta}_i^{HB}$ is given by

$$
\begin{aligned}
& V(\theta_i|\hat{\boldsymbol{\theta}}) \\
= \; & E\{V(\theta_i|\hat{\boldsymbol{\theta}}, A)|\hat{\boldsymbol{\theta}}\} + V\{E(\theta_i|\hat{\boldsymbol{\theta}}, A)|\hat{\boldsymbol{\theta}}\} \\
= \; & E\{(g_{1i}(A) + g_{2i}(A)|\hat{\boldsymbol{\theta}}\} + V\{\tilde{\boldsymbol{\theta}}heta_i^{HB}(A)|\hat{\boldsymbol{\theta}}\}.
\end{aligned}
$$

Note that unlike the EB,

$$
\hat{\theta}_i^{HB} \neq \tilde{\theta}_i^{HB}(\hat{A}),
$$

for an arbitrary estimator of A.

Implementation: Gibbs Sampling (MCMC)

Need the following full conditionals:

(a) $\theta_i|\boldsymbol{\beta}, A, \hat{\boldsymbol{\theta}} \overset{\text{ind}}{\sim} N\left[\hat{\theta}_i^B,\ \psi_i(1-B_i)\right],\ i = 1, \cdots, m$

(b) $\boldsymbol{\beta}|\boldsymbol{\theta}, A, \hat{\boldsymbol{\theta}} \sim N\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\theta},\ A(\mathbf{X}^T\mathbf{X})^{-1}\right]$

(c) $A|\boldsymbol{\beta}, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}} \sim IG\left[\frac{m-2}{2};\ \frac{1}{2}\sum(\theta_i - \mathbf{x}_i^T\boldsymbol{\beta})^2\right]$

Gibbs Sampling Algorithm

(i) Draw $\theta_i^{(1)},\ i = 1, \cdots, m$, from (a), using $\boldsymbol{\beta}^{(0)}$ & $A^{(0)}$ as starting values.

(ii) Draw $\boldsymbol{\beta}^{(1)}$ from (b) using $\boldsymbol{\theta}^{(1)}$ & $A^{(0)}$.

(iii) Draw $A^{(1)}$ from (c), using $\boldsymbol{\theta}^{(1)}$ & $\boldsymbol{\beta}^{(1)}$.

The steps (i)-(iii) complete one cycle. Perform a large number of cycles. The simulated samples after deleting the first t "burn-in" samples, i.e.

$$\left\{\boldsymbol{\beta}^{(t+r)}, A^{(t+r)}, \boldsymbol{\theta}^{(t+r)},\ r = 1, \cdots, R\right\}$$

are considered as R simulated samples from $[\boldsymbol{\beta}, A, \boldsymbol{\theta}|\hat{\boldsymbol{\theta}}]$.

Gibbs Sampling Algorithm

In small area estimation, our main interest is in $\boldsymbol{\theta}$. The posterior density of $\boldsymbol{\theta}$, i.e. $[\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}]$, is approximated using

$$\left\{\boldsymbol{\theta}^{(t+r)},\ r = 1, \cdots, R\right\}.$$

In particular, we can approximate the posterior means and variances as follows:

$$\begin{aligned}
\hat{\theta}_i^{HB} &\approx \frac{1}{R}\sum_{r=1}^{R}\boldsymbol{\theta}_i^{(t+r)} = \hat{E}(\theta_i|\hat{\boldsymbol{\theta}}),\ \text{say} \\
V(\theta_i|\hat{\boldsymbol{\theta}}] &\approx \frac{1}{R-1}\sum_{r=1}^{R}\left[\boldsymbol{\theta}_i^{(t+r)} - \hat{E}(\theta_i|\hat{\boldsymbol{\theta}})\right]^2 \\
&= \hat{V}(\theta_i|\hat{\boldsymbol{\theta}}),\ \text{say}
\end{aligned}$$

for $i = 1, \cdots, m$.
By the ergodic theorem for Markov chains, $\hat{E}(\theta_i|\hat{\boldsymbol{\theta}})$ converges to $E(\theta_i|\hat{\boldsymbol{\theta}}) = \hat{\theta}_i^{HB}$ and $\hat{V}(\theta_i|\hat{\boldsymbol{\theta}})$ to $V(\theta_i|\hat{\boldsymbol{\theta}})$ as $R \to \infty$.

Unit Level Models

Introduction

- y_{ij}: value of the study variable for the jth unit of the i small area population $(i = 1, \cdots, m; \;\; j = 1, \cdots, N_i)$.
- $g(y_{ij})$ is a known function of y_{ij}
- To estimate: $\theta_i = N_i^{-1} \sum_{j=1}^{N_i} g(y_{ij})$
- Ex: For the choice $g(y_{ij}) = y_{ij}$, θ_i is the finite population mean for area i.

Nested Error Regression Model

Ref: Battese, Harter and Fuller (JASA 1988) For $i = 1, \cdots, m;\ \ j = 1, \cdots, N_i$,

$$y_{ij} = x'_{ij}\beta + v_i + e_{ij},$$

where x_{ij} is a $p \times 1$ column vector of known auxiliary variables; $\{v_i\}$ and $\{e_{ij}\}$ are all independent with $v_i \overset{iid}{\sim} N(0, \sigma_v^2)$ and $e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2)$.

We can also write the model as a two-level model:

Level 1: $y_{ij}|v_i \overset{ind}{\sim} N(x'_{ij}\beta + v_i,\ \sigma_e^2);$

Level 2: $v_i \overset{iid}{\sim} N(0,\ \sigma_v^2).$

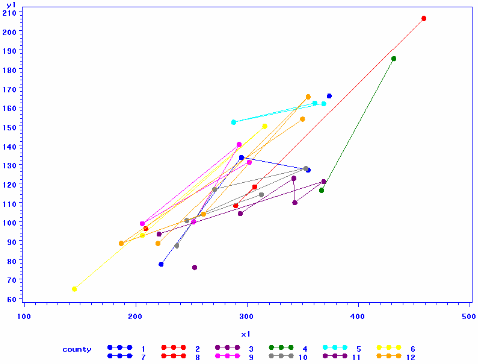
An Example

Estimation of the number of hectares of corn for 12 Iowa counties based on the 1978 June Enumerative Survey and satellite data.

Notations:

- y_{ij}: the number of hectares of corn in the jth segment of the ith county as reported in the June Enumerative Survey. Segments are about 250 hectares.
- $x'_{ij} = (1, x_{1ij}, x_{2ij})$, where x_{1ij} (x_{2ij}) is the number of *pixels* classified as corn (soybean) in the jth segment of the ith county. A *pixel* (a term for *picture elements*) is the unit for which satellite information is recorded. A pixel is about .45 hectares
- $\bar{X}' = (1, \bar{X}_{1i}, \bar{X}_{2i})$, where $\bar{X}_{1i}$ $(\bar{X}_{2i})$ is the mean number of pixels per segment classified as corn (soybean) for county i. This is the total number of pixels classified as corn divided by the number of pixels in that county.

Fig 2: Plot of Corn Hectares versus Corn Pixels by County



This plot also reflects the strong relationship between the reported hectares of corn and the number of pixels of corn for counties separately. But the slopes and/or intercepts seem differ by county.

BP/Bayes, EB and HB

Let $y_i = (y_{i,s},\ y_{i,ns})$ with $y_{i,s}$ and $y_{i,ns}$ denote the sampled and non-sampled parts, respectively. We assume a hierarchical model for $y_i,\ i = 1, \cdots, m$ (e.g., nested error model on y_{ij} or in a logarithmic scale). Then the Bayes/BP of θ_i for the general case can be approximated as follows:

- *Step 1:* Obtain L "census" files as $y^*_{i;l} = (y_{i,s}, y^*_{i,ns}),\ (l = 1, \ldots, L)$, where $y^*_{i,ns}$ is generated from the conditional distribution of $y_{i,ns}$ given $y_{i,s}$ with known hyperparameters.
- *Step 2:* Bayes/BP of θ_i is approximated by $L^{-1} \sum_{l=1}^{L} g(y^*_{i;l})$.

To obtain, EB or HB change step 1. For EB, $y^*_{i,ns}$ is generated from the conditional distribution of $y_{i,ns}$ given $y_{i,s}$ with estimated hyperparameters. For HB. $y^*_{i,ns}$ is generated from the conditional distribution of $y_{i,ns}$ given $y_{i,s}$ under some prior assumptions on the hyperparameters.

An Example: Nested Error Regression Model

Estimate $\bar{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij}$ or, equivalently, $\sum_{j=1}^{N_i} y_{ij}$ when N_i is known.
The Bayes estimator/BP of $\bar{Y}_i$:

$$
\begin{aligned}
\bar{Y}_i^B &\equiv \bar{Y}_i^B(\beta, \sigma_e^2, \lambda) \\
&= E(\bar{Y}_i | y_s; \beta, \sigma_e^2, \lambda) \\
&= f_i \hat{\bar{Y}}_i^{Reg}(\beta) + (1 - f_i) \left\{ [1 - B_i(\lambda)] \hat{\bar{Y}}_i^{Reg}(\beta) + B_i(\lambda) \hat{\bar{Y}}_i^{Syn}(\beta) \right\},
\end{aligned}
$$

where

$$
\begin{aligned}
\lambda &= \frac{\sigma_v^2}{\sigma_e^2} \\
B_i(\lambda) &= \frac{1}{1 + n_i \lambda} \\
\hat{\bar{Y}}_i^{Reg}(\beta) &= \bar{y}_i + (\bar{X}_i - \bar{x}_i)' \beta \\
\hat{\bar{Y}}_i^{Syn}(\beta) &= \bar{X}_i' \beta
\end{aligned}
$$

An Example: Nested Error Regression Model

- In an EB setting, one would estimate the hyperparameters using any classical method. For example, one can estimate β by the weighted least squares estimator with estimated variance components σ_e^2 and REML to estimate the variance components. One can then use a resampling method or Taylor series method to estimate the MSE. Confidence interval be obtained using the parametric bootstrap method of Chatterjee, Lahiri and Li (2008 AS).
- In a HB setting, one would put a prior on the hyperparameters. Typically, enough data will be available to estimate β and σ_e that one can use any reasonable noninformative prior distribution. For example, one can assume that apriori β and σ_e are independent and β and σ_e have improper uniform priors in the p-dimensional Euclidean space and positive pat of the real line, respectively. The prior on σ_v is less clear cut. See Gelman (2006). One suggestion is to put an improper uniform on σ_e. Apply MCMC.

FGT poverty measures

Ref: Foster, Greer and Thornbecke, 1984

- y: a welfare variable (income, expenditure, etc.) of interest.
- z threshold under(s) which a unit is under poverty
- For SGT poverty measure $g(y_{ij}) = \left(\frac{z-y_{ij}}{z}\right)^{\alpha} I(y_{ij} < z)$
- FGT poverty measure:

$$F_{\alpha i}(y_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{z - y_{ij}}{z}\right)^{\alpha} I(y_{ij} < z),$$

where

$$I(y_{ij} < z) = \begin{cases} 1 & \text{if } y_{ij} < z \ , \\ 0 & \text{otherwise,} \end{cases}$$

where α is a measure of the sensitivity of the index to poverty.

Examples

Examples of welfare variable

- Brazil: per-capita household expenditure.
- U.S. Small Area Income and Poverty Estimates (SAIPE) program: household income

Examples of threshold

- Brazil: IBGE used 20 different thresholds, varying by geographic region and rural/urban areas.
- U.S. SAIPE program: different thresholds are used depending on the household composition.

Poverty Incidence

$$F_{\alpha i}(y_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} I(y_{ij} < z)$$

Remarks:

- $\alpha = 0$
- proportion of units in that area living below the poverty line
- The headcount ratio merely measures the incidence of poverty, but not its intensity, i.e. measures how many poor individuals there are and not how poor they are.

Poverty Gap

$$F_{\alpha i}(y_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{z - y_{ij}}{z} \right) I(y_{ij} < z)$$

- $\alpha = 1$
- When the parameter is 1, the measure is the relative poverty gap, an index measuring poverty intensity;
- It can be interpreted as the cost of eliminating poverty (relative to the poverty line), because it shows how much would have to be transferred to the poor to bring their incomes up to the poverty line.

Poverty Severity

$$F_{\alpha i}(y_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{z - y_{ij}}{z} \right)^2 I(y_{ij} < z)$$

- $\alpha = 2$
- gives more emphasis to the very poor.

Design-Based Direct Estimation

Note that

$$F_{\alpha i}(y_i) = N_i^{-1} \sum_{j=1}^{N_i} u_{ij},$$

where

$$u_{ij} = \left(\frac{z - y_{ij}}{z} \right)^{\alpha} I(y_{ij} < z).$$

Let s_i be the set of units in the sample that belong to area i (size n_i) and w_{ij} be the survey weight associated with responding unit (ij). Then the survey-weighted direct estimator is given by

$$\hat{F}_{\alpha i}^{Dir} = \frac{\sum_{j \in s_i} w_{ij} u_{ij}}{\sum_{j \in s_i} w_{ij}}$$

Note: The direct estimators are highly unreliable due to small sample sizes in the areas.

The ELL Method (Elbers, Lanjouw and Lanjouw, 2003)

- Assume a linear mixed model on the log-transformed welfare variable of interest.
- Obtain L synthetic *census* files $\tilde{y}^*_{i;l}, \ (l = 1, \ldots, L)$.
- The ELL estimate of $F^*_{\alpha i}(y_i)$ is then obtained as $\bar{F}^*_{\alpha i} = L^{-1} \sum_{l=1}^{L} F_{\alpha i}(\tilde{y}^*_{i;l})$.
- The measure of uncertainty of the ELL estimate is given by

$$\frac{1}{L-1} \sum_{l=1}^{L} \left(F_{\alpha i}(\tilde{y}^*_{i;l}) - \bar{F}^*_{\alpha i} \right)^2 .$$

  A correction $1 + 1/L$ is often applied to capture variation due to imputation.

Remarks

- In the ELL model, area specific auxiliary variables from different administrative records can be incorporated.
- The ELL mixed model attempts to capture different features of the survey design, but not any small area specific effect.
- Just like any other synthetic small area methods, the ELL method is capable of producing poverty estimates even when there is no survey data from the area.
- In some public policymaking, unlike the EB/HB, the ELL method may be considered to be *fair* to all areas irrespective of the variation of the sample sizes across area.
- Basic data requirements: (i)Micro level census data, (ii)Micro level survey data containing the welfare variable of interest, (iii) Common auxiliary variables between the survey and the census
- Time gap between the census and the survey
- Incomparability of the auxiliary variables between the survey and the census

Time Series
Cross-Sectional Models

The Rao-Yu Model

Ref: Rao, J.N.K. and Yu, Y (1994)

For $i = 1, \cdots, m; t = 1, \cdots, T$,

$$\textit{Level 1:} : y_{it} = \theta_{it} + e_{it};$$
$$\textit{Level 2:} : \theta_{it} = x'_{it}\beta + v_i + u_{it}$$
$$\textit{Level 3:} : u_{it} = \rho u_{it-1} + \epsilon_{it} \; (|\rho| < 1)$$

where

- $e_i = (e_{i1}, \cdots, e_{iT})'$'s are independent multivariate normal with mean vector 0 and covariance matrix Ψ_i.
- An important extension of the Anderson-Hsiao model that incorporates sampling errors.
- Stationary model on the time Component

Datta-Lahiri-Maiti Model

Ref: Datta, Lahiri, Maiti (2002)

For $i = 1, \cdots, m; t = 1, \cdots, T$,

$$\textit{Level 1:} : y_{it} = \theta_{it} + e_{it};$$

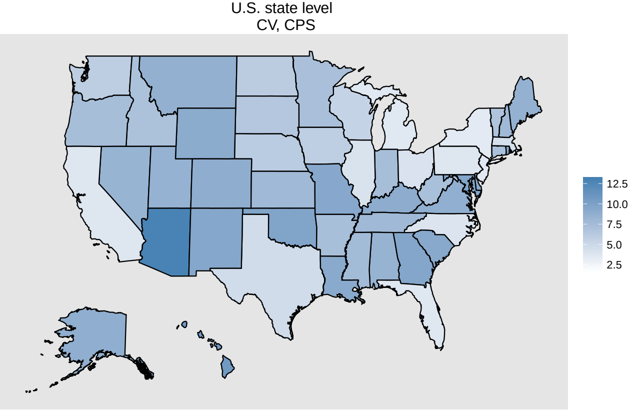
$$\textit{Level 2:} : \theta_{it} = x'_{it}\beta + v_i + u_{it}$$

$$\textit{Level 3:} : u_{it} = u_{it-1} + \epsilon_{it}$$

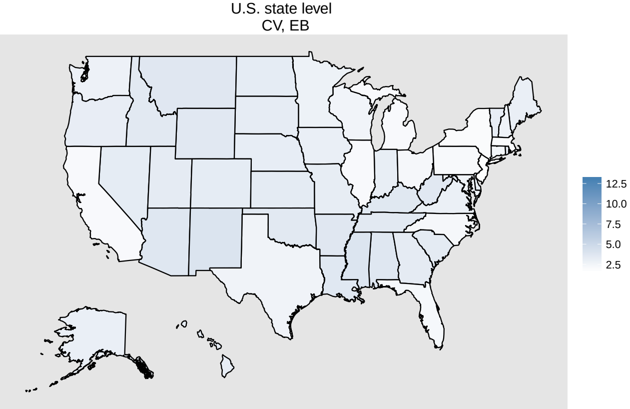
where

- This is a special case of linear mixed model.
- This model is not a special case of the Rao-Yu model
- No new theory needed. Just apply well-known results in linear mixed model.
- Ghosh and Nangia (1993) and Ghosh, Nangia and Kim (1996) also used random walk model for the time component, but their model does not include area specific random effects.

Estimates of Coefficient of Variations of CPS Direct estimates of Median Income of 4-person Families in the US States

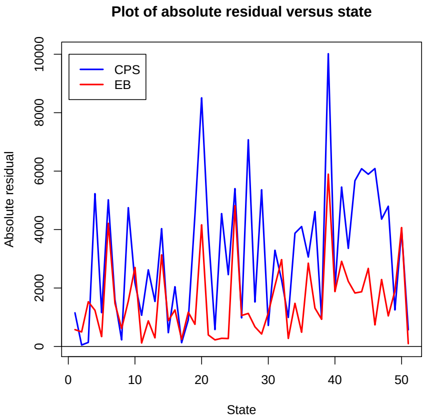


Estimates of Coefficient of Variations of EB estimates of Median Income of 4-person Families in the US States: Year 1989



A Plot of Absolute Residuals From a Simple Linear Regression

Dep Variable: 1989 Median Income Estimates from 1990 Census
Indep. Variable: CPS or EB Estimates for 1989



Other Time Series Cross-Sectional Models

For $i = 1, \cdots, m; t = 1, \cdots, T_i$,

$$\begin{aligned} \textit{Level 1:} &: y_{it} = \theta_{it} + e_{it}; \\ \textit{Level 2:} &: \theta_{it} = x'_{it}\beta_{A(it)} + v_i + u_{it} \\ \textit{Level 3:} &: u_{it} = u_{it-1} + \epsilon_{it} \text{ or a stationary model} \end{aligned}$$

- $\vec{\beta}_{A(it)}$ is a fixed effect of a bigger area at time t that covers small area i.
- Pramanik et al. (2014) considered a particular case of the model to estimate immunization rates for districts (small areas) in India.
- Spatio-temporal models can be tried (e.g., Singh et al. 2005; Pereira and Coelho 2012; Marhuenda, Molina and Morales 2013). The spatial component of the model may not be very effective in presence of reasonably good area specific auxiliary information (Vogt 2011 and work of Wayne Fuller back in the 80's)

Case Studies

Measuring Quality of Small Area Estimators in the U.S. Current Employment Statistics Survey

Partha Lahiri,
JPSM, University of Maryland, College Park

joint work with
Julie Gershunskaya, U.S. Bureau of Labor Statistics

Outline

- Overview of CES survey
- Properties of the variance estimator
- Proposed approach
- Empirical results
- Further work

Longitudinal Data Base (LDB)

- based on Quarterly Census of Employment and Wages (QCEW, formerly known as ES-202) program

- contains **monthly employment** data

  for every U.S. business establishment covered by Unemployment Insurance (UI) tax laws --

  virtually a **census**

- updated quarterly, on a **lagged** basis, approximately 6 to 9 months after the reference period

- provides a **sampling frame** and the **benchmark** data for the CES survey

CES Survey Overview

- Stratified simple random sample of Unemployment Insurance (UI) accounts:
  - State | NAICS Supersector | Size Class

  UI is cluster of establishments

- Optimal allocation minimizes variances of state-level monthly employment change

- Ests (National and for States and Areas) produced at various levels of industrial and geographical detail

Estimator for Employment Level

Weighted Link Relative (WLR) estimator:

$$\hat{Y}_t = Y_0 \hat{R}_1 \ldots \hat{R}_t ,$$

where

$$\hat{R}_t = \frac{\sum_{s_t} w_j y_{jt}}{\sum_{s_t} w_j y_{jt-1}}$$

Variance Estimation

- BHS for National Level Estimates
- RGBHS for States and Areas
- Taylor series method

Monte-Carlo Study

- 10,000 samples from Alabama
- 13 industries
- estimation for a fixed month t

Monte-Carlo Study (cont.)

$$E_d\left[\hat{X}\right] = \frac{1}{10{,}000}\sum_{s=1}^{10{,}000}\hat{X}_s$$

$$V_d\left[\hat{X}\right] = \frac{1}{9{,}999}\sum_{s=1}^{10{,}000}\left(\hat{X}_s - E_d\left[\hat{X}\right]\right)^2$$

Relative Bias: $RB\left[\hat{V}\right] = 100\%\,\dfrac{E_d\left[\hat{V} - V\right]}{V}$

Relative Variance: $CV\left[\hat{V}\right] = 100\%\,\dfrac{\sqrt{V_d\left[\hat{V}\right]}}{V}$

Relative Root MSE: $RRMSE\left[\hat{V}\right] = \sqrt{CV^2\left[\hat{V}\right] + RB^2\left[\hat{V}\right]}$

CV of the Point Estimator and CV of the Direct Variance Estimator, %

Industry	*Point Estimator*	*Variance Estimator*
1	2.5	152.6
2	1.5	73.3
3	0.5	47.2
4	0.5	48.0
5	1.0	91.3
6	0.6	29.3
7	1.2	195.6
8	1.4	94.4
9	0.7	59.5
10	1.0	193.1
11	0.6	38.6
12	0.8	34.2
13	1.7	46.0

Relative biases of the direct estimators, %

Industry	*Taylor Series*	*BHS*	*RGBHS*
1	**-2.2**	**1**	**1.3**
2	**-3**	**-2.3**	**-2.3**
3	**-0.6**	**0**	**0.2**
4	**-1.6**	**0.7**	**1.1**
5	**-1.6**	**0.4**	**-0.2**
6	**-4**	**-4.4**	**-3.9**
7	**-10.2**	**-7.1**	**-8.4**
8	**4.5**	**7**	**7.3**
9	**1.9**	**3.5**	**2.8**
10	**-18.5**	**-18.1**	**-17.2**
11	**-3.2**	**-3.2**	**-2.7**
12	**1.2**	**1.2**	**1.9**
13	**0.5**	**3.2**	**1.9**

CV of the direct variance estimators, %

Industry	*Taylor Series*	*BHS*	*RGBHS*
1	**152.6**	**173.3**	**158.3**
2	**73.3**	**100.0**	**75.9**
3	**47.2**	**73.4**	**49.7**
4	**48.0**	**73.9**	**51.3**
5	**91.2**	**125.8**	**93.9**
6	**29.0**	**64.7**	**32.5**
7	**195.4**	**221.7**	**198.7**
8	**94.3**	**104.1**	**97.3**
9	**59.4**	**92.6**	**62.7**
10	**192.2**	**202.1**	**198.7**
11	**38.5**	**75.4**	**42.5**
12	**34.2**	**69.0**	**37.6**
13	**46.0**	**89.7**	**49.6**

Synthetic Estimator of Variance

Model 1:

$$E_m[y_{j,t} \mid y_{j,t-1}] = R_{it} y_{j,t-1},$$

$$V_m[y_{j,t} \mid y_{j,t-1}] = \sigma_t^2 y_{j,t-1}.$$

$$\hat{V}_{it}^{S} = \hat{\sigma}_t^2 \frac{\sum_{s_{it}} w_j^2 y_{j,t-1}}{\left(\sum_{s_{it}} w_j y_{j,t-1} \right)^2},$$

where $\hat{\sigma}_t^2 = \dfrac{1}{\sum_{s_t} w_j} \sum_{s_t} w_j \dfrac{(y_{j,t} - \hat{R}_{it} y_{j,t-1})^2}{y_{j,t-1}}$

- Synthetic estimator reduces variance, introduces bias
- Looking for compromise

Composite Method

Take log transformation: $u_i = \ln(\hat{V}_i)$

Model 2:

$$\textbf{Level 1:} \quad E_d[u_i] = \theta_i \ ; \quad V_d[u_i] = \gamma_i^2$$

$$\textbf{Level 2:} \quad E_m[\theta_i] = \mu_i \ ; \quad V_m[\theta_i] = \tau^2$$

Composite Method (cont.)

Model 2 $\Rightarrow$ $$\hat{\theta}_i^{BLUP} = B_i u_i + (1 - B_i)\mu_i,$$

where $$B_i = \frac{\tau^2}{\tau^2 + \psi_i},$$

$$\psi_i = E_m\left[\gamma_i^2\right]$$

Estimation of parameters $\boldsymbol{\varphi} = (\psi_i, \mu_i, \tau^2)'$

- Assume: $\mu_i = \beta x_i$, where $x_i = \ln[\hat{V}_i^S]$

- Estimate $\hat{\psi}_i = \dfrac{\sqrt{\hat{V}_d[\hat{V}_i]}}{\hat{V}_i}$, where $\hat{V}_d[\hat{V}_i]$ est of $V_d[\hat{V}_i]$

- Estimation of τ^2 and β:

$$\sum_i \frac{1}{\tau^2 + \hat{\psi}_i}[u_i - \beta(\tau^2)x_i]^2 = I - 1$$

where I is the number of industries

Finally,

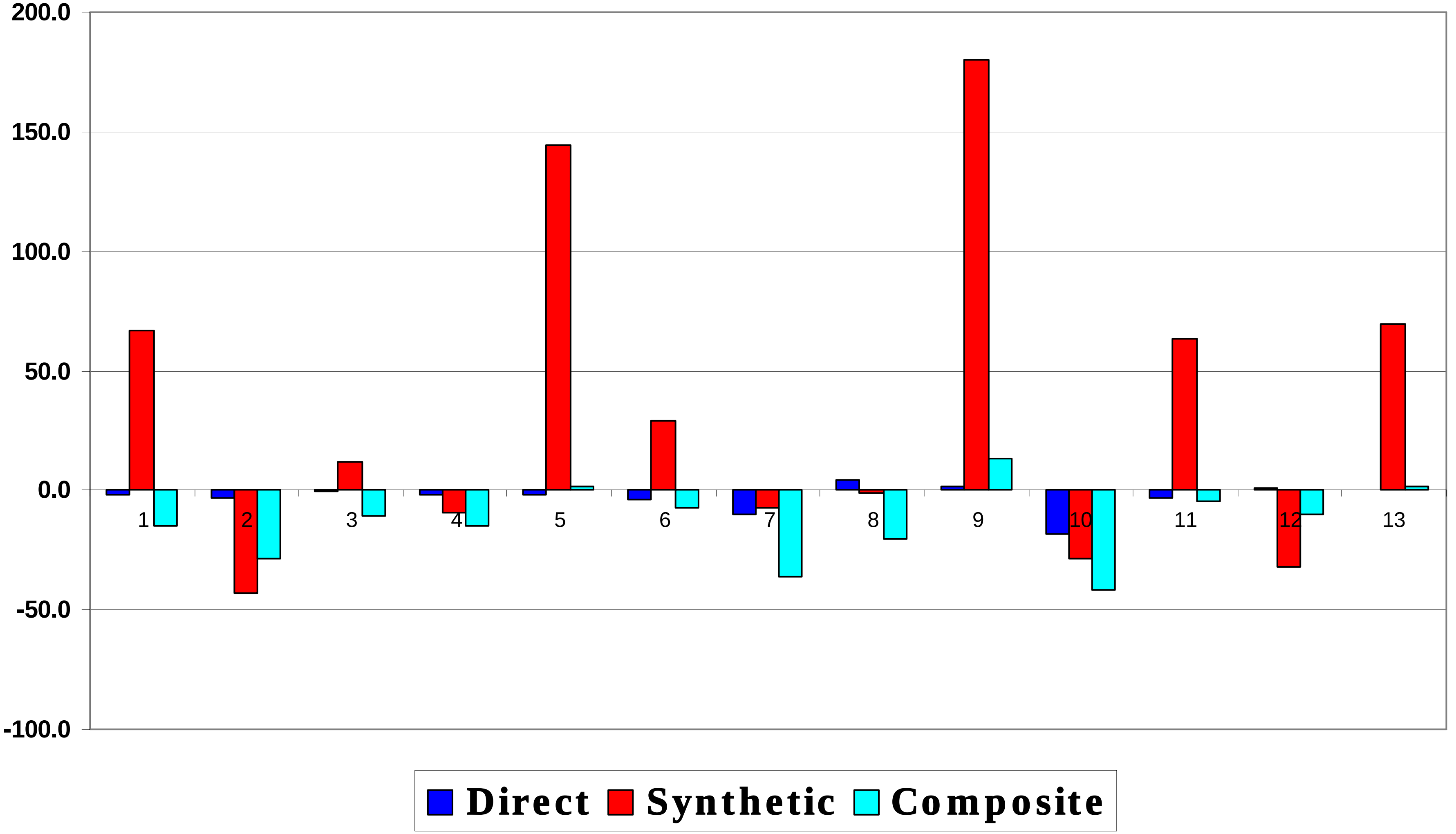
EBLUP of θ_i :

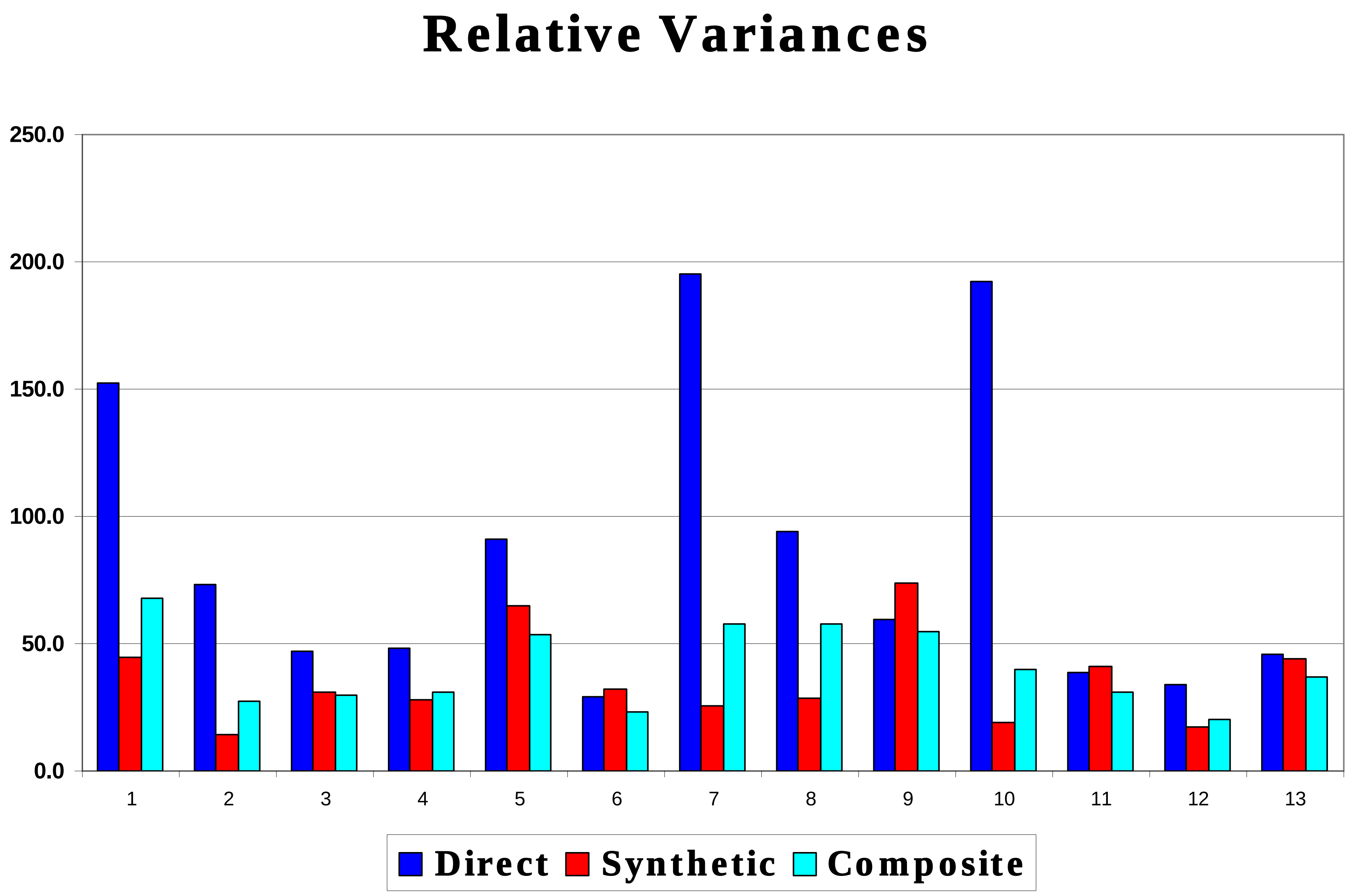
$$\hat{\theta}_i = \theta_i^{BLUP}(u_i; \hat{\boldsymbol{\varphi}}).$$

Take reverse transformation:

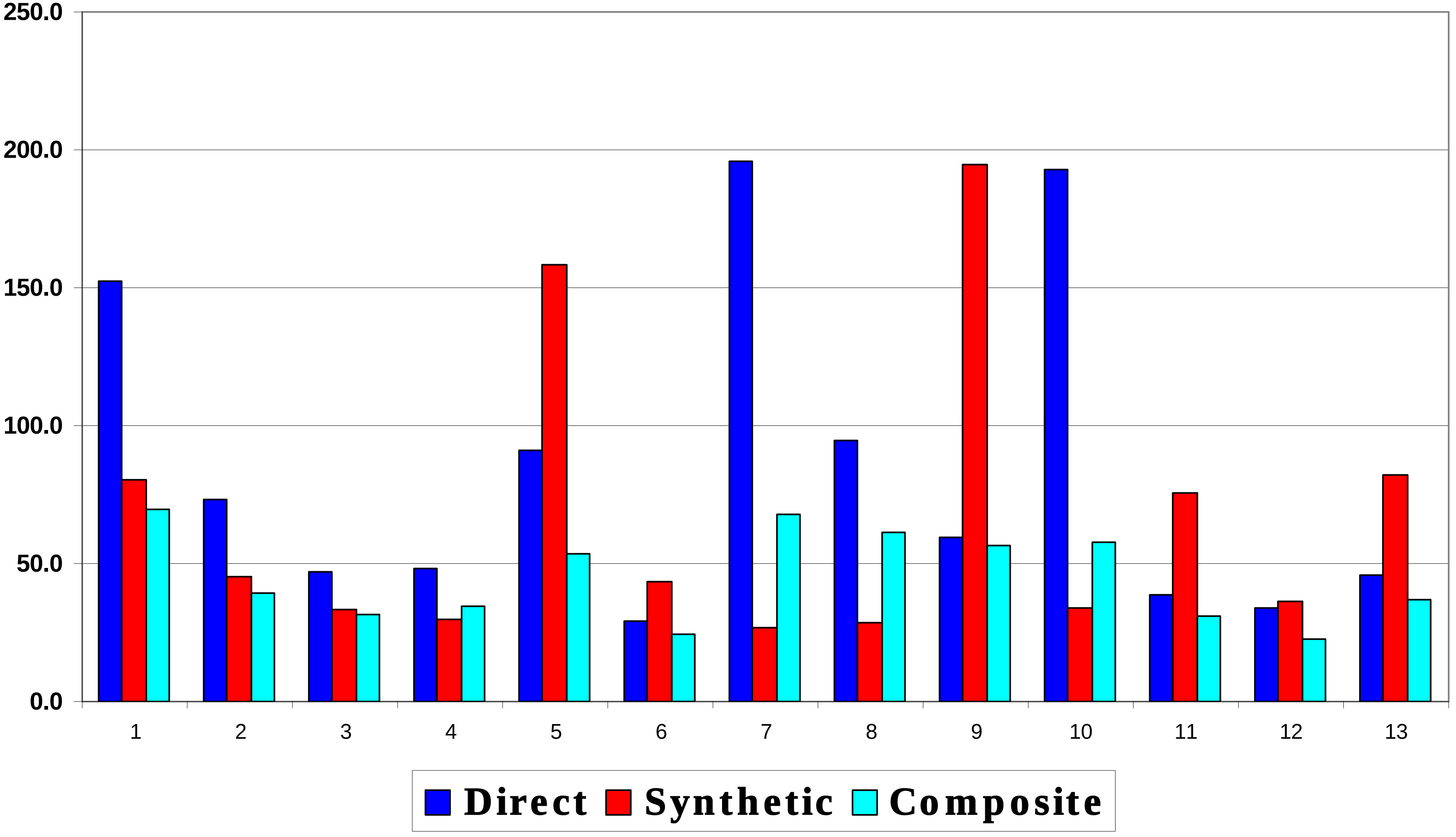
$$\hat{V}_i^C = \exp(\hat{\theta}_i)$$

Relative Bias




Relative Variances
250.0
200.0
150.0
100.0
50.0
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
Direct
Synthetic
Composite

Relative Root MSE



Relative Root MSE

Industry	*Direct*	*Synthetic*	*Composite*
1	152.6	80.4	69.6
2	73.3	45.3	39.5
3	47.2	33.4	31.4
4	48.0	29.8	34.7
5	91.3	158.6	53.3
6	29.3	43.5	24.5
7	195.6	26.7	67.9
8	94.4	28.8	61.1
9	59.5	194.7	56.3
10	193.1	34.2	57.5
11	38.6	75.5	31.1
12	34.2	36.4	22.9
13	46.0	82.2	36.8

Coverage & Length properties

Industry	*Direct*	*Synthetic*	*Composite*
1	88.4(.073,54.7)	95.4(.108, 12.8)	89.7(0.074, 33.5)
2	89.0(.046,31.2)	79.0(0.036, 12.1)	84.1(0.041, 17.6)
3	89.8(.016,21.8)	91.1(0.017, 13.2)	88.1(0.015, 16.0)
4	88.5 (.015,23.3)	87.6 (0.015, 14.4)	86.4 (0.014, 18.0)
5	89.5 (.031,34.8)	98.3 (0.051, 12.8)	91.2 (0.032, 24.7)
6	88.9 (.019,14.8)	93.1 (0.022, 12.0)	88.5 (0.019, 12.6)
7	88.7 (.032,59.0)	91.3 (0.037, 13.3)	86.3 (0.030, 31.9)
8	84.9 (.041,50.8)	88.6 (0.045, 13.9)	84.2 (0.038, 38.5)
9	89.3 (.024,27.5)	99.2 (0.040, 12.5)	91.5 (0.025, 24.0)
10	89.3 (0.025,53.1)	90.7 (0.027, 12.6)	86.9 (0.024, 23.2)
11	89.7 (0.018,18.7)	95.7 (0.023, 12.1)	89.5 (0.018, 15.7)
12	89.9 (.025, 15.5)	82.0 (0.021, 12.3)	87.9 (0.024, 11.1)
13	89.2 (.055, 22.3)	96.2 (0.072, 12.4)	89.7 (0.055, 18.4)

Summary

- Direct variance estimators may be very unstable even in domains where point ests are good
- SAE approach improves efficiency of ests of variances, with comparable coverage properties

Further research

- Alternative synthetic estimators (e.g., using historical data)
- Alternative estimator of Level 1 variance
- Extension to small domains
- Extension to the ests of variances of level ests

Poverty Mapping for the Chilean Comunas

Partha Lahiri
JPSM, University of Maryland, College Park, USA

Livorno, Italy, June 16, 2015

[Based on joint work with Carolina Casas-Cordero and Jenny Encina]

Introduction

- The eradication of poverty has been at the center of various public policies in Chile and has guided public policy efforts.

- The nationwide survey estimate of the poverty rate has declined since the early 90´s suggesting some progress towards this goal. Erratic time series patterns, however, have emerged for small *comunas* - the smallest territorial entity in Chile.

- For a handful of extremely small comunas, survey estimates of poverty rates are unavailable for some or all time points simply because the survey design, which traditionally focuses on precise estimates for the nation and large geographical areas, excludes these comunas for some or all of the time points.

- Direct survey estimates of poverty rates typically do not meet the desired precision for small comunas and thus the assessment of implemented policies is not straightforward at the comuna level.

- In order to successfully monitor trends, identify influential factors, develop effective public policies and eradicate poverty at the comuna level, there is a growing need to improve on the methodology for estimating poverty rates at this level of geography.

- The need for socioeconomic data at lower levels of geography found its way into the Chilean legislation in 2007 when an amendment to the law of the *Fondo Común Municipal* (FCM) established a new set of indicators for its fund allocation algorithm among comunas. The regulation passed in 2009 required the Ministry to provide poverty rate estimates for all comunas in Chile.

- Regarding the production of comuna level estimates, an Expert Commission, appointed by the Ministry of Social Development (henceforth referred to as the *Ministry*) in 2010, raised concerns because of (1) the significant costs associated with sampling almost all comunas in the country, and (2) the relatively low precision for some comuna level estimates making the planned comparison among comunas and/or across time useless.

- The Commission recommended to (i) reduce the overall sample significantly, (ii) stop the production of comuna level direct estimates, and (iii) search for alternative data sources such as administrative records or develop a new data collection effort specifically designed for comuna level representation of social indicators of interest for various public policies.

- In 2010, the Ministry produced for the first time poverty rate estimates for all 345 comunas in Chile using both standard design-based and the Ministry-PNUD synthetic method.

The Poverty Measure Used in Chile

- In Chile, poverty is measured using the poverty rate, also known as Headcount Index, defined as the proportion of households with *income* below the *poverty threshold* or *poverty line*.

- The first ingredient of the poverty rate is the *poverty line*. For most Latin American countries, the poverty line is the cost of a basket of essential food and non-food items. This poverty line is expressed in per-capita terms. The methodology for estimating Chile´s poverty line was developed by the Comisión Económica para América Latina y el Caribe (CEPAL). Data from the Chilean expenditure survey *Encuesta de Presupuestos Familiares 1987-1988* was used to estimate the value of the food basket. Two different poverty lines were derived from the food basket ---- one for rural areas and the other for urban areas.

- The second ingredient of the poverty rate, the *per-capita income*, is the ratio of the *total household income* and the *household size*. Households whose per-capita income falls below the poverty line are considered in poverty. The poverty rate is then the percent of households in each region/comuna that are in poverty.

The Casen Survey

- Chile's official data source for poverty statistics is the National Socioeconomic Characterization Survey (Casen) - a survey sponsored every two or three years by the Ministry since 1987 with sample in most of the comunas.

- The Casen survey is a cross-sectional multipurpose household survey designed to understand the socioeconomic conditions of the population and the evaluation of social programs. The survey has been fielded regularly every two or three years since 1987.

- The 2009 Casen survey collected data from 246,924 persons in 71,460 households, representing a total of 16,607,007 persons living in private dwellings in Chile in November, 2009. The sampling design used was as follows:

  -The target population was defined to cover 334 out of the 345 comunas in the country.

  -Samples were drawn independently from 602 sampling strata formed by the comuna´s urban/rural subdivisions.

  -Using a two-stage sampling design, small geographic entities, known as *secciones*, were sampled at the first stage (Primary Sampling Units, PSUs) and housing units were sampled at the second stage (Secondary Sampling Units, SSUs) within each sampling strata.

  -The PSU´s were selected with probability proportional to size, measured in terms of the number of occupied housing units. A variable number of SSU´s were selected with equal probability using a systematic sampling algorithm with a random start within each selected PSU. Within each housing unit interviews were attempted with all households (*i.e.* no subsampling was implemented beyond the selection of the housing units).

Data Preparation

- Comuna level data derived from Casen 2009

  p_i : direct estimate of poverty rate for the *i*th comuna;

  $y_i = \sin^{-1}\sqrt{p_i}$; n_i : effective sample size

  $D_i = 1/(4n_i)$, an approximated sampling variance of y_i

- Comuna level administrative data

  -average wage for dependent workers

  -percentage of rural population

  -percentage of illiterate population

  -percentage of school attendance

  -the average of the comuna-level poverty rates from Casen 2000, 2003 and 2006

  -region-level indicators for the 7^{th}, 8^{th} and 9^{th} regions of the country

Description of SAE Method Implemented in Chile

Four Guidelines:

- method must use the Casen survey data directly to the extent possible since this is the largest data that collect information on most current poverty related variables
- poverty rate estimates should be close to the survey-weighted direct estimates for comunas with reasonably large samples
- method must not produce poverty rate estimates that considerably deviate from the corresponding direct survey estimates even for small comunas
- poverty count estimates, when aggregated over all the comunas in a given region, must produce the official survey-weighted count for that region.

Modeling

Level 1 (Sampling Model):

Given θ_i, y_i's are independent with $y_i \sim N(\theta_i, D_i)$;

Level 2 (Linking Model):

θ_i's are independent with $\theta_i \sim N(x_i{}'\beta, A)$,

- m is the number of comunas in Chile covered by Casen;
- $\theta_i = \sin^{-1}\sqrt{P_i}$; P_i is the true poverty rate;
- $x_i{}' = (x_{i0}, \cdots, x_{is-1})$ is a $s \times 1$ vector of s known fixed comuna specific auxiliary variables with $x_{i0} = 1$; $\beta = (\beta_0, \cdots, \beta_{s-1})$ is a $s \times 1$ column vector of unknown regression coefficients where β_0 denotes the intercept;
- A is the unknown model variance $(i = 1, \cdots, m)$.

Empirical Bayes Estimator of θ_i

Bayes estimator:

$$\hat{\theta}_i^B = (1 - B_i)y_i + B_i x_i{}'\beta, \text{ where } B_i = D_i/(A + D_i).$$

An Empirical Bayes (EB) estimator of θ_i:

$$\hat{\theta}_i^{EB} = (1 - \hat{B}_i)y_i + \hat{B}_i x_i{}'\hat{\beta}, \text{ where } \hat{B}_i = D_i\Big/\left(\hat{A} + D_i\right).$$

- The weight the EB estimator puts on the direct estimator y_i depends on the ratio $\hat{A}/D_i$.
- The choice of the adjusted maximum profile likelihood estimator of A over the usual residual maximum likelihood (REML) estimator was intentional and was used to assign more weight on the direct estimator since adjusted profile likelihood tends to have more upward bias than the REML.

- Since the adjusted maximum profile likelihood estimator is strictly positive, it avoids the common problem of the full shrinkage $(i.e., \hat{B}_i = 1)$ that is often encountered with the REML-based empirical Bayes estimator of θ_i.
- In theory, EB estimates can go out of the admissible range $[0, \pi/2]$. Thus, $\hat{\theta}_i^{EB}$ is truncated to 0 if $\hat{\theta}_i^{EB}$ is negative and to $\pi/2$ if $\hat{\theta}_i^{EB}$ is greater than $\pi/2$.

Limited Translation Empirical Bayes Estimator of θ_i

$$\hat{\theta}_i^{LT} = \begin{cases} \hat{\theta}_i^{EB} & \text{if} \quad y_i - \sqrt{D_i} \le \hat{\theta}_i^{EB} \le y_i + \sqrt{D_i}, \\ y_i - \sqrt{D_i} & \text{if} \quad \hat{\theta}_i^{EB} \le y_i - \sqrt{D_i}, \\ y_i + \sqrt{D_i} & \text{if} \quad \hat{\theta}_i^{EB} \ge y_i + \sqrt{D_i}, \end{cases}$$

Back-transformation and raking

Back-transform: $\hat{P}_i = \sin^2 \hat{\theta}_i^{LT}$.

For a few comunas with no sample in the Casen 2009 survey, the estimates of the poverty rate were computed using the Ministry-PNUD synthetic method.

Whether a comuna is in the Casen sample or not, the final official raked SAE estimates of poverty rates for all the comunas that belong to the *r*th region are given by:

$$\hat{P}_i^{SAE} = \hat{P}_i \times R_r,$$

where

- $R_r = p_r^{regn} N_r^{regn} \Big/ \sum_{i=1}^{m_r^*} \hat{P}_i N_i$ is the raking factor common to all comunas in the region $r; m_r^*$ is the total number of comunas in region r; p_r^{regn} is the direct design-based estimate of the regional-level poverty rate using the original regional weights; N_i is an estimate of the population projection in comuna i belonging to region $r; N_r^{regn}$ is an estimate of the population projection in region $r;$ $N_r^{regn} = \sum_{i=1}^{m_r^*} N_i$.

Confidence Intervals for the Poverty Rates

Step 1: Generate R independent parametric bootstrap samples$\{(y_i^{(r)}, \theta_i^{(r)}),\ i=1,\cdots m\}$, $r=1,\cdots,R$ as follows: $\theta_i^{(r)} \sim N(x_i^T\hat{\beta},\hat{A}),\ y_i^{(r)} \mid \theta_i^{(r)} \sim N(\theta_i^{(r)}, D_i),\ i=1,\ldots,m.$

Step 2: Produce estimates $\hat{A}^{(r)}$, $\hat{B}_i^{(r)}$ and $\hat{\beta}^{(r)}$ by replacing the original data with the parametric bootstrap samples generated in Step 1. We repeat this step R times.

Step 3: For each bootstrap simple, calculate the following pivotal quantity: $t_i^{(r)} = \left(\theta_i^{(r)} - \hat{\theta}_i^{EB(r)}\right) \Big/ \sqrt{D_i(1-\hat{B}_i^{(r)})}$, where $\hat{\theta}_i^{EB(r)} = (1-\hat{B}_i^{(r)})y_i^{(r)} + \hat{B}_i^{(r)} x_i{}' \hat{\beta}^{(r)}.$

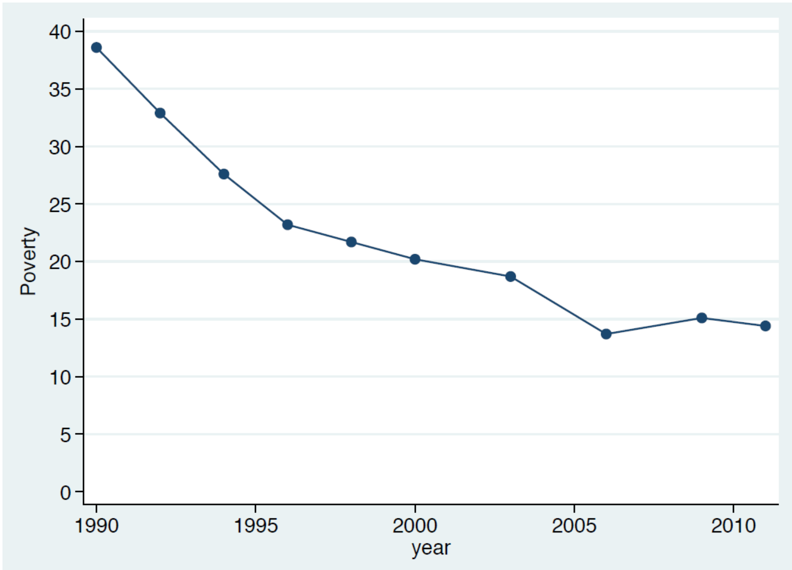
Step 4: For comuna i, obtain q_{1i} and q_{2i}, the $100\alpha/2$ and $100(1-\alpha/2)$ percentiles of $\{t_i^{(r)},\ r=1,\cdots,R\}$.

Step 5: For comuna i, an approximate $100(1-\alpha)\%$ confidence interval for θ_i is obtained as: $(L_i, U_i),$ where $L_i = \hat{\theta}_i^{EB} + q_{1i}\sqrt{D_i(1-\hat{B}_i)}$ and $U_i = \hat{\theta}_i^{EB} + q_{2i}\sqrt{D_i(1-\hat{B}_i)}$. Note that the admissible range for θ_i is $[0, \pi/2]$. Thus, L_i is truncated to 0 if L_i is negative and U_i is truncated to $\pi/2$ if U_i is greater than $\pi/2$. The probability that (L_i, U_i) is not contained in $(0, \pi/2)$ is expected to be negligible unless $4n_i$ is very small. The truncated confidence interval for θ_i is denoted by (L_i^*, U_i^*).

Step 6: Finally, the lower and upper limits of the confidence interval (L_i^*, U_i^*) in Step 5 are back-transformed to yield the following approximate $100(1-\alpha)\%$ confidence interval of the poverty rate P_i : $\left(\sin^2 L_i^*,\ \sin^2 U_i^*\right)$. Note that the parametric bootstrap confidence interval for any one-to-one transformed parameter can be easily obtained using the simple back-transformation. In our case, the motivation for this back-transformed confidence interval comes from the fact that for any $0 < p < 1$ and $0 < \theta < \pi/2$, $\sin^{-1}\sqrt{p}$ and $\sin^2\theta$ are monotonically increasing functions of p and θ, respectively.

Appendix

Figure 21.1.
Estimates of national poverty rates in Chile, by year.



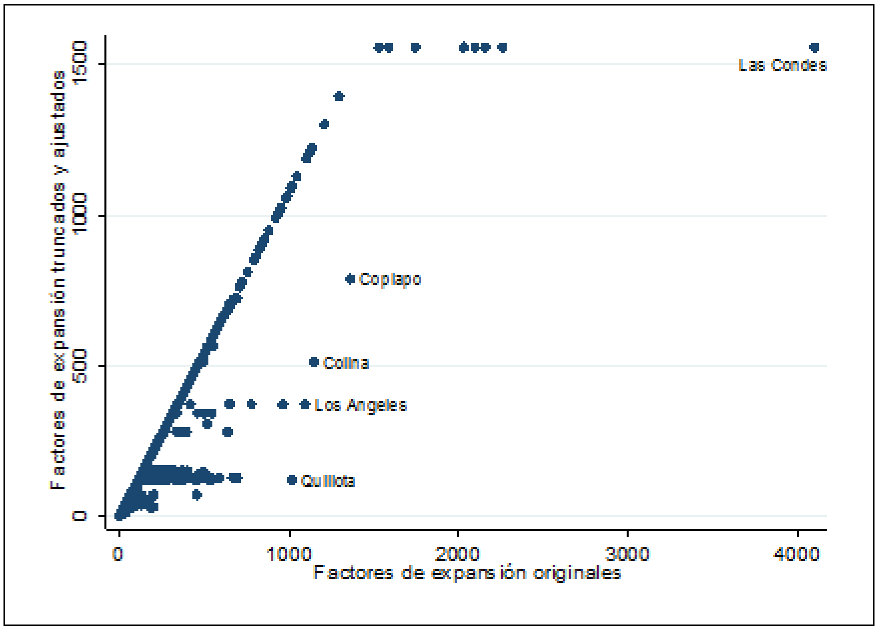
Source: Compiled by the authors based on Casen 1990, 1992, 1994, 1996, 1998, 2000, 2003, 2006, 2009 and 2011 data.

Figure 21.2.
Descriptive statistics for the original survey weights, cut-off point and total number of original comuna weights truncated, by region and zonal group. Casen 2009 data.

Truncation Groups	Descriptive statistics original comuna weights			Truncation point	Number of original comuna weights truncated
	Average	Minimum	Maximum		
1	87.4902	5	501	137.6	748
2	10.1405	2	34	63.0	0
3	94.3949	3	692	672.7	32
4	4.5907	1	27	32.5	0
5	59.8353	6	1.363	731.7	12
6	14.3449	4	47	83.4	0
7	95.7634	7	558	524.6	82
8	27.3082	4	134	127.6	58
9	74.0826	5	1.020	112.6	4,117
10	23.0577	3	100	75.7	105
11	53.7106	2	637	262.0	229
12	22.6640	2	110	87.7	171
13	69.0955	3	405	141.5	1,471
14	26.0215	4	160	49.8	1,760
15	66.7037	4	1,095	346.3	105
16	20.2520	4	203	30.6	2,929
17	60.0223	4	548	318.7	225
18	28.2342	6	183	37.2	2,152
19	70.5936	2	693	118.5	1,547
20	22.8171	3	461	67.9	673
21	37.9711	5	183	52.8	497
22	9.9334	3	34	11.8	328
23	85.7458	8	544	777.8	0
24	9.8642	2	41	14.1	67
25	147.5910	5	4,103	1,445.2	81
26	38.0404	2	1,147	476.6	18
27	53.8487	6	520	287.1	86
28	31.1182	6	237	135.4	54
29	106.1280	1	475	133.2	268
30	14.3313	1	57	103.1	0
Total	-	-	-	-	17,815

Source: Ministerio de Desarrollo Social [42].

Figure 21.3.
A plot of original weights (x-axis) and trimmed survey weights (y-axis) for all observations in Casen 2009.



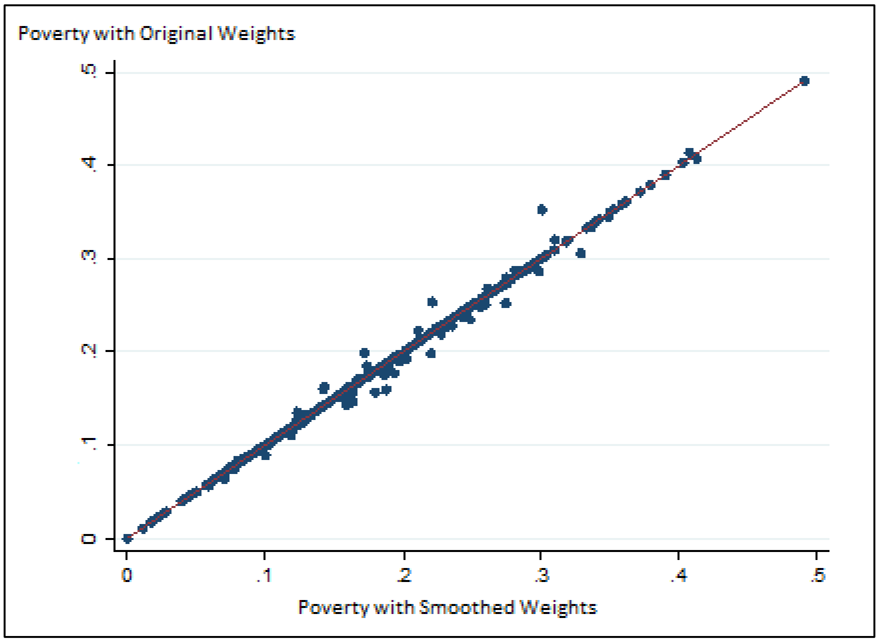
Source: Ministerio de Desarrollo Social [42].

Figure 21.4.
Descriptive statistics of number of cases at the Respondent level and the Household level. Casen 2009 data.

Quartiles of comunas respondent sample	Respondent level Sample			Household level Sample		
	Min	Mean	Max	Min	Mean	Max
1	53	491.0	610	20	152.7	198
2	612	654.9	692	155	195.7	239
3	693	752.2	864	177	214.1	265
4	873	1,064.3	1,608	211	294.4	409

Source: Compiled by authors based on Casen 2009 data.

Figure 21.5.
A plot of direct survey estimates with original survey weights (x-axis) and trimmed survey weights (y-axis) for comunas in Chile. Casen 2009 data.



Source: Ministerio de Desarrollo Social [42].

Figure 21.6.
Estimates of the design effect of the direct poverty rate in Chile using trimmed comuna weights, by region. Casen 2009 data.

N°	Region	Design Effect Estimates
1	Tarapacá	3.280
2	Antofagasta	5.750
3	Atacama	6.477
4	Coquimbo	4.665
5	Valparaíso	3.390
6	O´Higgins	4.307
7	Maule	4.870
8	Biobío	5.506
9	Araucanía	5.618
10	Los Lagos	6.095
11	Aysén	2.843
12	Magallanes	2.323
13	Metropolitana	3.290
14	Los Ríos	8.681
15	Arica y Parinacota	2.864

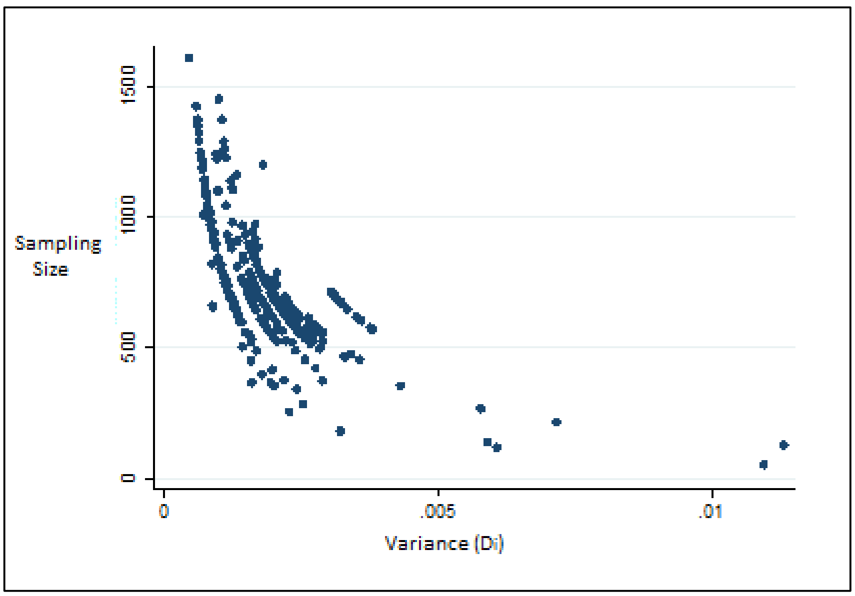
Source: Ministerio de Desarrollo Social [42].

Figure 21.7.
Descriptive statistics of D_i and B_i by groups of comunas formed using quartiles of the distribution of the sampling variances (D_i). Casen 2009 data.

Quartiles of D_i	Descriptive statistics of D_i			Descriptive statistics of B_i		
	Mean	Minimum	Maximum	Mean	Minimum	Maximum
Group 1 (lowest 25% of D_i)	0.0009047	0.0004453	0.0012006	0.2765	0.1598	0.3390
Group 2	0.0014632	0.0012023	0.0016854	0.3836	0.3393	0.4186
Group 3	0.0019475	0.001686	0.0021987	0.4535	0.4186	0.4843
Group 4 (highest 25% of D_i)	0.0030616	0.0022082	0.0113181	0.5474	0.4854	0.8286
All	0.0018417	0.0004453	0.0113181	0.4150	0.1598	0.8286

Source: Ministerio de Desarrollo Social [42].

Figure 21.8.
Comuna-level sample size (y-axis) and comuna-level estimate of variance (x.-axis) of the direct estimate of the poverty rates. Casen 2009 data.



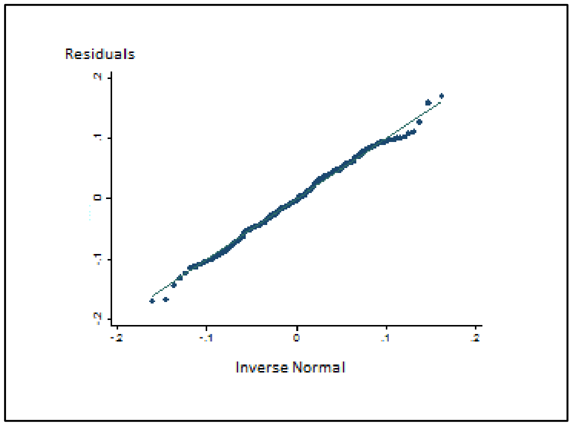
Source: Ministerio de Desarrollo Social [42].

Figure 21.9.
Initial set of auxiliary variables reviewed for their possible inclusion as comuna-level auxiliary variables in the area level model.

Number and Name of the auxiliary variable	Institution responsible for data collection	Frequency of publication of the data
#1. Subsidio Familiar	Unidad de Prestaciones Monetarias, Ministerio de Desarrollo Social.	monthly and yearly
#2. Subsidio al Pago del Consumo de Agua Potable y Servicio de Alcantarillado de Aguas Servidas	Unidad de Prestaciones Monetarias, Ministerio de Desarrollo Social.	monthly and yearly
#3. Bono Chile Solidario	Unidad de Prestaciones Monetarias, Ministerio de Desarrollo Social.	monthly and yearly
#4. Subsidio de Discapacidad Mental	Unidad de Prestaciones Monetarias, Ministerio de Desarrollo Social.	monthly and yearly
#5. Pensión Básica Solidaria (vejez e invalidez)	Unidad de Prestaciones Monetarias, Ministerio de Desarrollo Social.	December
#6. Aporte Previsional Solidario (vejez e invalidez)	Unidad de Prestaciones Monetarias, Ministerio de Desarrollo Social.	December
#7. Bonificación al Ingreso Ético Familiar	Unidad de Prestaciones Monetarias, Ministerio de Desarrollo Social.	monthly and yearly
#8. Beca de Apoyo a la Retención Escolar, BARE	Unidad de Prestaciones Monetarias, Ministerio de Desarrollo Social.	monthly and yearly
#9. Afiliados Sistema de Capitalización Individual	Superintendencia de Pensiones	monthly and yearly
#10. Matrícula	Ministerio de Educación	Yearly
#11. Rendimiento	Ministerio de Educación	Yearly
#12. SIMCE	Ministerio de Educación	Yearly or every two years
#13. Titulados Educación Superior	Ministerio de Educación	Yearly
#14. Índice de Vulnerabilidad del Establecimiento (IVE-SINAE)	Junta Nacional Escolar y Becas (Junaeb)	Yearly
#15. Situación Nutricional estudiantes básica y media	Junta Nacional Escolar y Becas (Junaeb)	Yearly
#16. Población beneficiaria Fonasa	Ministerio de Salud	Yearly
#17. Atenciones sector privado	Ministerio de Salud	Yearly
#18. Razón de analfabetos respecto a la población de 10 y más años en la comuna	CENSO, INE	Every 10 years
#19. Porcentaje de Población Rural	CENSO, INE	Every 10 years
#20. Porcentaje de Asistencia Escolar Comunal	SINIM	monthly
#21. Tamaño promedio del hogar	CENSO, INE	Every 10 years
#22. Tasa de pobreza histórica	CASEN	Every 2 or 3 years
#23. Contribuciones de Vivienda	SII (http://www.sii.cl/avaluaciones/estadisticas/estadisticas_bbrr.htm#2)	Yearly
#24. Remuneraciones promedio de los trabajadores dependientes		Yearly

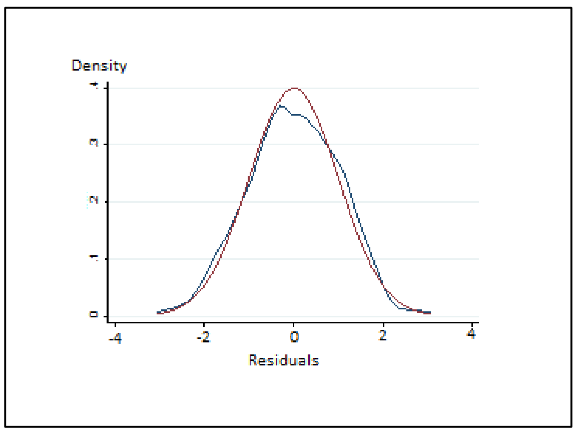
Source: Ministerio de Desarrollo Social [42].

Figure 21.10
QQ plot of the standardized residuals



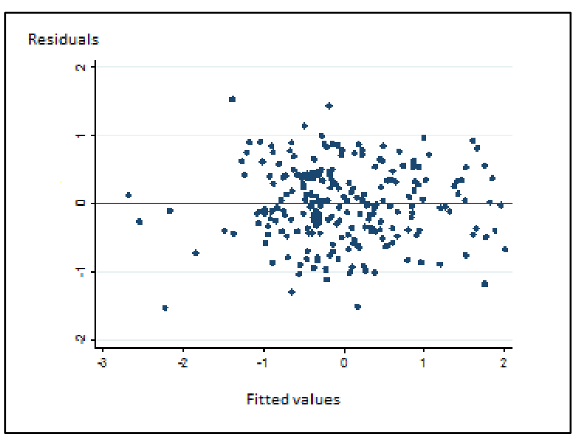
Source: Ministerio de Desarrollo Social [42].

Figure 21.11
Distribution of the standardized residuals (blue line)



Source: Ministerio de Desarrollo Social [42].

Figure 21.12
Plot of standardized residuals against fitted values



Source: Ministerio de Desarrollo Social [42].

Figure 21.13.
Estimates of Spearman correlation coefficients and p-values for the squared standardized residuals of the OLS regression model in Figure 21.14.

Auxiliary Variable	Sperarman Correlation	P-values
Average wage of dependent workers	-0.0144	0.8264
Average of the poverty rate from Casen 2000, 2003 and 2006	-0.0148	0.8214
% of population in rural areas	-0.0065	0.9214
% of illiterate population	-0.0092	0.8882
% of population attending school	0.0072	0.9126
Dummy for region 7	0.0337	0.607
Dummy for region 8	-0.095	0.1467
Dummy for region 9	-0.0061	0.9256

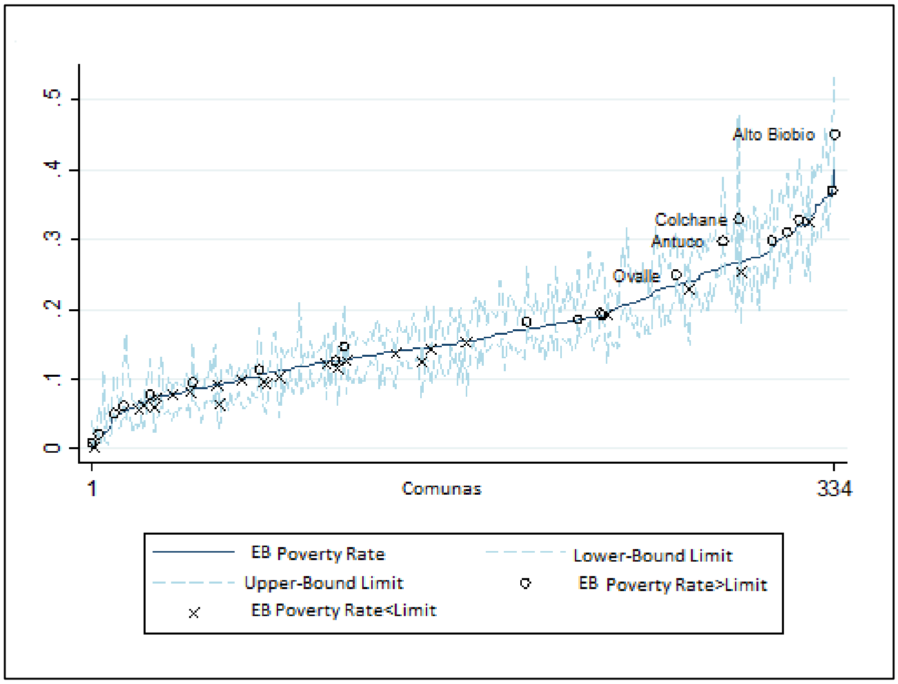
Source: Ministerio de Desarrollo Social [42].

Figure 21.14.
Output of regression analysis based on comunas with population more than 10.000 inhabitants (dependent variable: arcsine transformed direct survey estimate of the poverty rate with original and trimmed weights; independent variables: a set of variables used in the comuna level model with arcsine transformation for proportions and logarithmic transformation for the rest).

Independent variables	Regression coefficient estimate (t-statistics): original comuna weights	Regression coefficient estimate (t-statistics): trimmed comuna weights
Average wage of dependent workers (log)	-0.09575646 (3.52**)	-0.21927953 (3.52**)
Average of the poverty rate from Casen 2000, 2003 and 2006 (arcsin)	0.49548266 (7.92**)	0.48474029 (7.92**)
% of population in rural areas (arcsin)	-0.13409847 (4.96**)	-0.39252745 (4.96**)
% of illiterate population (arcsin)	0.40349163 (2.57*)	0.25176513 (2.57*)
% of population attending to school (arcsin)	-0.21883535 (2.23*)	-0.0938032 (2.23*)
Dummy for region 7 (=1)	0.03442978 (2.11*)	0.08671043 (2.11*)
Dummy for region 8 (=1)	0.03882056 (2.67**)	0.12474226 (2.67**)
Dummy for region 9 (=1)	0.105632 (6.04**)	0.28328927 (6.04**)
Constant	1.61477028 (4.24**)	-0.00203088 (0.06)
Number of observations	235	235
Adjusted R^2	0.67	0.67

Notes: * statistically significant at the 5% level; ** statistically significant at the 1% level.
Source: Ministerio de Desarrollo Social [42].

Figure 21.15
Limited translation empirical Bayes estimates of the comuna level poverty rates, and the upper and lower thresholds.



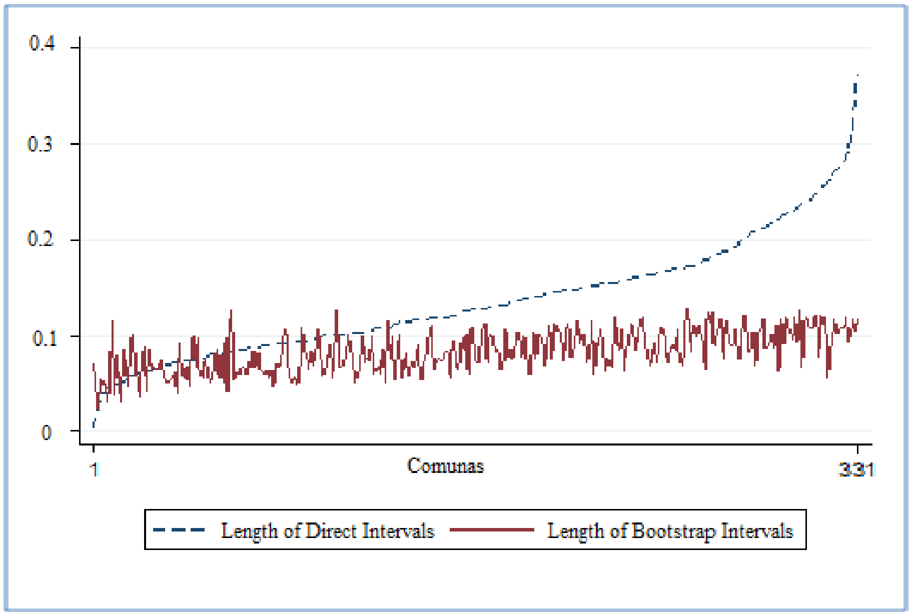
Source: Ministerio de Desarrollo Social [42].

Figure 21.16.
Raking factors used so that the total of the model-based estimates within each region matches the standard design-based official estimate for the region, by region.

N°	Region	R_r
1	Tarapacá	1.12172
2	Antofagasta	0.97455
3	Atacama	1.06685
4	Coquimbo	1.04309
5	Valparaíso	1.00387
6	O´Higgins	1.00430
7	Maule	1.05292
8	Biobío	0.99010
9	Araucanía	1.01628
10	Los Lagos	1.04088
11	Aysén	1.06255
12	Magallanes	0.97368
13	Metropolitana	0.97765
14	Los Ríos	1.08572
15	Arica y Parinacota	0.99486

Source: Ministerio de Desarrollo Social [42].

Figure 21.17.
Length of the direct and parametric bootstrap confidence intervals of the comuna-level poverty rates for comunas sorted by the limited translation empirical Bayes estimates of the poverty rate.



Note: The three comunas with the largest estimates of the length of the direct confidence interval were excluded from the graph. Source: Compiled by authors based on Casen 2009 data and Ministerio de Desarrollo Social [43].

Figure 21.18a.
Histograms of pivots in the parametric bootstrap method with 5,000 bootstrap samples.
Comuna of Puchucavi

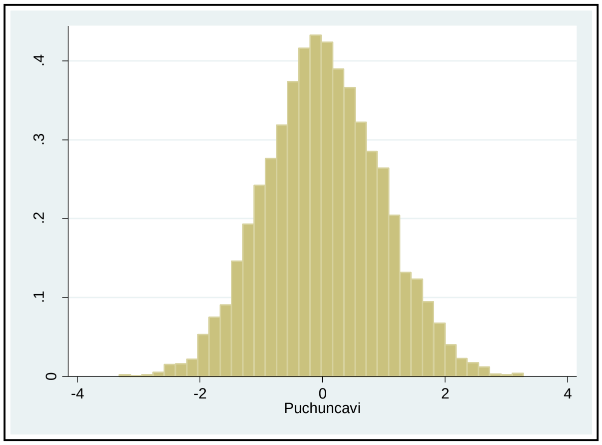


Figure 21.18b.
Histograms of pivots in the parametric bootstrap method with 5,000 bootstrap samples.
Comuna of Providencia

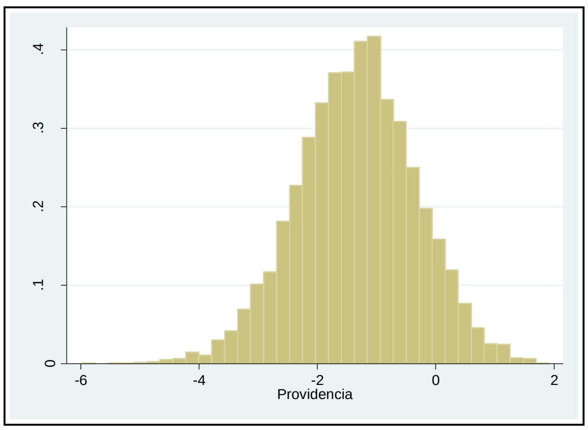
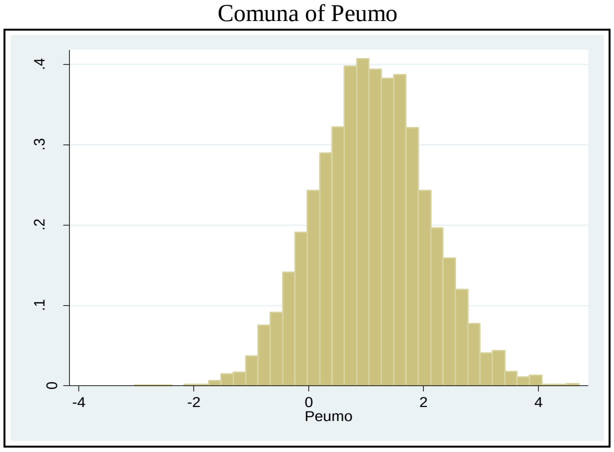


Figure 21.18c.
Histograms of pivots in the parametric bootstrap method with 5,000 bootstrap samples.
Comuna of Peumo



Source: Ministerio de Desarrollo Social [42].

References

1. Ley 18.695 Art.1. *Fija el Texto Refundido, Coordinado y Sistematizado de la Ley Nº 18.695, Orgánica Constitucional de Municipalidades*. Ministerio del Interior; Subsecretaría de Desarrollo Regional y Administrativo. Santiago, Chile, Publicación: 26-07-2006.
2. Ministerio de Desarrollo Social (2012). *Metodología del Diseño Muestral y Factores de Expansión Encuesta de Caracterización Socioeconómica Nacional*. Serie de Documentos Metodológicos N° 1, Observatorio Social.
3. United Nations Development Programme (UNDP) (1990). *Human Development Report 1990*. New York Oxford. Oxford University Press.
4. Programa de las Naciones Unidas para el Desarrollo (PNUD) Chile y Ministerio de Planificación de Chile (2000). *Desarrollo Humano en las comunas de Chile,* Santiago, Chile.
5. Ley 20.237 Art.1 (1). *Modifica el Decreto Ley Nº 3.063, de 1979, Sobre Rentas Municipales la Ley Nº 18.695, Orgánica Constitucional de Municipalidades, y Otros Cuerpos Legales, en Relación con el Fondo Común Municipal y Otras Materias Municipales*. Ministerio del Interior; Subsecretaría de Desarrollo Regional y Administrativo. Santiago, Chile, Publicación: 17-12-2007.
6. Decreto 1293 Art. 13(f). *Reglamento para la Aplicación Del Artículo 38 Del Decreto Ley N° 3.063, De 1979, Modificado por el Artículo 1° de la Ley N° 20.237*. Ministerio del Interior; Subsecretaría de Desarrollo Regional y Administrativo. Santiago, Chile, Publicación: 02-01-2009.
7. Comisión de Técnicos de la Encuesta Casen (2010). *Informe de la Comisión de Técnicos sobre la Encuesta Casen*. Santiago, Chile.
8. United Nations Expert Group on Poverty Statistics (Rio Group) (2006). *Compendium of Best Practices in Poverty Measurement*. United Nations Economic Commission for Latin America and the Caribbean (ECLAC) and Brazilian Institute for Geography and Statistics (IBGE). Rio de Janeiro, September 2006. ISBN 85-240-3908-6.
9. Comisión para la Medición de la Pobreza (2014). *Informe Final de la Comisión para la Medición de la Pobreza*. Santiago, Chile.
10. Comisión Económica para América Latina y el Caribe (CEPAL) (1990). *Una estimación de la medición de pobreza para Chile, 1987*, Santiago de Chile.
11. Ministerio de Planificación (2010). *Informe Metodológico Casen 2009*. División Social, Santiago, Chile.
12. Potter F.J. (1993). *The effect of weight trimming on nonlinear survey estimates*, San Francisco, CA: American Statistical Association.
13. Bell, W. (1997). *Models for county and state poverty estimates*. Preprint, Statistical Research Division, U. S. Census Bureau.
14. Bell, W., Basel, W., Cruse, C, Dalzell, L., Maples, J., OHara, B., and Powers, D. (2007). *Use of ACS Data to Produce SAIPE Model-Based Estimates of Poverty for Counties*, Census Report.
15. National Research Council (2000). *Small-area estimates of school-age children in poverty: Evaluation of Current Methodology*. Citro, C. and Kalton, G. (Eds.), National Academy Press, Washington DC.
16. Efron, B. and Morris, C. (1975). *Data analysis using Stein's estimator and its generalizations*, Journal of the American Statistical Association 70, 311-319.
17. Carter, G. and Rolph, J. (1974). *Empirical Bayes methods applied to estimating fire alarm probabilities*, Journal of the American Statistical Association 69, 880-885.
18. Jiang, J., Lahiri, P., Wan, S., and Wu, C. (2001). *Jackknifying in the Fay-Herriot Model with an example*, Proceedings of the Seminar on Funding Opportunity in Survey Research, Council of Professional Associations on Federal Statistics.
19. Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W., and Feuer, E.J. (2007). *Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening*. Journal of the American Statistical Association 102(478): 474.
20. Xie, D., Raghunathan, T.E., and Lepkowski, J.M. (2007). *Estimation of the proportion of overweight individuals in small areas - a robust extension of the Fay-Herriot model*. Statistics in Medicine 26(13): 2699-2715.
21. StataCorp. 2009. *Stata Statistical Software: Release 11*. College Station, TX: StataCorp LP.
22. Box, G.E.P. (1979). *Robustness in the Strategy of Scientific Model Building* in Robustness in Statistics: Proceedings of a Workshop (1979) edited by RL Launer and GN Wilkinson.

23. Elbers, C., Lanjouw, J., and Lanjouw, P. (2003). *Micro-Level Estimation of Poverty and Inequality*, Econometrica 71:1 (2003), 355–364.
24. Elbers, C., Lanjouw, P., and Leite, P. (2008). *Brazil within Brazil: Testing the Poverty Map Methodology in Minas Gerais*, World Bank policy research working paper no. 4513.
25. Molina, I. and Rao, J.N.K. (2010). *Small area estimation of poverty indicators*, Canadian Journal of Statistics, 38, 369-385.
26. Agostini, C., Brown, P., and Gongora, D. (2008). *Distribución Espacial De La Pobreza En Chile*. Estudios De Economía, 35, N.1, 79-110.
27. Agostini, C., Brown, P., and Roman, A. (2010). *Estimando Indigencia y Pobreza Indígena Regional con Datos Censales y Encuestas de Hogares*. Cuadernos de Economía, 47(135), 125-150.
28. Rao, J. N. K. (2003). *Small Area Estimation*, Wiley, New York.
29. Pfeffermann, D. (2013). *New Important Developments in Small Area Estimation*, Statistical Science, 28, 40-68.
30. Fay, R.E. and Herriot, R.A. (1979). *Estimates of income for small places: An application of James_Stein procedure to census data*, Journal of the American Statistical Association 74, 269_277.
31. Li, H. and Lahiri, P. (2010). *An adjusted maximum likelihood method for solving small area estimation problems*. J. Multivariate Anal. 101 882–892. MR2584906
32. Efron, B. and Morris, C. (1972). *Limiting the risk of Bayes and empirical Bayes estimators-Part II: the empirical Bayes case.* Journal of the American Statistical Association 67: 130-139.
33. Chatterjee, S., Lahiri, P., and Li, H. (2008). *Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models*, The Annals of Statistics 36, 1221-1245.
34. Brown, L.D., Cai, T.T. and DasGupta, A. (2001). Interval Estimation for a Binomial Proportion, Statistical Science, 16, 101-133.
35. Ha, N. and Lahiri, P. (2014a). *Hierarchical Bayesian Estimation of Small Areas Proportions Using Complex Survey Data*, unpublished manuscript.
36. Pfeffermann, D., Sikov, A., and Tiller, R. (2014). Single and two-stage cross-sectional and time series benchmarking procedures for small area estimation, *Test*, DOI: 10.1007/s11749-014-0400-8
37. Ha, N. and Lahiri, P. (2014b). *Comment: Single and two-stage cross-sectional and time series benchmarking procedures for small area estimation*, *Test*, DOI: 10.1007/s11749-014-0400-8
38. Chatterjee, S. and Lahiri, P. (2013), *A Simple Computational Method for Estimating Mean Squared Prediction Error in General Small-Area Model*, paper presented at the SAE 2013, Bangkok.
39. Alkire, S. and Foster, J. (2007). *Counting and Multidimensional Poverty Measurement*. OPHI Working Paper Series, N°7.5, OPHI.
40. Statistics Netherlands (2008). *Model-Based Estimation for Official Statistics*, Discussion paper (08002), The Hage, Netherlands.
41. Heady, P. and Ralphs, M. (2005). *EURAREA: an overview of the project and its findings*. Statistics in Transition 7, 557—570.
42. Ministerio de Desarrollo Social (2013a). *Procedimiento de cálculo de la Tasa de Pobreza a nivel Comunal mediante la aplicación de Metodología de Estimación para Áreas Pequeñas (SAE)*. Serie de Documentos Metodológicos N° 1, Observatorio Social.
43. Ministerio de Desarrollo Social (2013b). *Incidencia de la Pobreza a nivel Comunal, según Metodología de Estimación para Áreas Pequeñas. Chile 2009 y 2011*. Serie de Informes Comunales N° 1, Observatorio Social.

An Evaluation of Different Small Area Estimators and Benchmarking for the Annual Survey of Public Employment and Payroll

Bac Tran
Governments Division
U.S. Census Bureau
Joint work with Partha Lahiri, JPSM
University of Maryland, College Park, U.S.A

Target Population

- Individual governments

  A government is an organized entity which, in addition to having governmental character, has sufficient discretion in the management of its own affairs to distinguish it as separate from the administrative structure of any other governmental unit

- Types
  - Counties
  - Municipalities
  - Townships
  - Special Districts
  - Schools Districts

United States™ Census Bureau
Economic Statistics

Parameters of Interest
Annual Survey of Employment and Payroll (ASPEP)

Full-time Employees

Full-time Pay

Part-time Employees

Part-time Pay

Part-time Hours

Parameters of Interest (Cont'd)
ASPEP Publication

Statistics on the number of federal, state, and local government employees and their gross payrolls

http://www2.census.gov/govs/apes/10locmd.txt
2010 Public Employment and Payroll Data
Local Governments
MARYLAND

SOURCE: 2010 Annual Survey of Public Employment and Payroll. For information on sampling and nonsampling errors and definitions, see http://www.census.gov/govs/apes/how_data_collected.html. Data users who create their own estimates from these tables should cite the U.S. Census Bureau as the source of the original data only.

Government Function	Full-time employees	Full-time pay ($)	Part-time employees	Part-time pay ($)	Full-Time Equivalent Employment	Total March Pay ($)
Total	189,620	984,236,113	59,634	89,231,689	214,213	1,073,467,802
Financial Administration	2,285	11,454,282	147	268,486	2,350	11,722,768
Other Government Administration	3,300	16,966,287	844	1,565,802	3,692	18,532,089
Judicial and Legal	3,233	16,149,220	363	681,272	3,438	16,830,492
Police Protection Total	15,983	93,050,897	1,381	1,603,148	16,620	94,654,045
Police Officers Only	12,278	75,342,746	148	249,672	12,362	75,592,418
Other Police Employees	3,705	17,708,151	1,233	1,353,476	4,258	19,061,627
Fire Protection Total	6,772	40,058,581	153	252,374	6,845	40,310,955
Firefighters Only	6,222	37,071,603	43	52,648	6,242	37,124,251
Other Fire Employees	550	2,986,978	110	199,726	603	3,186,704
Corrections	3,559	17,501,794	73	144,404	3,608	17,646,198
Highways	5,267	21,153,791	99	165,216	5,313	21,319,007
Air Transportation	39	150,081	45	46,878	56	196,959
Water Transport and Terminals	3	18,757	8	3,388	5	22,145
Public Welfare	2,579	11,536,891	1,321	2,456,860	3,455	13,993,751
Health	3,934	18,597,016	1,114	2,591,935	4,706	21,188,951

Parameters of Interest
Statistical Aggregation

- Totals

  by (state, function)

- Level of government totals
  - Local, state, state and local
  - Nation

Parameters of Interest (Cont'd) Some Function Codes of ASPEP

001, Airport

002, Space Research & Technology (Federal)

005, Correction

006, National Defense and International Relations (Federal)

012, Elementary and Secondary - Instruction

112, Elementary and Secondary - Other Total

014, Postal Service (Federal)

016, Higher Education - Other

018, Higher Education - Instructional

021, Other Education (State)

022, Social Insurance Administration (State)

023, Financial Administration

024, Firefighters

124, Fire - Other

025, Judicial & Legal

029, Other Government Administration

032, Health

040, Hospitals

044, Streets & Highways

050, Housing & Community Development (Local)

052, Local Libraries

059, Natural Resources

061, Parks & Recreation

062, Police Protection - Officers

162, Police-Other

079, Welfare

080, Sewerage

081, Solid Waste Management

087, Water Transport & Terminals

089, Other & Unallocable

090, Liquor Stores (State)

091, Water Supply

092, Electric Power

093, Gas Supply

094, Transit

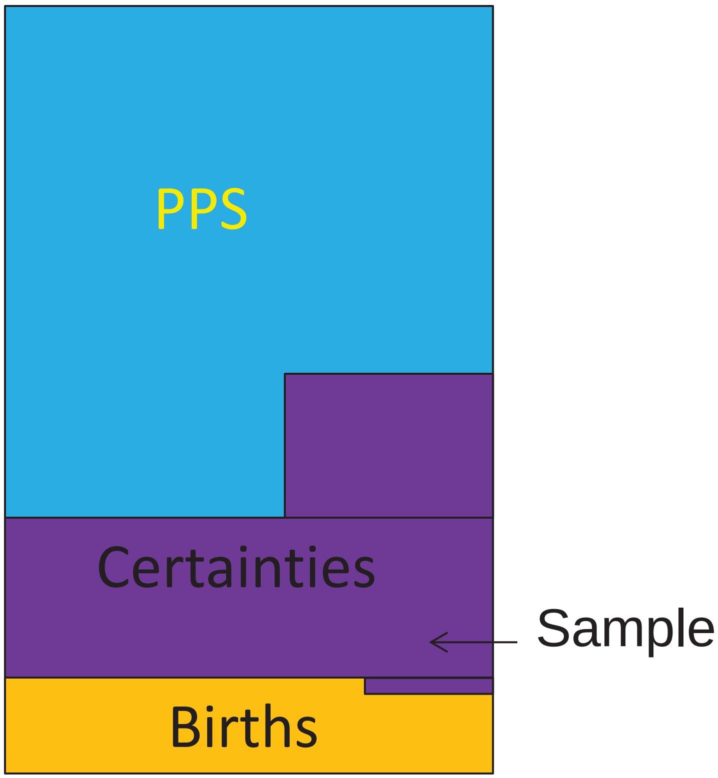
Sample Design

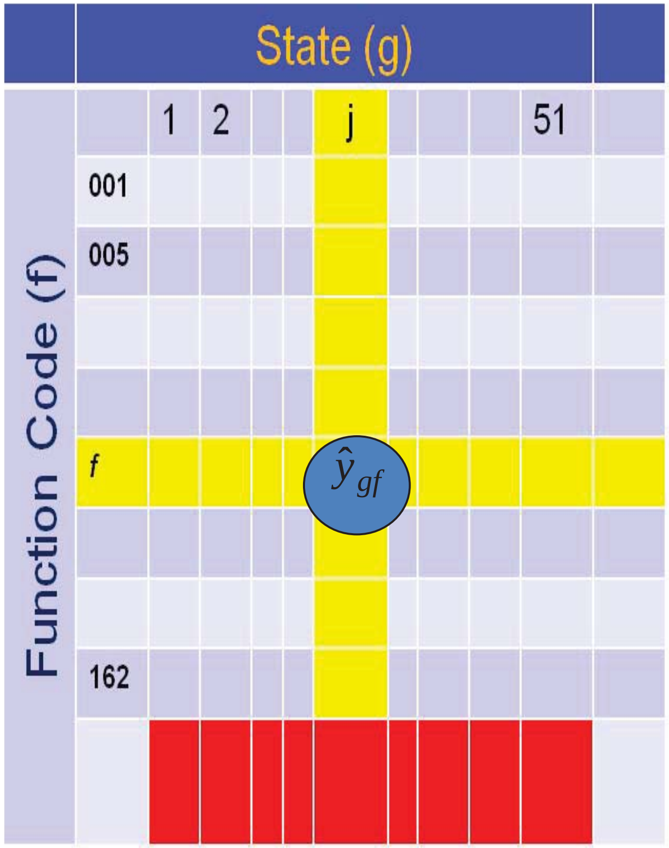
Multistage sample design

- PPS sample
  - Stratified PPS (state x type) based on Total Pay
- Cut-off sampling method in sizable (state, type) strata
  - Construct a cut-off point to determine small and large size units (two strata)
- Modified cut-off sampling (a stratified PPS sample method)
  - Sub-sampling on small strata

Sample

Sampling Frame





Small Area Challenge

- Designed at (state, type) level, estimated at function level

- Estimate the total of employees and payroll at state by function level

$$Y_{gf} = \sum_{i \in U_{gf}} Y_{gfi} \ \ where\ g = state, and\ f = function$$

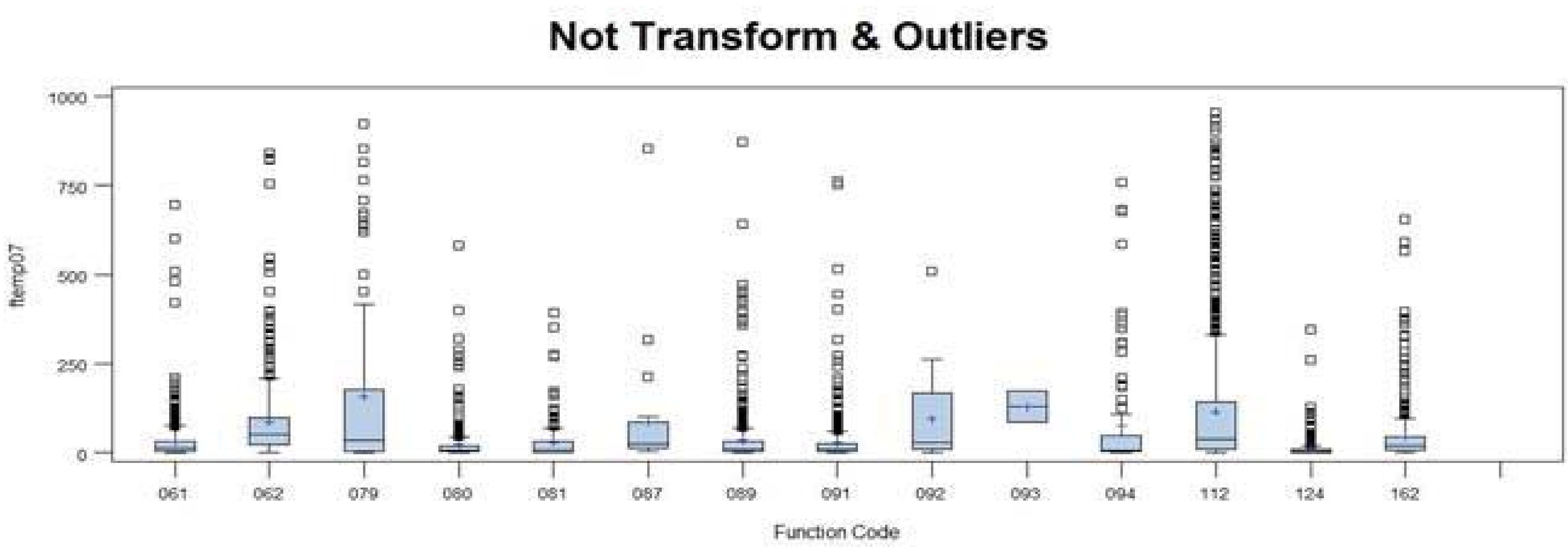
Small Area Challenge (Cont'd)

- Small area: a small geographic area within a larger geographic area or a small demographic group within a larger group

- Most small area estimation methods borrow strength from related or similar small areas using auxiliary data

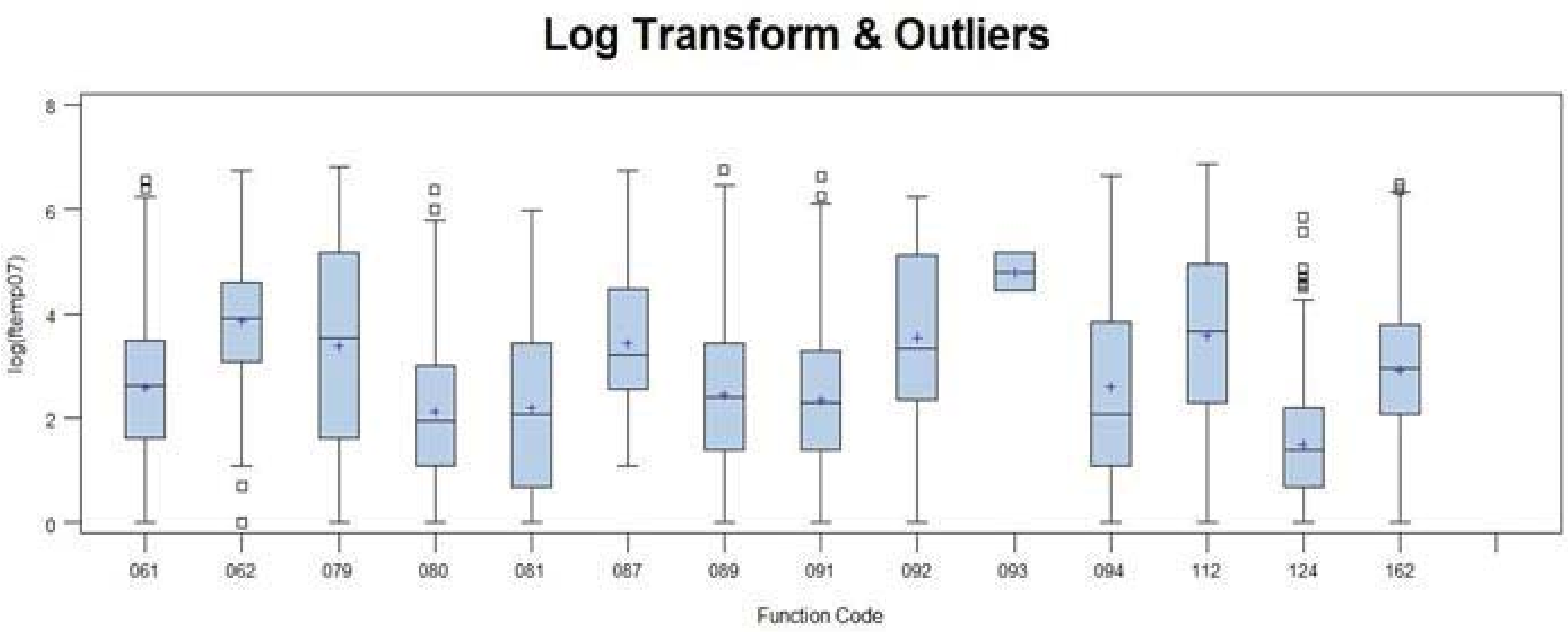
Other Challenges

Figure 1: Skew data -Not Transform (California)
(Full-Time Employees, Function)



Other Challenges (Cont'd)

Figure 2: Skew data - Log Transform (CA)
(Log(Full-Time Employees), Function)



United States™ Census Bureau
Economic Statistics

Estimators- ASPEP

- Direct

  → Horvitz-Thompson: $\hat{y}_{gf}^{HT} = \sum w_{gf} y_{gf}$

- Battese, Harter, Fuller (BHF) Model

- Our Proposed Model

Estimators (Cont'd)
Battese, Harter, Fuller (BHF) Model

$$y_{ij} = \beta_0 + \beta_1 x_i + v_i + \varepsilon_{ij}$$

y_{ij} : the number of full-time employees for the j^th governmental unit within the i^th small area

x_i : number of full-time employees for the *i*^th small area obtained from the previous census

β_0, and β_1: unknown intercept and slope, respectively; v_i are small area specific random effects

ε_{ij} : errors in individual observations

Estimators (Cont'd)
Our Proposed Model

$$\log(y_{ij}) = \beta_0 + \beta_1 \log(x_i) + v_i + \varepsilon_{ij}$$

where

$$v_i \overset{iid}{\sim} N(0, \tau^2) \quad \text{and } \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

Evaluation Data

- California 2002 & 2007 Census

  ASPEP

  government units that overlap between the 2002 and 2007 Census of Governments reporting strictly positive numbers of full-time employees.

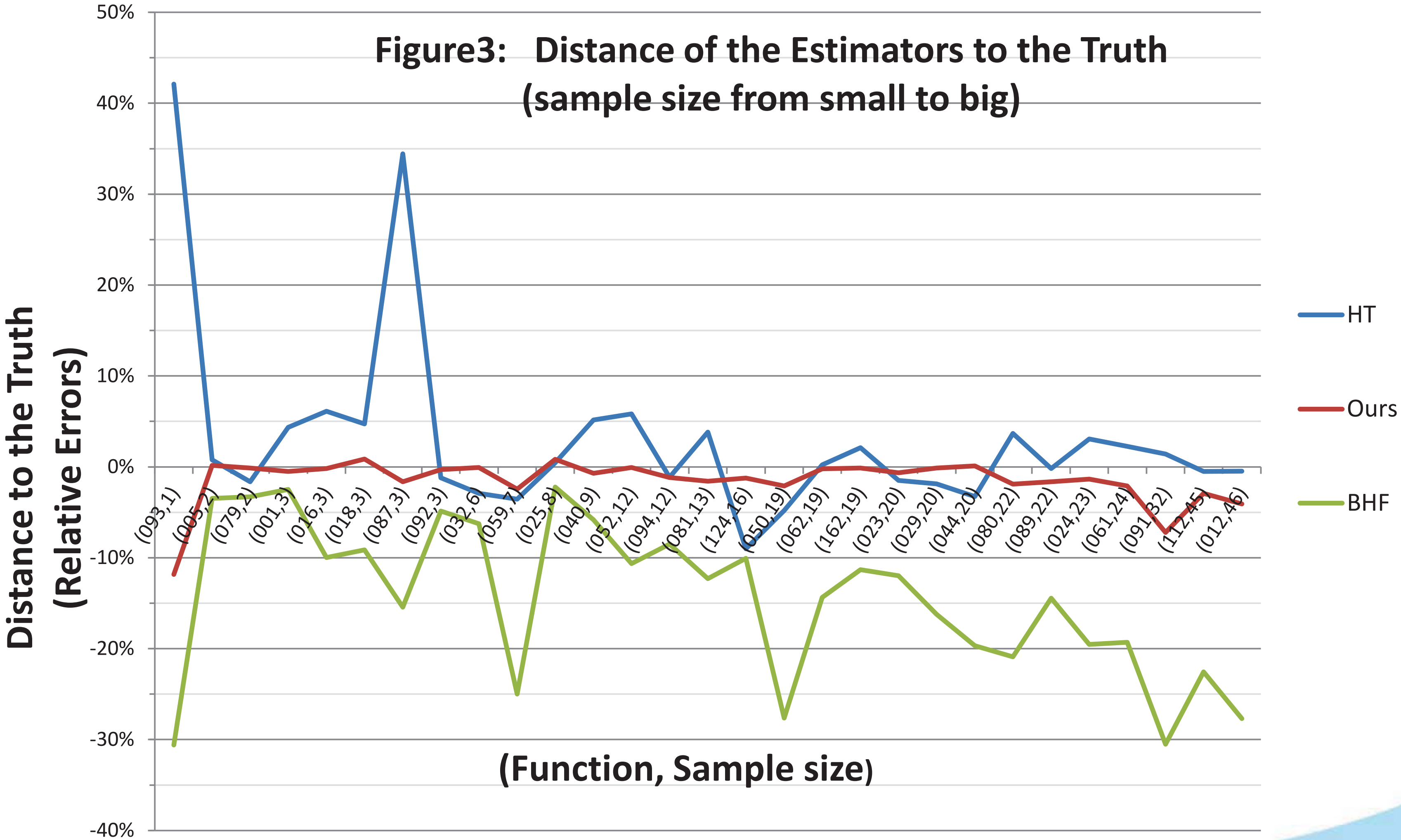
Evaluation- Results

Table 1: Percent Relative Error for Differences Estimates of Full Time Employees to the Truth (California)

Function	HT	Proposed	BHF	n_pps	n_pps/n
Gas Supply	42.1%	(11.8%)	(30.6%)	1	50.0%
Correction	0.77%	0.17%	(3.46%)	2	5.41%
Welfare	(1.65%)	(0.14%)	(3.30%)	2	3.45%
Water Transport & Terminals	34.4%	(1.64%)	(15.4%)	3	27.3%
Higher Education - Other	6.12%	(0.19%)	(9.97%)	3	5.66%
Higher Education - Instructional	4.72%	0.86%	(9.14%)	3	5.66%
Electric Power	(1.22%)	(0.30%)	(4.87%)	3	15.8%
Airports	4.35%	(0.49%)	(2.49%)	3	6.67%
Health	(2.93%)	(0.08%)	(6.26%)	6	9.09%
Natural Resources	(3.56%)	(2.46%)	(25.0%)	7	14.0%
Judical & Legal	0.44%	0.82%	(2.21%)	8	7.77%
Hospitals	5.17%	(0.71%)	(5.81%)	9	23.1%
Transit	(1.15%)	(1.18%)	(8.49%)	12	21.8%
Local Libraries	5.82%	(0.06%)	(10.6%)	12	13.3%
Solid Waste Management	3.81%	(1.58%)	(12.3%)	13	13.1%
Fire - Other	(9.02%)	(1.23%)	(10.1%)	16	17.0%
Housing & Community Development (Local)	(4.80%)	(2.11%)	(27.6%)	19	14.5%
Police-Other	2.10%	(0.12%)	(11.3%)	19	13.8%
Police Protection - Officers	0.21%	(0.21%)	(14.4%)	19	14.4%
Streets & Highways	(3.27%)	0.11%	(19.7%)	20	13.3%
Other Government Administration	(1.87%)	(0.12%)	(16.2%)	20	13.2%
Financial Administration	(1.50%)	(0.65%)	(12.0%)	20	13.1%
Sewerage	3.68%	(1.91%)	(20.9%)	22	20.6%
Other & Unallocable	(0.20%)	(1.65%)	(14.5%)	22	15.4%
Firefighters	3.08%	(1.36%)	(19.5%)	23	22.1%
Parks & Recreation	2.26%	(2.11%)	(19.3%)	24	16.2%
Water Supply	1.42%	(7.20%)	(30.5%)	32	28.3%
Elementary and Secondary - Other Total	(0.51%)	(2.92%)	(22.6%)	45	19.3%
Elementary and Secondary - Instruction	(0.48%)	(4.08%)	(27.7%)	46	19.7%

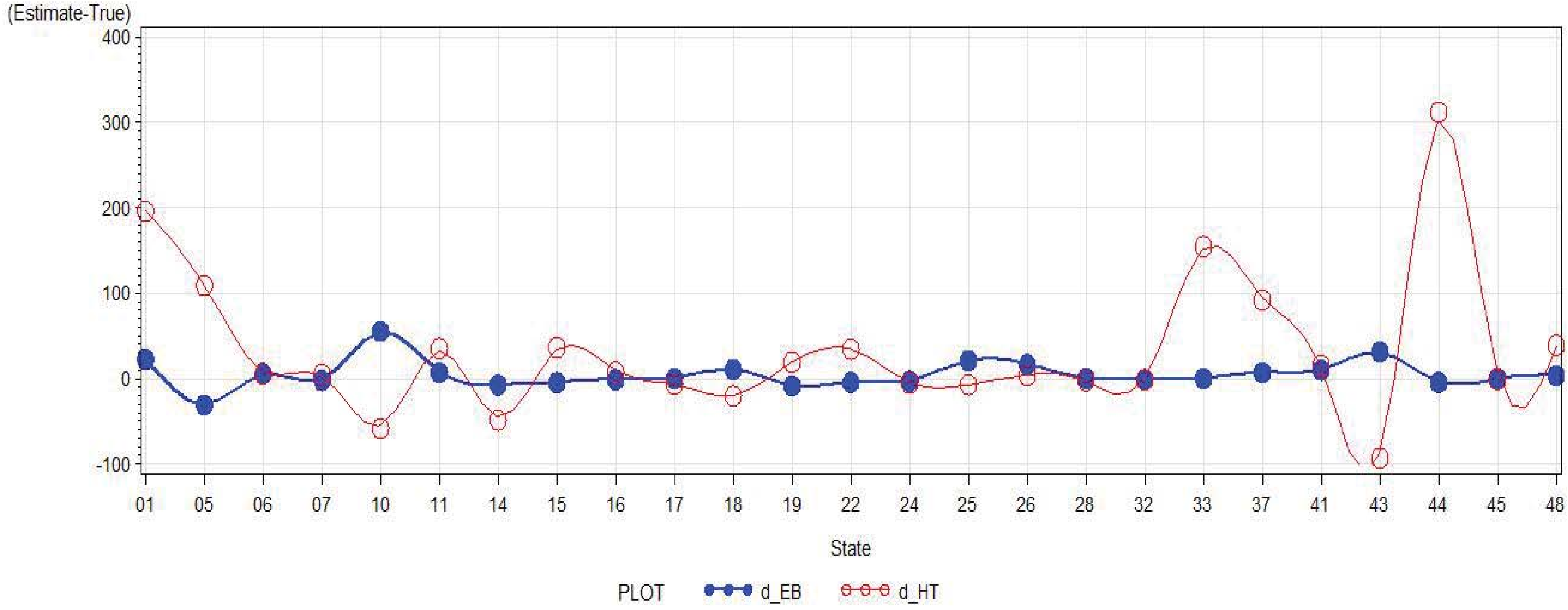
United States Census Bureau
Economic Statistics

Evaluation (Cont'd)
Visualization of Table 1



Evaluation- Results
(For Gas Supply, All States, Average n= 4)

Figure 4: **Distances of EB, HT to the Truth**

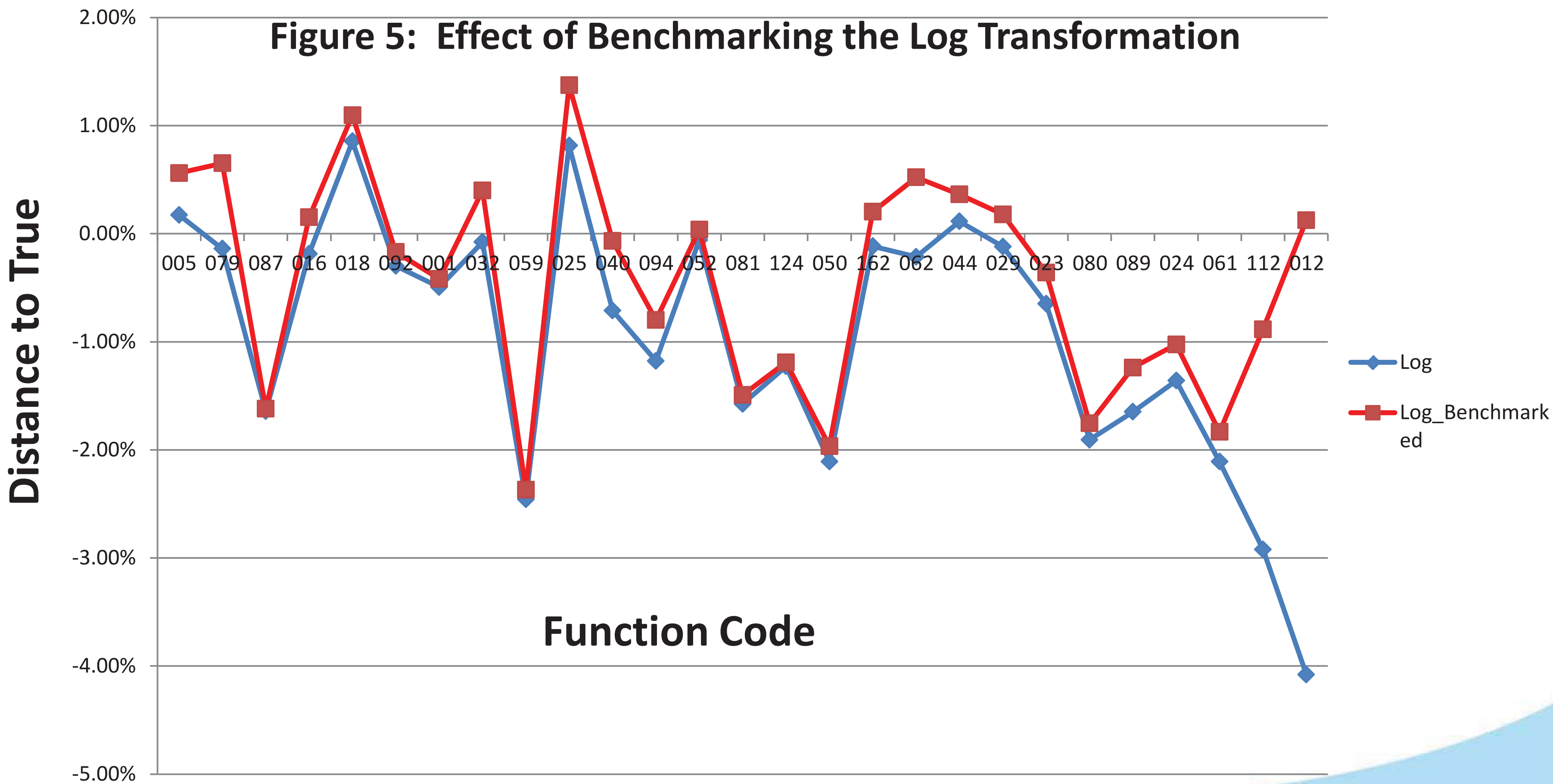


Evaluation (Cont'd)
Overall- Relative Errors

Table 2: Comparison of Overall Relative Errors (CA)

Overall - Absolute Relative Errors			
Σ|(HT-True)/True|	Σ|(EB-True)/True|	Σ|(EB_benchmarked -True)/True|	Σ|(BHF-True)/True|
5.26%	1.67%	1.44%	14.35%
Overall - Relative Errors			
Σ(HT-True)/True	Σ(EB-True)/True	Σ(EB_benchmarked-True)/True	Σ(BHF-True)/True
3.05%	-1.5%	-1%	-14.35%

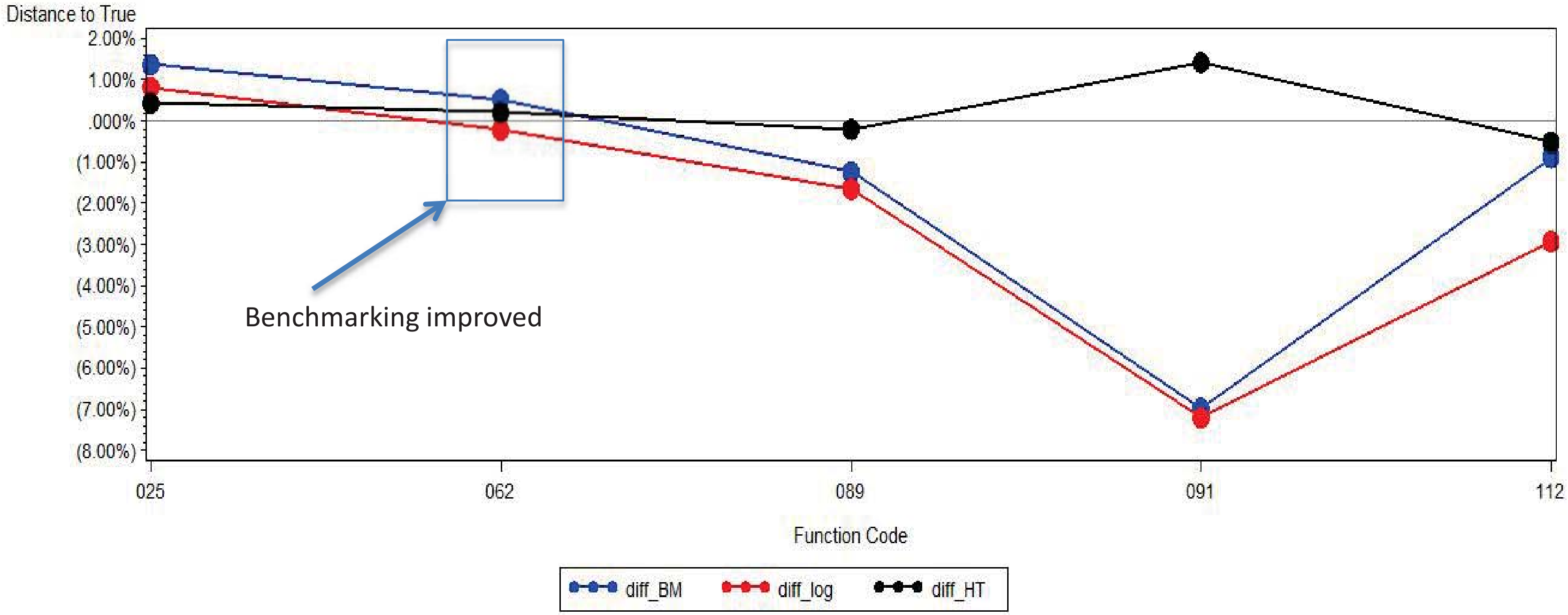
Evaluation (Cont'd)
Raking Log-transformed to HT Base (CA)



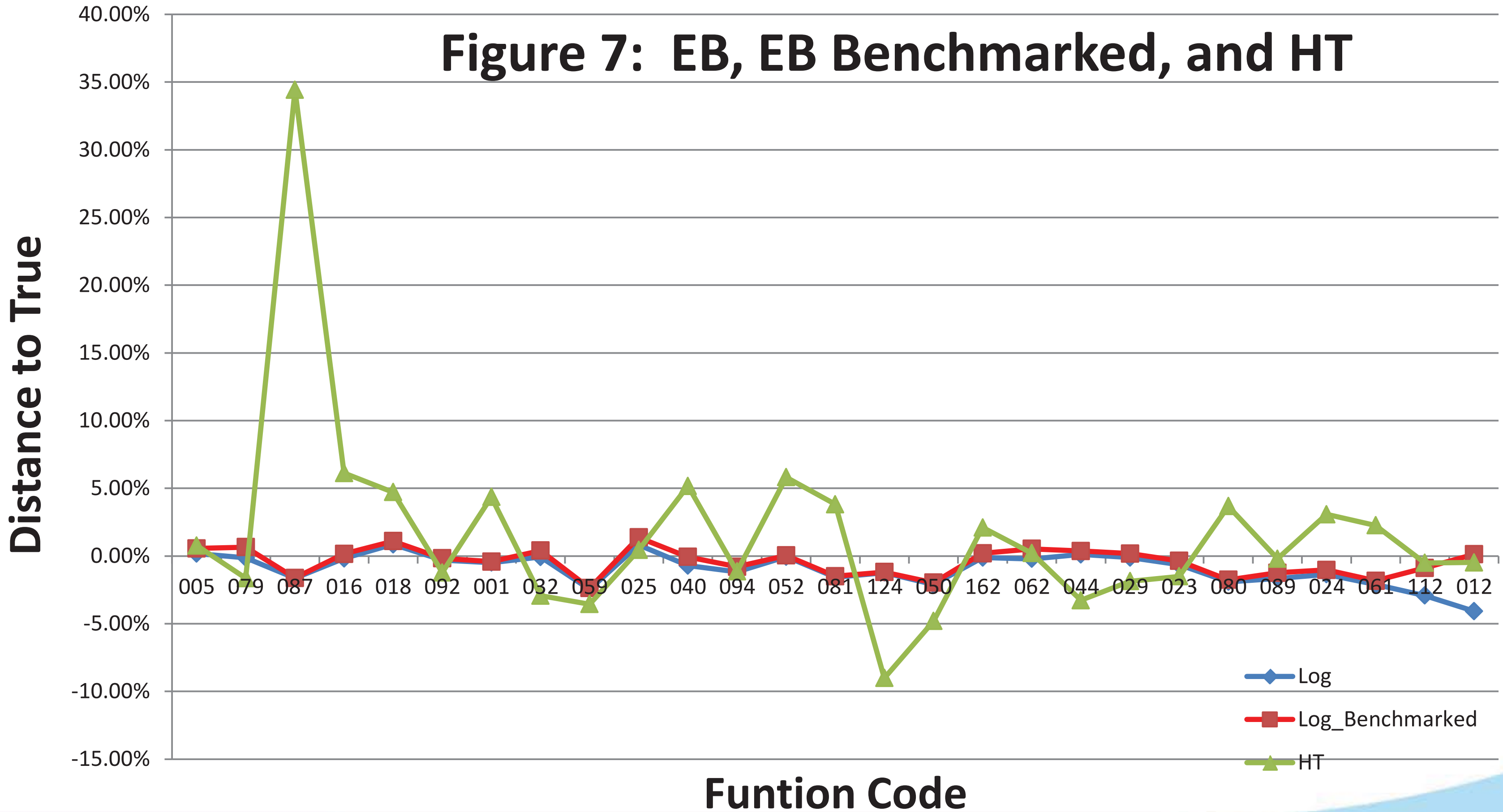
United States Census Bureau
Economic Statistics

Evaluation (Cont'd)

Figure 6: The Effect of Benchmarking the Log Transform Where the HT is Better



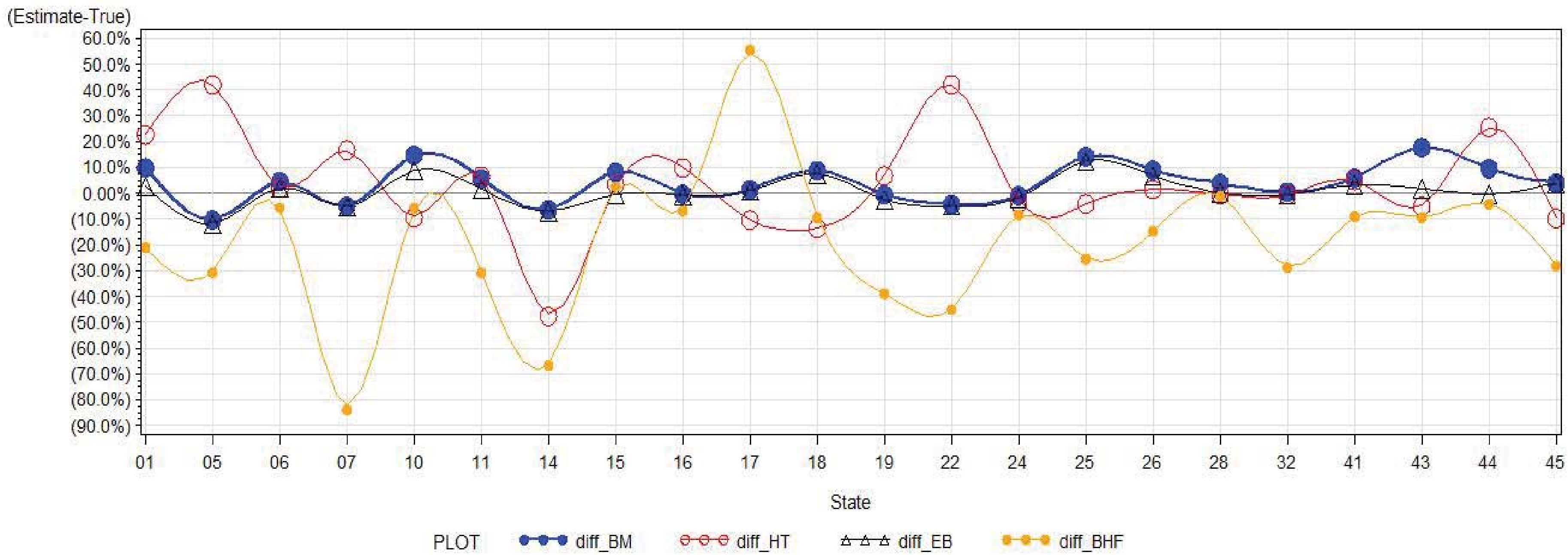
Evaluation (Cont'd)
Comparison: EB, EB Benchmarked and HT



Evaluation (Cont'd)
Domain Analysis (Gas Supply, AVG n=4)

EB= log(full-time employees), Benchmarked-EB= EB benchmarked to HT (one-way raking to nation total)

Figure 8: EB, Benchmarked-EB, HT, and BHF



Evaluation (Cont'd)
Results

- 24 out of 29 function codes (CA), our estimator outperforms the BHF, especially in small area (n <= 8)
- Benchmark Ratio (BR)
  - o BR= $|\sum(\text{estimate-HT})/\text{HT}|$
  - o Indicating how close the estimate is to the HT when considering large areas

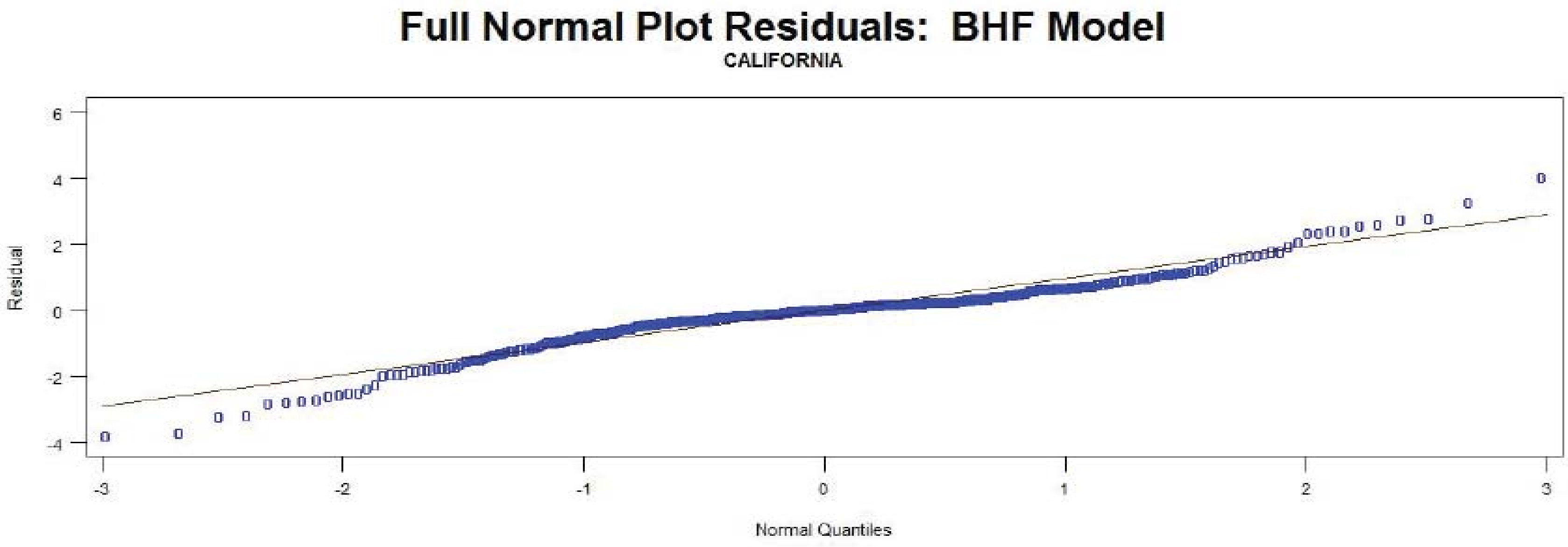
Evaluation (Cont'd)
Results

Comparison of Benchmark Ratios (Nation)

Size	BR for the EB	BR for the BHF	Number of units
< 50	1.5	1.6	1086
≥ 50	1.1	1.5	212

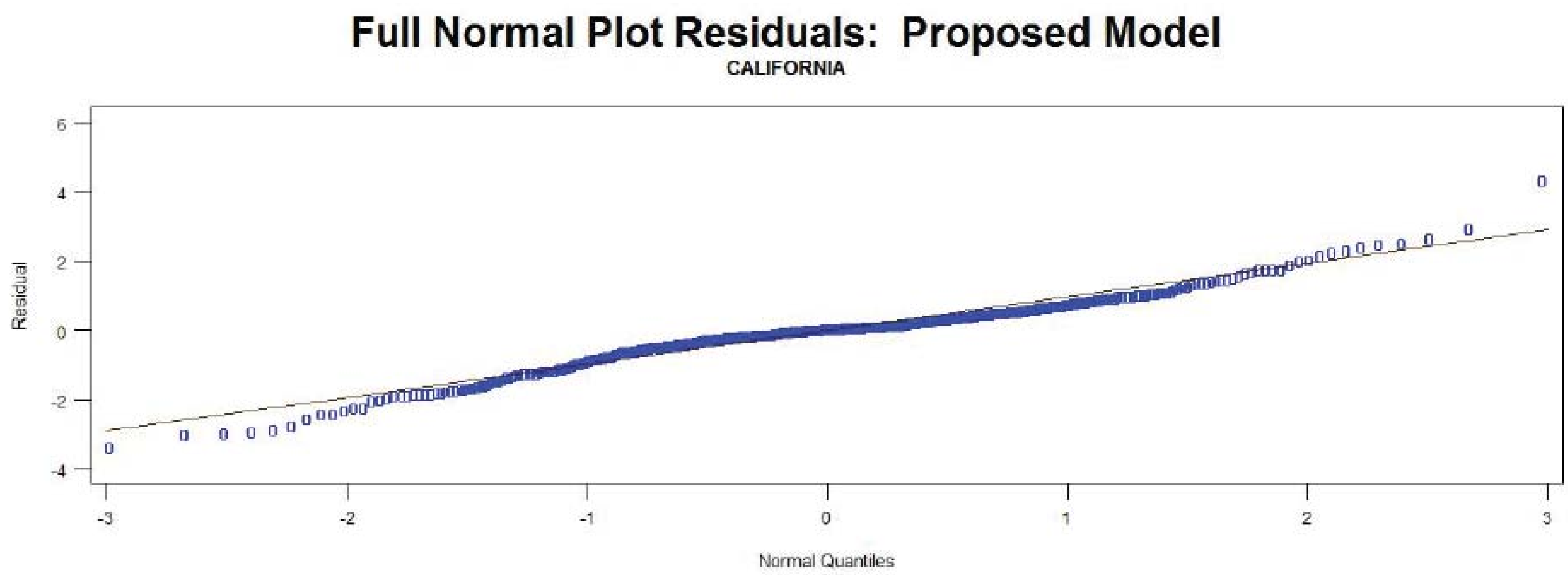
Evaluation (Cont'd)
Results- Diagnostic Analysis

Figure 9: QQ Plot for BHF Model



Evaluation (Cont'd)
Results- Diagnostic Analysis

Figure 10: QQ Plot for Our Model



Robust Small Area Estimation Using a Mixture Model

Julie Gershunskaya
U.S. Bureau of Labor Statistics

Partha Lahiri
JPSM, University of Maryland, College Park, USA

ISI Meeting, Dublin, August 23, 2011

Parameter of Interest: Small Area Means

y_{ij} : **value of a characteristic of interest for the jth unit in area i** $(\mathbf{i = 1,\ldots,m; j = 1,\ldots,N_i})$

Parameter of interest:

$$\overline{Y}_i = N_i^{-1} \sum_{j=1}^{N_i} y_{ij} = f_i \overline{y}_i + (1 - f_i) \overline{Y}_{ir},$$

$\overline{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$**;** $f_i = n_i / N_i$**;** N_i **and** n_i **are the population size and sample size for area i**

Estimator of Small Area Means

$$\hat{\bar{Y}}_i = f_i \bar{y}_i + (1 - f_i)\hat{\bar{Y}}_{ir}$$

- $\hat{\bar{Y}}_{ir}$ **is a model-dependent predictor of the mean of the non-sampled part of area *i*** $(i = 1, \cdots, m)$.
- **If** $f_i \approx 0$**,** $\hat{\bar{Y}}_i \approx \hat{\bar{Y}}_{ir}$
- **Let** $n = \sum_{i=1}^{m} n_i$ **and** $N = \sum_{i=1}^{m} N_i$.

The Nested Error Regression Model (Battese, Harter, Fuller, 1988)

For $\mathbf{i = 1,\ldots,m; j = 1,\ldots,N_i}$,

$$\mathbf{y_{ij} = x_{ij}^T\beta + v_i + \varepsilon_{ij}}$$

- $\mathbf{x}_{ij}$ **is a vector of known auxiliary**
- $\boldsymbol{\beta}$ **is the corresponding vector of parameters;**
- v_i **are random effects**
- ε_{ij} **are errors in individual observations**
- $v_i \overset{iid}{\sim} N(0,\tau^2)$ **and** $\varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$,
- **We assume that sampling is non-informative**

EBLUP

BLUP of $\bar{Y}_{ir}$:

$$\hat{\bar{Y}}_{ir} = \bar{\mathbf{x}}_{ir}^T \hat{\boldsymbol{\beta}} + \hat{v}_i \ ,$$

- $\bar{\mathbf{x}}_{ir}^T = (N_i - n_i)^{-1} \sum_{j=n_i+1}^{N_i} \mathbf{x}_{ij}^T ,$
- $\hat{\boldsymbol{\beta}}$ **is the BLUE of $\boldsymbol{\beta}$,**
- $\hat{v}_i = \tau^2(\sigma^2/n_i + \tau^2)^{-1}(\bar{y}_i - \bar{\mathbf{x}}_i^T \hat{\boldsymbol{\beta}})$**.is the BLUP of** v_i

- **EBLUP of $\bar{Y}_{ir}$ after plugging in estimates of $\sigma^2$and τ^2.**

A Robust Unit-Level Model: An Extension of the BHF Model

For $\mathbf{j = 1, \ldots, N_i; i = 1, \ldots, m}$,

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + v_i + \varepsilon_{ij},$$

- $v_i \overset{iid}{\sim} N(0, \tau^2)$, $\varepsilon_{ij} \mid z_{ij} \overset{iid}{\sim} (1 - z_{ij}) N(0, \sigma_1^2) + z_{ij} N(0, \sigma_2^2)$,
- $z_{ij} \mid \pi \overset{iid}{\sim} Bin(1; \pi)$,
- π**: probability of belonging to mixture part 2.**
- $\sigma_1^2 \leq \sigma_2^2$

Empirical Best Predictor (EBP) of $\bar{Y}_i$

$$\hat{\bar{Y}}_{ir} = \bar{\mathbf{x}}_{ir}^T \hat{\boldsymbol{\beta}} + \hat{v}_i$$

- $$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij}^T (y_{ij} - \hat{v}_i) \Big/ \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij}^T \mathbf{x}_{ij},$$
- $$w_{ij} = \hat{\sigma}_1^{-2}(1 - \hat{z}_{ij}) + \hat{\sigma}_2^{-2} \hat{z}_{ij}, \quad \hat{z}_{ij} = E\left[z \mid y_{ij}, \mathbf{x}_{ij}, \hat{\boldsymbol{\theta}} \right]$$
- $$\hat{v}_i = \frac{\hat{\tau}^2}{D_i + \hat{\tau}^2} (\hat{\bar{y}}_i - \hat{\bar{\mathbf{x}}}_i^T \hat{\boldsymbol{\beta}}), \quad D_i = \left(\sum_{j=1}^{n_i} w_{ij} \right)^{-1}$$
- $$\hat{\bar{y}}_i = \left(\sum_{j=1}^{n_i} w_{ij} \right)^{-1} \sum_{j=1}^{n_i} w_{ij} y_{ij}, \quad \hat{\bar{\mathbf{x}}}_i = \left(\sum_{j=1}^{n_i} w_{ij} \right)^{-1} \sum_{j=1}^{n_i} w_{ij} \mathbf{x}_{ij}^T$$

,

Overall Bias-corrected REB

$$\hat{\bar{Y}}_{ir}^{REB+OBC} = \hat{\bar{Y}}_{ir}^{REB} + n^{-1} s^{REB} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \phi_b \left(\frac{e_{ij}^{REB}}{s^{REB}} \right),$$

s^{REB}**: a robust measure of scale for the set of residuals** $\left\{ e_{ij}^{REB}; j = 1, \ldots, n_i, i = 1, \ldots, m \right\}$**,**

e.g., $s^{REB} = med \left| e_{ij}^{REB} - med(e_{ij}^{REB}) \right| / 0.6745$

ϕ_b**: a bounded Huber's function with the tuning parameter *b* = *5*.**

Estimation of Crop Indication

- **USDA-NASS has been publishing county level crop and livestock estimates since 1917**
- **County indications of crops such as harvested yield are needed to assist farmers, agribusinesses and government agencies in local agricultural decision making.**
- **Most NASS Field Offices conduct a separate County Estimates Survey every year. Data from multiple sample surveys are used to estimate harvested yield for various crops at the county level.**

Estimators Compared

For seven mid-western states in the year 2007, we compared the following estimates, treating the 2007 agriculture census as the gold standard.

- **EBLUP under the BHF model**
- **EBP under NER Mixture Model [N2]**
- **Kott-Busselberg Model-Based Direct [KB]**
- **USDA-NASS official estimates**

Criteria for Evaluation

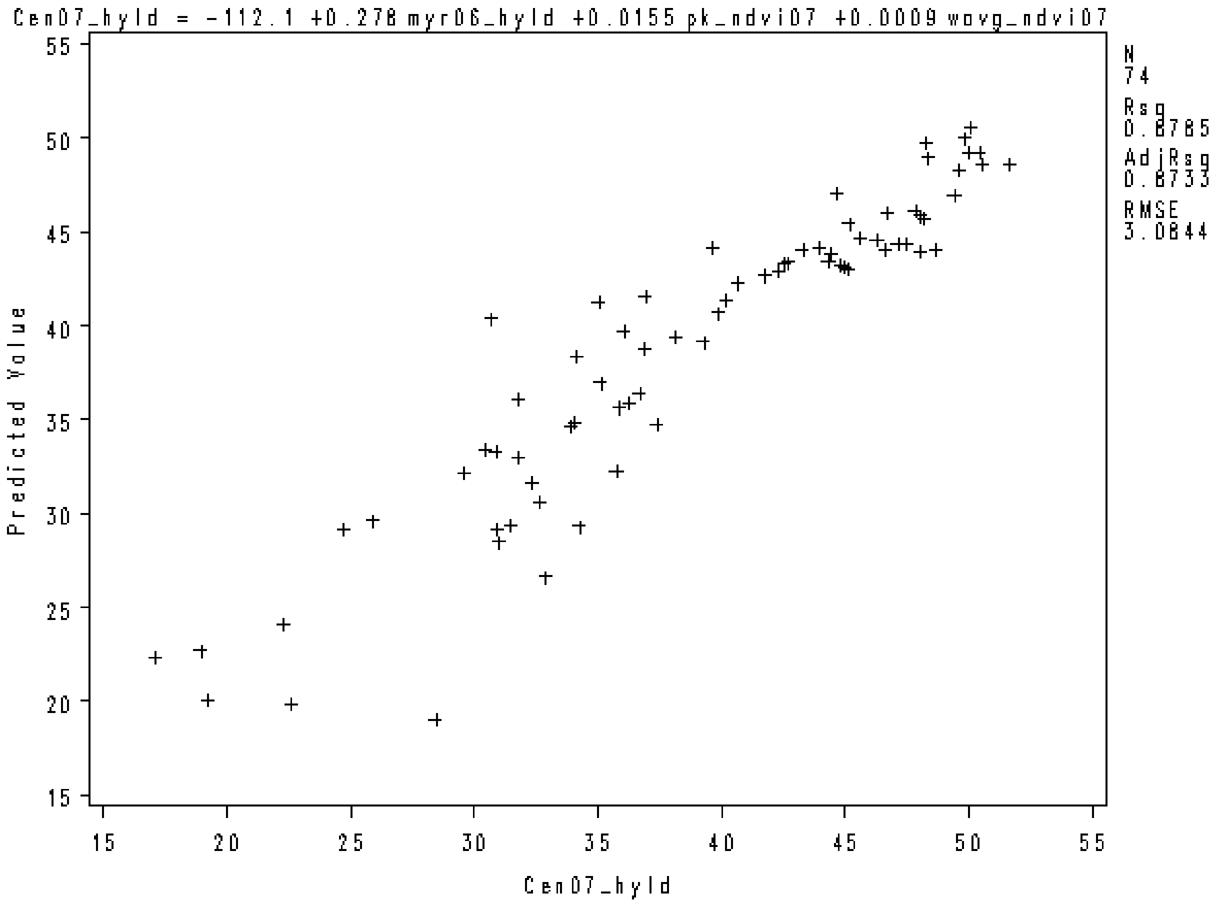
- **AAD: the mean of absolute deviations between county estimates and corresponding 2007 census (PC) values**
- **ASD: the mean of squared deviations between estimates and PC values**
- **AARD: the mean of ratios between absolute deviations and PC values**
- **ASRD: the mean of squared ratios between absolute deviations and PC values**
- **PBC: the proportion of counties with estimate less than the corresponding PC value.**

Results

- **The BHF and N2 estimates are clearly superior to the direct estimates for all the states considered.**

- **EBPs are also better than the official estimates in all but one state (Minnesota.)**

- **The OBC* correction to N2 provides similar results for most of the seven states. However, it provides slightly better results for Iowa, but slightly worse results for Minnesota.**



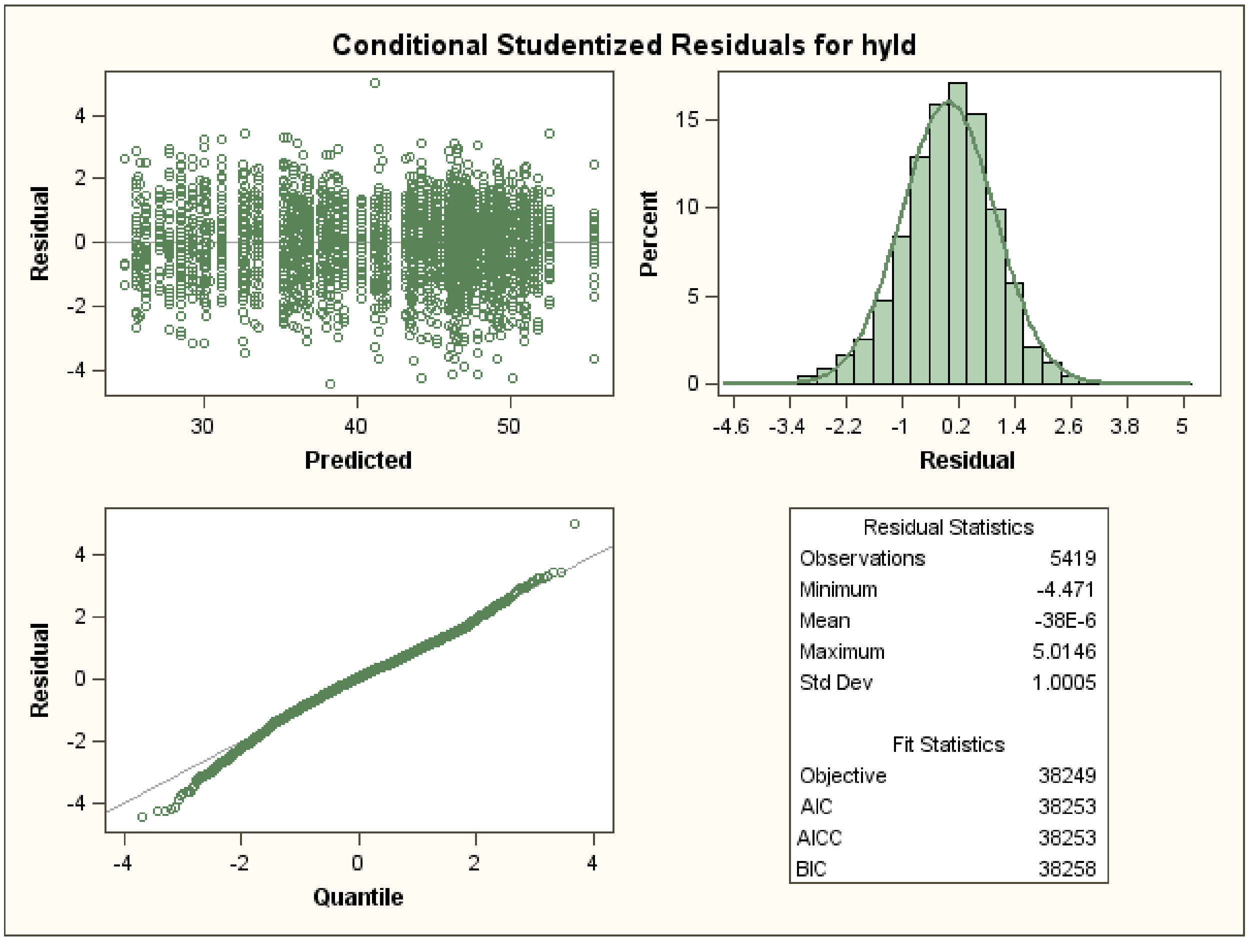
Level 2 Regression for Harvested Yield: Minnesota

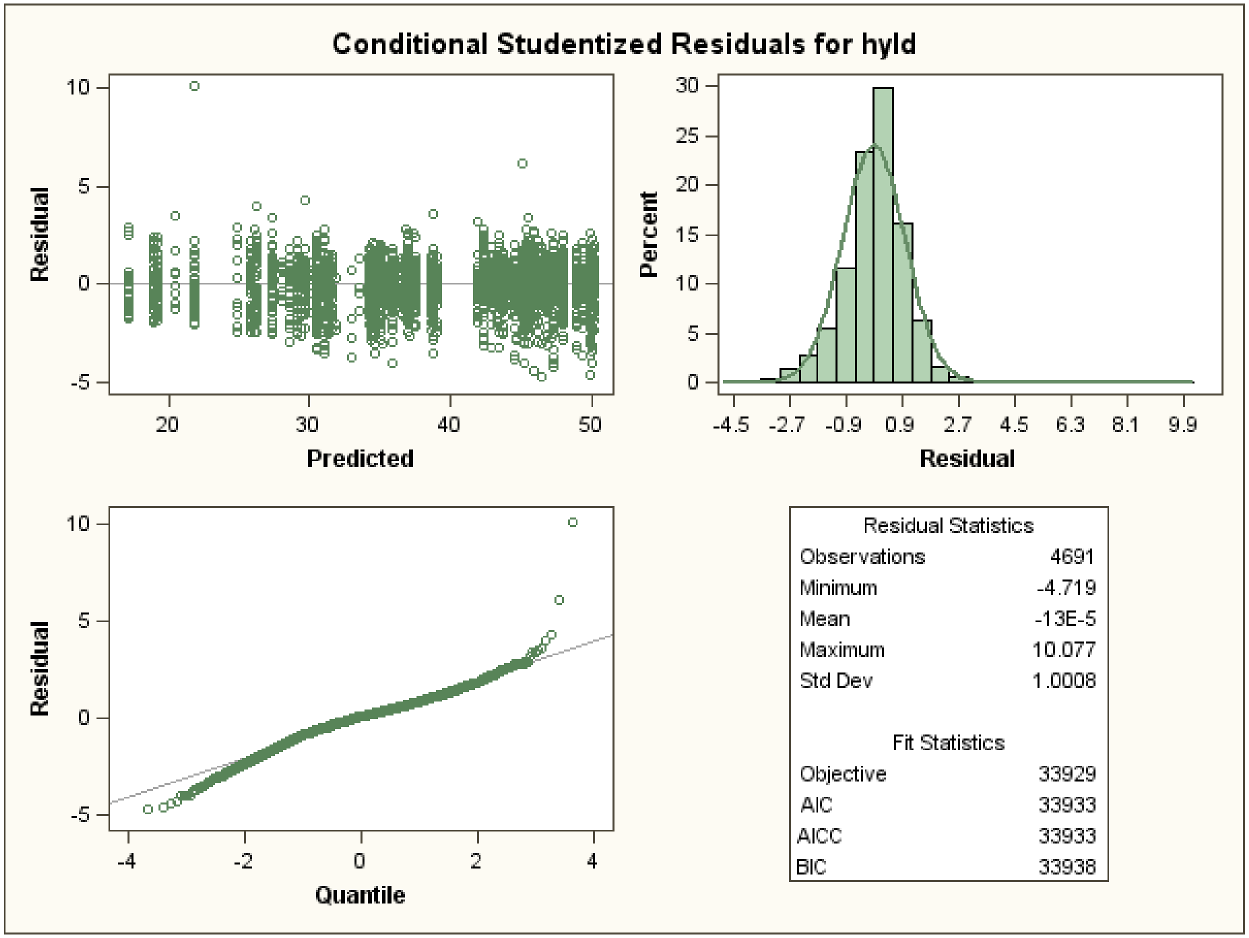
<u>Table:</u> Estimation Accuracy Measures for Harvested Yield*

State	Estimator	Metric				
		AAD	ASD	AARD	ASRD	PBC
Illinois	EBLUP	<u>1.34</u>	<u>2.85</u>	0.036	0.002	0.32
	KB	2.7	12.6	0.07	0.009	0.85
	N2	<u>1.33</u>	<u>2.8</u>	0.036	0.002	0.33
	N2+OBC*	<u>1.33</u>	2.8	0.036	0.002	0.32
	Official	1.82	5.18	0.048	0.004	0.42
Iowa	EBLUP	1.10	1.81	0.022	0.001	0.69
	KB	2.7	13.5	0.055	0.006	0.82
	N2	1.24	2.15	0.025	0.001	0.83
	N2+OBC*	0.95	1.48	0.019	0.001	0.72
	Official	2.12	5.94	0.043	0.002	0.08
Minnesota	EBLUP	1.32	<u>3.92</u>	0.037	0.004	0.31
	KB	3.46	26.0	0.095	0.022	0.85
	N2	<u>1.23</u>	<u>4.04</u>	0.036	0.004	0.36
	N2+OBC*	1.38	<u>4.58</u>	0.040	0.005	0.28
	Official	1.32	2.67	0.034	0.002	0.19

Residual Plots for BHF model for Soybeans yield: Minnesota

Residual Plots for BHF model for Soybeans yield: Indiana




Conditional Studentized Residuals for hyld
Residual
Predicted
Percent
Residual
Residual
Quantile
Residual Statistics
Observations 4691
Minimum -4.719
Mean -13E-5
Maximum 10.077
Std Dev 1.0008
Fit Statistics
Objective 33929
AIC 33933
AICC 33933
BIC 33938

Future Research:

- Develop refined area level covariates using NDVI
- Incorporate non-response model
- Use robust methods to estimate for harvested acreage to deal with outliers in the size variable
- Develop a unified benchmarked method that produces estimates of all crop indications

References:

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data, Journal of the American Statistical Association, 83, 28-36

Bellow, M.E. (2007), Comparison of Methods for Estimating Crop Yield at the County Level, United States Department of Agriculture, National Agricultural Statistics Service, RDD Research Report.

Gershunskaya, J. (2010), Robust Small Area Estimation Using a Mixture Model, Proc. SRMS

Gershunskaya, J. and Lahiri, P., (2008). Robust Estimation of Monthly Employment Growth Rates for Small Areas in the Current Employment Statistics Survey. Proceedings of the Section on Survey Research Methods, American Statistical Association.

Iwig, W.C. (1993), The National Agricultural Statistics Service County Estimates Program, National Agricultural Statistics Service.

Kott, P.S. (2008), Some ideas for a New Set of County-Estimates Crop Indications: an Update.

Prusacki, J. (2008), County Estimates/Small Area Estimation.

Rao, J.N.K. (2003). ***Small Area Estimation*****, New-York, John Wiley & Sons, Inc.**

Stasny, E. A., Goel, P.K., & Rumsey, D.J. (1991). County Estimates of Wheat Production. ***Survey Methodology*****,** ***17*****, 211-225.**

David Salsburg, ASA Connect Discussion

"...D.J. Finney once wrote about the statistician whose client comes in and says, "Here is my mountain of trash. Find the gems that lie therein." Finney's advice was to not throw him out of the office but to attempt to find out what he considers "gems". After all, if the trained statistician does not help, he will find some one who will...."

Workshop on Statistical Data Integration
Singapore, 5th - 8th August, 2019.

- **Topics**:
  - Record Linkage
  - Statistical Matching
  - Small Area Estimation
  - Statistical Disclosure Avoidance
  - Synthetic Population
  - Big Data
  - Combining Multiple Surveys
- **Organisers**: Sanjay Chaudhuri (chair), Partha Lahiri, Pedro Luis do Nascimento Silva, Danny Pfeffermann.
- **Sponsor**: Institute for Mathematical Sciences, National University of Singapore.
- More Information:

https://ims.nus.edu.sg/events/2019/data/index.php.

Conference on Current Trends in Survey Statistics
Singapore, August 13-16, 2019.

- **Topics**:
  - Statistical Data Integration with Complex Survey Data
  - Statistical Methods for Non-sampling Errors
  - Mixed Mode Mixed Frame Surveys
  - Resampling Methods with Survey Data
  - Informative Sampling
  - Empirical Likelihood for Survey Data
  - Bayesian Methods for Survey Data
  - Non-probability Sampling
  - Others
- The conference is partly sponsored by the Institute for Mathematical Sciences, National University of Singapore and endorsed by the International Association of Survey Statisticians (IASS).

- **Scientific Advisory Board**:
  - Raymond Chambers, University of Wollongong
  - Malay Ghosh, University of Florida
  - Graham Kalton, Westat
  - Partha Lahiri, (Chair), University of Maryland
  - Danny Pfeffermann, National Statistician of Israel and University of Southampton,
  - J. N. K. Rao, Carleton University,
  - Pedro Luis do Nascimento Silva, IBGE, Brazil.
- **Tentative Scientific Programme Committee:**
  - Sanjay Chaudhuri (Chair), NUS
  - William Bell, US Census Bureau
  - Yang Cheng, US Census Bureau
  - Cinzia Cirillo, University of Maryland, College Park
  - Jiming Jiang, University of California, Davis
  - Ralph Munnich, University of Trier, Germany
  - Santanu Pramanik, Delhi Centre of National Data Innovation, India,
  - Jan Van Barkel, University of Maastricht
  - Rebecca Steort, Duke University
  - Dongchu Sun, University of Missouri
  - Jiraphan Suntornchost, Chulalongkorn University,Thailand
  - Nikos Tzavidis, University of Southampton
  - Li-Chun Zhang, University of Southampton, Oslo, Stat Norway

THANK YOU!