Small Area Estimation Methods, Applications and Practical Demonstration

Part 1: Traditional Methods

J.N.K. Rao
School of Mathematics and Statistics, Carleton University

Isabel Molina
Department of Statistics, Universidad Carlos III de Madrid

Introduction

- U large finite population of size N.
- $y_1, \ldots, y_N$ variable of interest for the population units.
- Target quantity: population total

$$Y = \sum_{k=1}^{N} y_k.$$

- $s \subset U$ sample drawn from the population.
- $r = U - s$ non-sample units.

Basic direct estimator

- π_k probability of inclusion of unit k in the sample.
- $d_k = 1/\pi_k$ sampling weight for unit k.
- Basic direct estimator of Y:

$$\hat{Y} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} d_k y_k$$

- It is design-unbiased:

$$E(\hat{Y}) = Y$$

Variance of direct estimator

- Design-unbiased estimator of the variance:

$$v(\hat{Y}) = \sum_{k \in s} \sum_{\ell \in s} \frac{\pi_{k,\ell} - \pi_k \pi_\ell}{\pi_{k,\ell}} \frac{y_k}{\pi_k} \frac{y_\ell}{\pi_\ell},$$

  $\pi_{k,\ell}$ joint inclusion probability for units k and ℓ.

  ✓ *Särndal, Swensson and Wretman (1992), equation (5.8.5)*

- Under the approximation $\pi_{k,\ell} \cong \pi_k \pi_\ell$, $k \neq \ell$,

$$v(\hat{Y}) \cong \sum_{k \in s} \left(\frac{1 - \pi_k}{\pi_k^2} \right) y_k^2 = \sum_{k \in s} d_k (d_k - 1) y_k^2.$$

FGT poverty indicators

- E_k welfare measure for individual k: for instance, normalized annual net income.
- z poverty line: Spanish Statistical Institute (INE) uses:

$$z = 0.6 \times \text{Median}(E_k).$$

- FGT Family of poverty indicators

$$F_\alpha = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{z - E_k}{z} \right)^\alpha I(E_k < z), \quad \alpha = 0, 1, 2.$$

- $\alpha = 0 \Rightarrow$ Poverty incidence
- $\alpha = 1 \Rightarrow$ Poverty gap
- $\alpha = 2 \Rightarrow$ Poverty severity

✓ *Foster, Greer & Thornbecke (1984)*

Direct estimators of FGT poverty indicators

- Poverty indicator:

$$F_\alpha = \frac{1}{N}\sum_{k=1}^{N} F_{\alpha k}, \quad F_{\alpha k} = \left(\frac{z - E_k}{z}\right)^\alpha I(E_k < z), \quad \alpha = 0, 1, 2.$$

- Basic direct estimator:

$$\hat{F}_\alpha = \frac{1}{N}\sum_{k\in s} d_k F_{\alpha k}, \quad \alpha = 0, 1, 2.$$

- Estimator of the variance:

$$v(\hat{F}_\alpha) = \frac{1}{N^2}\sum_{k\in s} d_k(d_k - 1)F_{\alpha k}^2, \quad \alpha = 0, 1, 2.$$

Adjustments to direct estimators

- g_k adjustment factor for design weights, $k \in s$.
- $w_k = d_k g_k$ final weight, $k \in s$.
- New direct estimator:

$$\hat{Y}^{DIR} = \sum_{k \in s} w_k y_k$$

Post-stratified direct estimator

- J large population subgroups (post-strata) with homogeneous units: for instance, age-sex groups.
- N_{+j} (projected) census count in post-stratum j.
- s_{+j} sample in post-stratum j.
- $\hat{N}_{+j} = \sum_{k \in s_{+j}} d_k$ basic direct estimator of N_{+j}.
- Adjustment factor: $g_k = N_{+j}/\hat{N}_{+j},\ k \in s_{+j}$.

Ratio estimator

- X known total of auxiliary variable x.
- $\hat{X} = \sum_{k \in s} d_k x_k$ basic direct estimator of X.
- Adjustment factor: $g_k = X/\hat{X}$ for $k \in s$.
- Ratio direct estimator:

$$\hat{Y}^{DIR} = (\hat{Y}/\hat{X})\, X$$

Calibration estimator

- p auxiliary variables with known population totals $X_l,\ l = 1, \ldots, p$.
- **Idea:** Find weights $w_k,\ k \in s$, which minimize the distance

$$\begin{aligned} \min \quad & \sum_{k \in s} \left(w_k - d_k\right)^2 / d_k \\ s.t. \quad & \sum_{k \in s} w_k x_{kl} = X_l,\ l = 1, \ldots, p. \end{aligned}$$

- **Solution:** $w_k = d_k g_k$, where $g_k = 1 + \mathbf{x}_k^T \hat{\mathbf{T}}^{-1}(\mathbf{X} - \hat{\mathbf{X}})$,

$$\mathbf{x}_k = (x_{k1}, \ldots, x_{kp})^T, \quad \mathbf{X} = (X_1, \ldots, X_p)^T, \quad \hat{\mathbf{T}} = \sum_{k \in s} d_k \mathbf{x}_k \mathbf{x}_k^T$$

✓ *Deville and Särndal (1992)*

Calibration estimator

Special case: Post-stratified direct estimator

- Post-stratified direct estimator is a calibration estimator when the auxiliary variables are post-stratum indicators:

$$x_{kj} = \begin{cases} 1, & \text{if unit } k \text{ belongs to post-stratum } j, \text{ for } j = 1, \ldots, J \\ 0, & \text{otherwise} \end{cases}$$

- Note: $\sum_{j=1}^{J} x_{kj} = 1$ for each unit k.
- Extension: two or more post-stratification variables with known marginal population counts.

Introduction to small domain estimation

- U partitioned into m domains U_i, $i = 1, \ldots, m$.
- Terminology: Domain, Area, Sub-population.
- N_i size of domain i.
- $s_i \subset U_i$ sample within domain i, $r_i = U_i - s_i$ non-sample.
- Total of domain i:

$$Y_i = \sum_{k \in U_i} y_k$$

- **Direct estimators:** Use only area-specific sample data.
- Basic direct estimator of Y_i and variance estimator:

$$\hat{Y}_i = \sum_{k \in s_i} d_k y_k, \qquad v(\hat{Y}_i) = \sum_{k \in s_i} d_k (d_k - 1) y_k^2.$$

Introduction to small domain estimation

Examples of Domains

- *Geographic*: county, school district, even a state, municipality, census tract, "tehsil": group of villages in India.

  US survey: $n = 10,000$ persons (self-weighting sample)

  Expected state sample sizes:

California state	1207	Washington DC	22
New York state	698	Wyoming	18

- *Socio-economic*: age $\times$ sex $\times$ race $\times$ poverty status

Indirect estimation

- Sample is not planned to give accurate direct estimators for the domains: there are domains with few sample observations.
- **Small domain**: domain for which $\text{cv}(\hat{Y}_i) > 20\%$.
- **Indirect or small area estimators:** Borrow strength from sample observations of related areas to increase "effective" sample size.

Properties of direct estimators

- Additivity: $\sum_{i=1}^{m} \hat{Y}_i^{DIR} = \hat{Y}^{DIR}$
- Measure of accuracy of $\hat{Y}_i^{DIR}$:

$$\mathrm{MSE}(\hat{Y}_i^{DIR}) = \mathrm{V}(\hat{Y}_i^{DIR}) + \left[\mathrm{B}(\hat{Y}_i^{DIR})\right]^2 =: (a) + (b),$$

  (a) is of the order of the i-th area sample size.
  (b) negligible for large overall sample size n.

- $\mathrm{CV} = \sqrt{\mathrm{V}(\hat{Y}_i^{DIR})}/Y_i = \mathrm{SE}(\hat{Y}_i^{DIR})/Y_i$
- CV$< 20 - 25\%$ considered adequate for small areas.

Design issues to improve efficiency

Design issues that help to improve efficiency of direct estimators for planned small domains:

- Clustering reduces efficiency: better to use list frames.
- Use many strata so that all domains are well covered.
- Optimal allocation for large domains may lead to small domains poorly represented: compromise sample allocation between large and small domains.

Design issues to improve efficiency

Example: Canadian Labor Force Survey

- Monthly sample of 59,000 households optimized at the provincial level: CV as high as 17.7% for Unemployment Insurance (UI) regions.
- *Compromise two-step sample allocation*:
  42,000 for province level, 17,000 for UI level.

  (i) For UI regions, maximum CV reduced from 17.7% to 9.4%.
  (ii) For provinces and Canada, small increase of CV:
  Ontario: from 2.8% to 3.2%
  Canada: from 1.4% to 1.5%

✓ *Singh, Gambino and Mantel (1994)*

Design issues to improve efficiency

- *Other suggestions to "reduce" the use of indirect estimators:*

  (i) Integration (or harmonization) of surveys.

  (ii) Use of multiple frame surveys.

  (iii) Rolling samples: In the American Community Survey (ACS), independent samples are drawn each month from each county.

  ✓ *Marker (2001)*; ✓ *Kish (1999)*.

- *Planned domains: sample allocation:*

  Minimize a weighted sum of variances of direct estimators subject to fixed overall sample size. Weights are called "inferential priorities". ✓ *Longford (2006)*

  Problem: Difficult to specify the weights.

Design issues to improve efficiency

- *Alternative solution:* Minimize total sample size subject to desired tolerances on the area sampling variances and on the aggregate sampling variance. ✓ *Choudhry and Rao (2009)*
- *"Client will always require more than is specified at the design stage"* ✓ *Fuller (1999)*.
- So we cannot avoid unplanned small domains.

First application of indirect estimation

Example: 1945 Radio Listening Survey

- Target: to estimate the median num. of radio stations heard during the day in 500 U.S. counties.
- Mail survey: From each of 500 counties, 1000 families sampled and sent mailed questionaire.
- Response rate only 20% and incomplete coverage.
- x_i median no. of stations heard during day (mail survey) in the i-th county, for $i = 1, \ldots, 500$.
- Intensive interview survey of 85 counties: Probability sample of 85 counties subsampled and subject to personal interviews.

✓ *Hansen, Hurwitz & Madow, 1953, p. 483;* ✓ *Rao, 2003, p. 36*

First application of indirect estimation

Example: 1945 Radio Listening Survey

- y_i median no. of stations heard during day (interview) in the i-th sample county, for $i = 1, \ldots, 85$.
- $\text{corr}(y, x) = 0.70$
- Linear Regression:

$$y_i = \alpha + \beta\, x_i + e_i, \quad i = 1, \ldots, 85.$$

- Predicted values: For non-sampled counties, predicted values $\hat{y}_i^{SYN} = 0.52 + 0.74 x_i$ used as indirect (synthetic) estimators.
- This is an example of explicit linking model

Definition of synthetic estimator

Synthetic estimator:

An unbiased estimator is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas under the assumption that the small areas have the same characteristics as the large area, we identify these estimates as synthetic estimates. ✓ *González (1973)*

Example: Simplest possible model

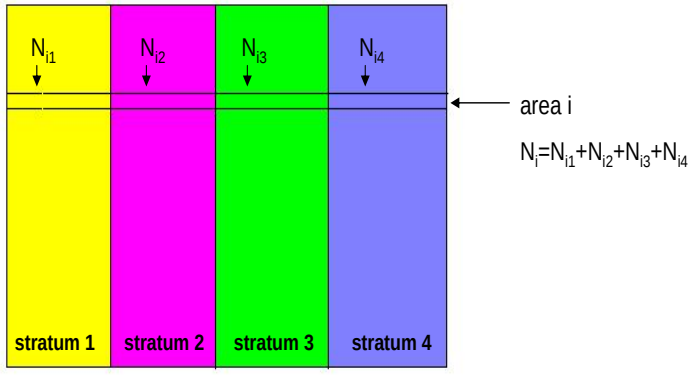
- Total: $Y_i = N_i \bar{Y}_i$, where $\bar{Y}_i = Y_i/N_i$ is area mean.
- N_i, N known.
- Implicit model: $\bar{Y}_i = \bar{Y}$, $i = 1, \ldots, m$.
- $\hat{Y}$, $\hat{N}$ reliable direct estimators of Y, N
- Synthetic estimator: $\hat{Y}_i^{SYN} = N_i(\hat{Y}/\hat{N})$

Post-stratified synthetic estimator

- J post-strata ($j = 1, \ldots, J$) which cut across the areas.

- N_{ij} count in the intersection of domain i and post-stratum j (known).
- Total of domain i:

$$Y_i = \sum_{j=1}^{J} N_{ij} \bar{Y}_{ij}.$$



- Implicit model:

$$\bar{Y}_{ij} = \bar{Y}_{+j} = Y_{+j}/N_{+j}$$

Post-stratified synthetic estimator

- Post-stratified synthetic estimator:

$$\hat{Y}_i^{SYN} = \sum_{j=1}^{J} N_{ij}(\hat{Y}_{+j}/\hat{N}_{+j})$$

- $\hat{Y}_{+j}, \hat{N}_{+j}$ reliable direct estimators of Y_{+j}, N_{+j}.
- Need homogeneity only within post-strata.
- Special case: When $y \in \{0, 1\}$, interest is in the proportion $P_i = Y_i/N_i$. Synthetic estimator of P_i:

$$\hat{P}_i^{SYN} = \frac{1}{N_i} \sum_{j=1}^{J} N_{ij} \hat{P}_{+j}, \quad N_i = \sum_{j=1}^{J} N_{ij}$$

Example: 1980 U.S. National Natality Survey (NNS)

- Target: state estimates of percent jaundiced live births.
- Data from $n = 9941$ live births collected from birth certificates, questionnaires and hospitals.
- 25 post-strata: mother's race and age group, live birth order.
- $\hat{P}_{+j}$ direct national est. for post-stratum j.
- N_{ij} no. of hospital births (from State Vital Registration data)

$$\text{Pennsylvania: } \sum_{j=1}^{J} N_{ij}\hat{P}_{+j} = 33,806, \quad \sum_{j=1}^{J} N_{ij} = 156,799$$

- Synthetic est. of percent jaundiced live birth:

$$\text{Pennsylvania: } \hat{P}_i^{SYN} = \frac{33,806}{156,799} = 21.6$$

✓ *González, Placek and Scott (1996)*

Evaluation of synthetic estimates:

- Compare estimates to true values for five states and for selected health characteristics:
  - % low birth weight
  - % late or no prenatal care
  - % low 1-minute Apgar score
- True values from the State Vital Registration data.
- est. MSE=(syn. estimate-true value)^2
- SE (direct est.) using balanced repeated replication.
- Synthetic estimate better especially for smaller states (e.g., Montana).
- Relative root mean squared error (RRMSE):

  Dir. est: 14% (Penn.) to 62% (Montana)

  Syn. est: 0% (Penn.) to 24% (Indiana)

  Exception: 32% (Kansas) exceeds 25%

Table 1: RRMSE of Direct and Synthetic Estimates

Characteristic	Estate	True (%)	Direct est. Est. (%)	Direct est. RRMSE (%)	Syn. est. Est. (%)	Syn. est. RMSE (%)
Low birth:	Pennsylvania	6.5	6.6	15	6.5	0
	Indiana	6.3	6.8	22	6.5	3
	Tennessee	8.0	8.5	23	7.2	10
	Kansas	5.8	6.8	36	6.4	10
	Montana	5.6	9.2	71	6.3	13
Prenatal care:	Pennsylvania	3.9	4.3	21	4.3	10
	Indiana	3.8	2.0	21	4.7	24
	Tennessee	5.4	4.7	26	5.0	7
	Kansas	3.4	2.1	35	4.5	32
	Montana	3.7	3.0	62	4.3	16
Apgar score:	Pennsylvania	7.9	7.7	14	9.4	19
	Indiana	10.9	9.5	16	9.4	14
	Tennessee	9.6	7.3	18	9.7	1
	Kansas	11.1	12.3	25	9.4	15
	Montana	11.6	12.9	40	9.4	19

MSE of synthetic estimator

- Synt. est. depends on direct estimators for large domains. Hence, design variance of synt. est. small in comparison with that of the direct est. for small domain.
- But synt. est. are biased because they depend on strong assumptions.
- Hence, full MSE (accounting for bias and var.) is relevant.

MSE of synthetic estimator

- Approximate MSE:

$$\begin{aligned}\mathrm{MSE}(\hat{Y}_i^{SYN}) &= E(\hat{Y}_i^{SYN} - Y_i)^2 \\ &= E(\hat{Y}_i^{SYN} - \hat{Y}_i^{DIR})^2 + 2\mathrm{Cov}(\hat{Y}_i^{SYN}, \hat{Y}_i^{DIR}) - V(\hat{Y}_i^{DIR}) \\ &\approx E(\hat{Y}_i^{SYN} - \hat{Y}_i^{DIR})^2 - V(\hat{Y}_i^{DIR})\end{aligned}$$

- Estimated MSE:

$$\mathrm{mse}(\hat{Y}_i^{SYN}) = (\hat{Y}_i^{SYN} - \hat{Y}_i^{DIR})^2 - v(\hat{Y}_i^{DIR}).$$

- $\mathrm{mse}(\hat{Y}_i^{SYN})$ is approximately unbiased but unstable.
- Average over domains: *(✓ González and Wakesberg, 1973)*

$$\mathrm{mse}_a(\hat{\bar{Y}}_i^{SYN}) = \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{N_\ell^2}(\hat{Y}_\ell^{SYN} - \hat{Y}_\ell^{DIR})^2 - \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{N_\ell^2}v(\hat{Y}_\ell^{DIR})$$

- Limitation: $\mathrm{mse}_a(\hat{\bar{Y}}_i^{SYN})$ is stable but not area-specific.

MSE of synthetic estimator

- Assumption:

$$b^2(\hat{\bar{Y}}_i^{SYN}) \approx \frac{1}{m}\sum_{\ell=1}^{m} b^2(\hat{\bar{Y}}_\ell^{SYN})$$

- Estimator of average bias:

$$b_a^2(\hat{\bar{Y}}_i^{SYN}) = \text{mse}_a(\hat{\bar{Y}}_i^{SYN}) - \frac{1}{m}\sum_{\ell=1}^{m} v(\hat{\bar{Y}}_\ell^{SYN})$$

- Area-specific MSE estimator: *(✓ Marker, 1995)*

$$\text{mse}_M(\hat{\bar{Y}}_i^{SYN}) = v(\hat{\bar{Y}}_i^{SYN}) + N_i^2 b_a^2(\hat{\bar{Y}}_i^{SYN})$$

- Assumption not satisfied for areas exhibiting strong individual effects.

SPREE (Structure Preserving Estimation)

Example: Estimation of counts in a contingency table

- Available data: Census counts in a three-way table $\{N_{iab}\}$:
  i area index,
  a categories of y (employed/unemployed),
  b categories of auxiliary variable (white/nonwhite)
- Reliable survey estimates of margins $\{\hat{M}_{+ab}\}$, $\{\hat{M}_{i++}\}$
- Target: Find current counts $\{M_{iab}\}$, as close as possible to $\{N_{iab}\}$, that confirm to margins $\{\hat{M}_{+ab}\}$ and $\{\hat{M}_{i++}\}$, and preserves association structure in $\{N_{iab}\}$ as much as possible.

SPREE (Structure Preserving Estimation)

Example: Estimation of counts in a contingency table

- **One-way SPREE:** Minimize w.r.t. $\{x_{iab}\}$ a distance measure:

$$\begin{aligned} \min \quad & D(N_{iab}, x_{iab}) = \sum_{iab} N_{iab} log\{N_{iab}/x_{iab}\} \\ s.t. \quad & \sum_{i} x_{iab} = \hat{M}_{+ab} \end{aligned}$$

- **Solution:** Rake census counts N_{iab}:

$$\bar{M}_{iab} = N_{iab} \frac{\hat{M}_{+ab}}{N_{+ab}}$$

- **Two-way SPREE:** Add the restriction $\sum_{ab} x_{iab} = \hat{M}_{i++}$. Solution: IPF (Iterative Proportional Fitting)

Evaluation of SPREE

- Comparison of SPREE estimates with true mortality counts from Vital Registration System.
- Four difference causes of death (a)
  State (i)
  36 age-sex-race groups (b)
- N_{iab} 1960 census counts.
- $\hat{M}_{+ab} = M_{+ab}$, $\hat{M}_{i++} = M_{i++}$ known current margins.
- ARE $= |\text{est.} - \text{true}|/\text{true}$.
- Two-step SPREE significantly better than one-step SPREE in terms of ARE.

✓ *Purcell & Kish (1980)*

Table 2: Median of Percent ARE of SPREE Estimates

Cause of death	Year	One-step	Two-step
Malignant	1961	1.97	1.85
Neoplasms	1964	3.50	2.21
	1967	5.58	3.22
	1970	8.18	2.75
Major CVR	1961	1.47	0.73
Diseases	1964	1.98	1.03
	1967	3.47	1.20
	1970	4.72	2.22
Suicides	1961	5.56	6.49
	1964	8.98	8.64
	1967	7.76	6.32
	1970	13.41	8.52
Total others	1961	1.92	1.39
	1964	3.28	2.20
	1967	4.89	3.36
	1970	6.65	3.85

Off-the-Shelf Methods

- Customary calibration weights $\{w_k,\ k \in s\}$ satisfy the calibration constraint $\sum_{k\in s} w_k x_k = X$.
- **Objective:** To produce weights w_{ik} for each sample unit k and each small area i such that:

  (i) $\sum_i w_{ik} = w_k$,
  (ii) $\sum_{k\in s} w_{ik} x_k = X_i$ (known area total)

- No unique solution $\Rightarrow$ Assume a multiplicative model on the weights:

$$w_{ik} = \gamma_{ik} \exp(\beta_i' x_k + \delta_k)$$

$$\gamma_{ik} = \begin{cases} 1 & \text{if area } i \text{ is allowed to borrow strength from the} \\ & \text{area in which unit } k \text{ belongs,} \\ 0 & \textit{otherwise}. \end{cases}$$

✓ *Schirm and Zaslavsky (1997)*

Off-the-Shelf method

- **Solution:** Iterative algorithm:

  Step 1: Constraint (i) gives δ_k in terms of β_i.

  Step 2: Replacing the expression for δ_k in terms of β_i in constraint (ii), then (ii) can be written as $g(\beta_i) = 0$. Use Newton-Raphson iterations to find β_i.

  Step 3: Go back to step 1 and so on until convergence.

- **Final estimator for area** i:

$$\hat{Y}_i^{SYN} = \sum_{k \in s} w_{ik} y_k$$

Composite estimators

To balance the bias of a synthetic estimator and the instability of a direct estimator for a domain, take:

$$\hat{Y}_i^C = \phi_i \hat{Y}_i^{DIR} + (1 - \phi_i)\hat{Y}_i^{SYN}, \quad 0 \leq \phi_i \leq 1.$$

- **Sample-size dependent estimator:** For a given $\delta > 0$,

$$\phi_i = \begin{cases} 1, & \text{if} \quad \hat{N}_i \geq \delta N_i; \\ \hat{N}_i/(\delta N_i), & \text{if} \quad \hat{N}_i < \delta N_i. \end{cases}$$

Sample-size dependent estimator

- Under SRS, $\hat{N}_i = Nn_i/n$ and then

$$\phi_i = \begin{cases} 1 & \text{if} \quad n_i \geq \delta E(n_i); \\ n_i/(\delta E(n_i)) & \text{if} \quad n_i < \delta E(n_i) \end{cases}$$

- Canadian LFS: $\delta = 2/3$. For most areas, $1 - \phi_i = 0$; for other areas weight attached to $\hat{Y}_i^{SYN}$ is about 0.1 but never larger than 0.2. *(✓ Drew et al., 1982)*
- All characteristics y get the same weight ϕ_i regardless of their differences w.r.t. between area homogeneity.
- $\hat{N}_i = \sum_{k \in s_i} w_k$ increases with the size of s_i. SSD uses direct estimator $\hat{Y}_i^{DIR}$ if $\hat{N}_i \geq \delta N_i$ even when the expected sample size of the area is small.

Optimal composite estimator

- Find ϕ_i that minimizes MSE($\hat{Y}_i^C$) $\Rightarrow \phi_i^*$
- Optimal weight depends on true MSEs of $\hat{Y}_i^{SYN}$ and $\hat{Y}_i^{DIR}$.
- Estimated optimal weight:

$$\hat{\phi}_i^* = \text{mse}(\hat{Y}_i^{SYN})/(\hat{Y}_i^{SYN} - \hat{Y}_i^{DIR})^2$$

- Limitation: $\hat{\phi}_i^*$ is unstable. Average over variables y or areas or both, but then ϕ_i^* is not area-specific or y-specific.
- Estimated optimal *common weight:*

$$\begin{aligned}\hat{\phi}^* &= \sum_{\ell=1}^{m} mse(\hat{\bar{Y}}_\ell^{SYN}) / \sum_{\ell=1}^{m} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell^{DIR})^2 \\ &= 1 - \left\{ \sum_{\ell=1}^{m} v(\hat{\bar{Y}}_\ell) / \sum_{\ell=1}^{m} (\hat{\bar{Y}}_\ell^{SYN} - \hat{\bar{Y}}_\ell^{DIR})^2 \right\}\end{aligned}$$

Optimal composite estimator

Evaluation of composite estimators

- **Target:** Comparison of direct, synthetic, SSD ($\delta = 1$) and optimal composite estimators of number of unemployed in Health Service Areas (HSAs) of Friuli region in Italy.
- From the 1981 census data, samples from the Italian Labor Force Survey (stratified two-stage design) where simulated.
- Optimal ϕ_i obtained from census.
- 14 Health Service Areas (HSA) which cut across design strata
- Sample: 13 psu's (municipalities), 2290 ssu's (households).

Optimal composite estimator

Evaluation of composite estimators

- Number of Monte Carlo simulations: $L = 400$
- Performance measures:

$$\text{ARB} = \left| \frac{1}{L} \sum_{\ell=1}^{L} \left(\frac{\hat{Y}_i^{C(\ell)}}{Y_i} - 1 \right) \right|$$

$$\text{RRMSE} = \sqrt{\frac{1}{L} \sum_{\ell=1}^{L} (\hat{Y}_i^{C(\ell)} - Y_i)^2 / Y_i}$$

- Averages over 14 HSA's: $\overline{\text{ARB}}$, $\overline{\text{RRMSE}}$

Optimal composite estimator

Table 3: $\overline{\text{ARB}}$ and $\overline{\text{RRMSE}}$

Est.	$\overline{\text{ARB}}$(%)	$\overline{\text{RRMSE}}$(%)
Direct	1.75	42.08
Synthetic	8.97	23.80
Composite	6.00	23.57
SSD($\delta = 1$)	2.39	31.08

- Synthetic and optimal composite estimators about half $\overline{\text{RRMSE}}$ of direct estimator.
- $\overline{\text{RRMSE}}$ of SSD about 30% higher than synthetic.

✓ *Falorsi, Falorsi and Russo (1994)*

James-Stein estimator

- Target parameters: $\bar{Y}_i$, $i = 1, \ldots, m$
- Transformation: $\theta_i = g(\bar{Y}_i)$, $i = 1, \ldots, m$, such that

$$\hat{\theta}_i^{DIR} = g(\hat{\bar{Y}}_i^{DIR}) \overset{ind}{\sim} N(\theta_i, \psi_i);\ \psi_i \text{ known}$$

  Examples: $\theta_i = \bar{Y}_i$, $\theta_i = \log \bar{Y}_i$
- θ_i^0 guess of θ_i, $i = 1, \ldots, m$
- **James-Stein estimator:** $(\psi_i = \psi,\ i = 1, \ldots, m)$

$$\hat{\theta}_i^{JS} = \hat{\phi}_{JS}\hat{\theta}_i^{DIR} + (1 - \hat{\phi}_{JS})\theta_i^0,$$

  with weight

$$\hat{\phi}_{JS} = 1 - [(m-2)\psi]/S, \quad m \geq 3, \quad S = \sum_{i=1}^{m}(\hat{\theta}_i^{DIR} - \theta_i^0)^2$$

James-Stein estimator

- **Choice of θ_i^0:**

  (a) Auxiliary information available:

  $\mathbf{x}_i$ p-vector linearly related to θ_i,
  $\mathbf{X}^T = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$,
  $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1^{DIR}, \ldots, \hat{\theta}_m^{DIR})^T$,
  $\theta_i^0 = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{LS} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \hat{\boldsymbol{\theta}}$

  (b) No x-information: $x_i = 1 \Rightarrow \theta_i^0 = \sum_i \hat{\theta}_i^{DIR}/m$, $i = 1, \ldots, m$

- **Total MSE over areas:** Performance of estimators is evaluated over all areas. For direct estimator,

$$\text{Total MSE}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{m} \text{MSE}(\hat{\theta}_i^{DIR}) = m\psi$$

James-Stein estimator

Theorem:

i) If guess $\boldsymbol{\theta}^0$ equals true value $\boldsymbol{\theta}$,

$$\text{Total MSE}(\hat{\boldsymbol{\theta}}^{JS}) = \sum_{i=1}^{m} \text{MSE}(\hat{\theta}_i^{JS}) = 2\psi$$

ii) JS est. never worse than dir. est. in terms of Total MSE:

$$\text{Total MSE}(\hat{\boldsymbol{\theta}}^{JS}) \leq \text{Total MSE}(\hat{\boldsymbol{\theta}})$$

✓ *James and Stein (1961)*

James-Stein estimator

- **Compromise JS estimator**: $(c > 0)$

$$\hat{\theta}_i^{*JS} = \begin{cases} \hat{\theta}_i^{JS}, & \text{if } \hat{\theta}_i^{DIR} - c\sqrt{\psi} \leq \hat{\theta}_i^{JS} \leq \hat{\theta}_i^{DIR} + c\sqrt{\psi}; \\ \hat{\theta}_i^{DIR} - c\sqrt{\psi}, & \text{if } \hat{\theta}_i^{JS} < \hat{\theta}_i^{DIR} - c\sqrt{\psi}; \\ \hat{\theta}_i^{DIR} + c\sqrt{\psi}, & \text{if } \hat{\theta}_i^{JS} > \hat{\theta}_i^{DIR} + c\sqrt{\psi} \end{cases}$$

- MSE properties of compromise JS estimator: When $c = 1$,

  i) $\text{MSE}(\hat{\theta}_i^{*JS}) < 2\text{MSE}(\hat{\theta}_i^{DIR})$
  ii) $\hat{\theta}_i^{*JS}$ retains more than 80% of the gain of $\hat{\theta}_i^{JS}$ over $\hat{\theta}_i^{DIR}$ in terms of Total MSE.

- Final JS estimator of $\bar{Y}_i$: $\hat{\bar{Y}}_i^{*JS} = g^{-1}(\hat{\theta}_i^{*JS})$
- MSE properties of $\hat{\theta}_i^{*JS}$ are not retained by $\hat{\bar{Y}}_i^{*JS}$

James-Stein estimator

Evaluation of JS estimator: Baseball example

- Target parameters: batting averages of $m = 18$ baseball players during 1970 season: $\theta_i = P_i$, $i = 1, \ldots, 18$
- Direct estimators: batting averages after 45 times at bat in that season: $\hat{\theta}_i^{DIR} = \hat{P}_i^{DIR}$
- Guess: $\theta_i^0 = \frac{1}{18} \sum_{i=1}^{18} \hat{P}_i^{DIR} = 0.265 = \hat{P}_\bullet$
- Sampling variance of dir. est: $\psi = \hat{P}_\bullet(1 - \hat{P}_\bullet)/45 = 0.0043$
- Compromise JS estimator obtained with $c = 1$.
- True batting averages taken from the remainder of season (about 370 more at bats).

Table 4: Batting Averages for 18 Baseball Players

Player	DIR	True	JS	Compromise JS
1	0.400	0.346	0.293	0.334
2	0.378	0.298	0.289	0.312
3	0.356	0.276	0.284	0.290
4	0.333	0.221	0.279	0.279
5	0.311	0.273	0.275	0.275
6	0.311	0.270	0.275	0.275
7	0.289	0.263	0.270	0.270
8	0.267	0.210	0.265	0.265
9	0.244	0.269	0.261	0.261
10	0.244	0.230	0.261	0.261
11	0.222	0.264	0.256	0.256
12	0.222	0.256	0.256	0.256
13	0.222	0.304	0.256	0.256
14	0.222	0.264	0.256	0.256
15	0.222	0.226	0.256	0.256
16	0.200	0.285	0.251	0.251
17	0.178	0.319	0.247	0.243
18	0.156	0.200	0.242	0.221

James-Stein estimator

Evaluation of JS estimator: Baseball example

- Performance measures: Relative accuracy with respect to direct est:

$$R_1 = \sum_{i=1}^{m} (\hat{P}_i^{DIR} - P_i)^2 / \sum_{i=1}^{m} (\hat{P}_i^{JS} - P_i)^2$$

$$R_2 = \sum_{i=1}^{m} (\hat{P}_i^{DIR} - P_i)^2 / \sum_{i=1}^{m} (P_i^{*JS} - P_i)^2$$

- Results: $R_1 = 3.50$, $R_2 = 4.09$
- Compromise estimator protects player 1's $\hat{P}_1^{DIR} = 0.40$ from overshrinking towards $\hat{P}_\bullet = 0.265$

✓ *Efron (1975)*

MSE of James-Stein estimator

- JS estimator can be expressed as

$$\hat{\theta}_i^{JS} = \hat{\theta}_i^{DIR} + h_i(\hat{\boldsymbol{\theta}})$$

- Mean squared error:

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}_i^{JS}) &= E[\hat{\theta}_i^{DIR} + h_i(\hat{\boldsymbol{\theta}}) - \theta_i]^2 \\ &= \psi + 2E[(\hat{\theta}_i^{DIR} - \theta_i)h_i(\hat{\boldsymbol{\theta}})] + Eh_i^2(\hat{\boldsymbol{\theta}}) \\ &= \psi + 2\psi E[\partial h_i(\hat{\boldsymbol{\theta}})/\partial\hat{\theta}_i^{DIR}] + Eh_i^2(\hat{\boldsymbol{\theta}}) \end{aligned}$$

- Unbiased MSE estimator:

$$\mathrm{mse}(\hat{\theta}_i^{JS}) = \psi + 2\psi\partial h_i(\hat{\boldsymbol{\theta}})/\partial\hat{\theta}_i^{DIR} + [h_i(\hat{\boldsymbol{\theta}})]^2$$

- It is unbiased under normality of the direct estimators $\hat{\theta}_i^{DIR}$.
- It is highly unstable and can take negative values.
- Even if $h_i(\hat{\boldsymbol{\theta}})$ has no explicit form, derivatives $\partial h_i/\partial\theta_i$ can be evaluated numerically.

Benchmarking

- Usually a reliable direct estimator for an aggregate A of areas $\hat{Y}_A^{DIR}$ is available.
- Indirect estimators of area totals Y_i do not necessarily add up to $\hat{Y}_A^{DIR}$.
- Ratio adjustment: $\tilde{Y}_i$ indirect estimator of Y_i

$$\tilde{Y}_i^* = \frac{\tilde{Y}_i}{\sum_{i \in A} \tilde{Y}_i} \hat{Y}_A^{DIR} \Rightarrow \sum_{i \in A} \tilde{Y}_i^* = \hat{Y}_A^{DIR}$$

Example

- $\tilde{Y}_i$ indirect estimator of number of school-age children in poverty in county i of State A.
- $\hat{Y}_A^{DIR}$ direct estimator of poverty count in State A.

References

- Choudhry, G.H. and Rao, J.N.K. (2009). On Sample Allocation for Planned Small Areas, Paper presented at the meeting of the International Statistical Institute, Durban, South Africa.
- Deville, J.C. and Särndal, C.E. (1992). Calibration estimation in Survey Sampling, *Journal of the American Statistical Association*, **87**, 376–382.
- Drew, D., Singh, M.P. and Choudhry, G.H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey, *Survey Methodology*, **8**, 17–47.
- Efron, B. (1975). Biased Versus Unbiased Estimation, *Advances in Mathematics*, **16**, 259–277.
- Falorsi, P.D., Falorsi, S. and Russo, A. (1994). Empirical Comparison of Small Area Estimation Methods for the Italian Labour Force Survey, *Survey Methodology*, **20**, 171–176.

References

- Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures, *Econometrica*, **52**, 761–766.
- Fuller, W.A. (1999). Environmental Surveys Over Time, *Journal of Agricultural, Biological and Environmental Statistics*, **4**, 331–345.
- González, M.E. (1973). Use and Evaluation of Synthetic Estimates, *Proceedings of the Social Statistics Section*, American Statistical Association, 33–36.
- González, J.F., Placek, P.J. and Scott, C. (1996). Synthetic Estimation of Followback Surveys at the National Center for Health Statistics, In W.L. Schaible (ed.), *Indirect Estimators in U.S. Federal Programs*, New York: Springer-Verlag, 16–27.
- González, M.E. and Wakesberg, J. (1973). Estimation of the Error of Synthetic Estimates, Paper presented at the first meeting of the International Association of Survey Statistitians, Vienna, Austria.

References

- Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953). Sample Survey Methods and Theory I, New York: Wiley.
- James, W. and Stein, C. (1961). Estimation with Quadratic Loss, *Proceedings of the 4th Berkeley Symposium of Mathematical Statistics and Probability*, Univ. of California Press, Berkeley, 361–379.
- Kish, L. (1999). Cumulating/Combining Population Surveys, *Survey Methodology*, **25**, 129–138.
- Longford, N.T. (2006). Sample Size Calculation for Small-Area Estimation, *Survey Methodology*, **32**, 87–96.
- Marker, D.A. (1995). Small Area Estimation: A Bayesian Perspective, Unpublished Ph.D. Dissertation, University of Michigan, Ann Arbor.
- Marker, D.A. (2001). Producing Small Area Estimates From National Surveys: Methods for Minimizing Use of Indirect Estimators, *Survey Methodology*, **27**, 183–188.

References

- Purcell, N.J. and Kish, L. (1980). Postcensal Estimates for Local Areas (or Domains), *International Statistical Review*, **48**, 3–18.
- Rao, J.N.K. (2003). Small Area Estimation, Hoboken, New Jersey: Wiley.
- Särndal, C.E., Swenson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling, New York: Springer-Verlag.
- Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and Strategies for Small Area Data, *Survey Methodology*, **20**, 13–22.
- Schirm, A.L. and Zaslavsky, A.M. (1997). Reweighting Households to Develop Microsimulation Estimates for States, In 1997 Proceedings of the Section on Survey Research Methods, American Statistical Association, 306–311.