SAE Methods in Official Data Production for Policy Making

Partha Lahiri

Joint Program in Survey Methodology &
Department of Mathematics

University of Maryland, College Park, USA

Pisa, Italy, May 9, 2018

Small Area Estimation in U.S. Federal Programs

1. Infant and maternal health for states (NCHS)
2. Personal income for states and counties (BEA)
3. Post-census populations for counties (USCB)
4. Employment and unemployment for states (BLS)
5. Livestock, crop production for counties (NASS)
6. Disabilities, hospital utilization, physician visits for states (NCHS)
7. Median income for 4-person families for states (USCB)

Figure 1: Administrative Uses of Local Area Unemployment Statistics

ADMINISTRATIVE USES OF LOCAL AREA UNEMPLOYMENT STATISTICS				
User Agency/Program	**2014 Funding (Millions)**	**Geographic Areas Used**	**Reference Period**	**Allocation Formulas/Qualifying Criteria**
Department of Labor – Employment and Training Administration				
Adult Employment and Training Activities (WIA, Title I, Chapter 5)	$766.1	States and Areas of Substantial Unemployment (ASUs). An ASU is a contiguous piece of geography consisting of counties, cities, and/or parts of each, with a population of at least 10,000 and an unemployment rate of at least 6.5 percent. (1) (2)	Most recent program year (July-June).	Funding based on the following proportions: 1/3 on relative number of unemployed in ASUs, 1/3 on relative number of excess unemployed (i.e., number of unemployed in excess of 4.5 percent of labor force), and 1/3 on relative number of economically disadvantaged adults, age 22-72. Not more than 0.25% of funds allocated to outlying areas. (Additional minimum/maximum provisions apply.)
Youth Activities (WIA, Title I, Chapter 4)	$820.4	States and ASUs. (1) (2) (3)	Most recent program year (July-June).	Funding based on the following proportions: 1/3 on relative number of unemployed in ASUs, 1/3 on relative number of excess unemployed, and 1/3 on relative number of economically disadvantaged youth, age 16-21. Not more than 0.25% of funds allocated to outlying areas. Up to 1.5% allocated to Native American programs. (Additional minimum/maximum provisions apply.)
Dislocated Worker Employment & Training Activities (WIA, Title I, Chapter 5)	$1,222.5	States. (1) (2)	Most recent program year (July-June) for unemployed and excess unemployed; most recent calendar year for unemployed 15+ weeks.	Funding based on the following proportions: 1/3 on relative number of unemployed, 1/3 on relative number of excess unemployed, and 1/3 on relative number of individuals unemployed for 15 weeks or more. Not more than 0.25% of funds allocated to outlying areas.
Employment Service Grants to States	$664.2	States. (1)	Most recent calendar year.	State funding algorithm is based on the following proportions: 2/3 of relative number of civilian labor force and 1/3 on relative number of unemployed.
Labor Surplus Areas	(4)	Counties, cities over 25,000 population, and county balances. (1)	Most recent 2-calendar year average.	An area qualifies as a LSA when its average unemployment rate is 20 percent or more above the national average rate (including Puerto Rico) for the period, with the threshold being no lower than 6 percent and no higher than 10 percent.
Federal-State Extended Unemployment Benefits (EB)	(5)	States. (1)	Most recent 3 months for total unemployment trigger (TUR) or most recent 13 weeks for insured unemployment trigger (IUR).	State is eligible to pay EB if: (1) the seasonally adjusted total unemployment rate (TUR) for the most recent 3-month period is at least 6.5 percent and at least 10 percent above the State TUR for the same 3-month period in either of the 2 preceding years, or (2) the insured unemployment rate (IUR) is at least 5 percent and at least 120 percent of the average IUR for the same 13-week period in either of the 2 preceding years.
Youthbuild Program	$77.5	Census tracts and non-metropolitan counties.	Not specified.	An area can qualify if it is an underserved area, which is defined as an area comprised of census tracts with the following distress criteria: (i) a census tract where the unemployment remains high (50 percent or more above the nation's unemployment rate) and (ii) a census tract where a high rate of poverty persists.
Senior Community Service Employment Program (or Community Service Employment for Older Americans)	$434.4 (6)	Counties and cities	Annual Average in 2 of the last 3 years	Participants must be unemployed, 55 years of age or older, and have incomes no more than 125 percent of the Federal poverty level. They qualify as most in need if they reside in an areas with persistent unemployment (Persistent unemployment means that the annual average unemployment rate for a county or city is more than 20 percent higher than the national average for two out of the last three years). Unemployed means an individual who is without a job and who wants and is available for work, including an individual who may have occasional employment that does not result in a constant source of income.
Department of Labor – Veterans' Employment and Training Service				
Jobs for Veterans Act of 2002	$175	States. (1)	Most recent 3-calendar year average.	Funding is based on an estimate of the number of veterans seeking employment in a State as a portion of the number of veterans seeking employment nationwide.
Department of Agriculture				
The Emergency Food Assistance Program (TEFAP)	$268.8	States. (1) (2) (3)	Ten-month average of most recent October-July period.	Farm commodities and funds are allocated based on the following proportions: 3/5 on relative number of persons in households below the poverty line and 2/5 on relative number of unemployed persons.
Waivers to Supplemental Nutrition Assistance Program (SNAP) Time Limits for Able-Bodied Adults Without Dependents (ABAWD)	$73,916.5	States, metropolitan areas (MAs), counties, cities, Indian reservations, and specially designated areas (e.g., census tracts). (1)	Generally 12-month periods, but no less than 3 months for unemployment rate. Not specified for insufficient jobs criterion.	Waivers are granted to areas with: (1) an unemployment rate over 10 percent for the latest 12-month (or 3-month) period (2) insufficient jobs (3) designed as Labor Surplus Area by DOL (4) a 24-mo avg. UR of 20% above national average (5) a low and declining employment - population ratio (6) a lack of jobs in declining occupations or industries (7) described in an academic study/publication as area with lack of jobs or (9) qualifies for extended unemployment benefits.

http://www.bls.gov/lau/lauadminuses.pdf

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.

Figure: Source: Developed by Deepawansa (2018) based on literature review

Sample Surveys vs. Census

- lower cost.
- can be conducted more frequently.
- measurements are more accurate.
- more topics can be covered.
- estimators for large areas are very accurate.

A Caution

Sample size in a domain of interest is too small to use a standard estimator and its standard variance estimator.

Example 1 (Survey of drug use in Nebraska).

- Total sample size was about $4,300$.
- The sample size for Boone County was 14 and only 1 white, female age 25-44 in that county was sampled.

Example 2 (State Sample Sizes with an epsem sample of 10,000 persons).

State	1994 Population (in thousands)	Expected sample size
California	31,431	1,207
Texas	18,378	706
New York	18,169	698
.	.	.
.	.	.
.	.	.
Vermont	580	22
DC	570	22
Wyoming	476	18
Total	260,341	10,000

Figure: Time Series Plots of direct poverty rate estimates for selected Comunas in Chile

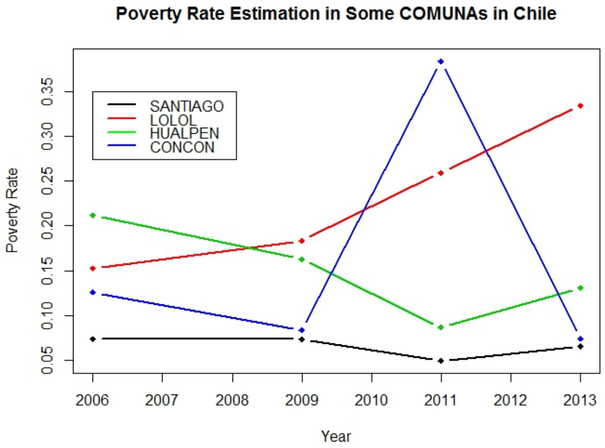
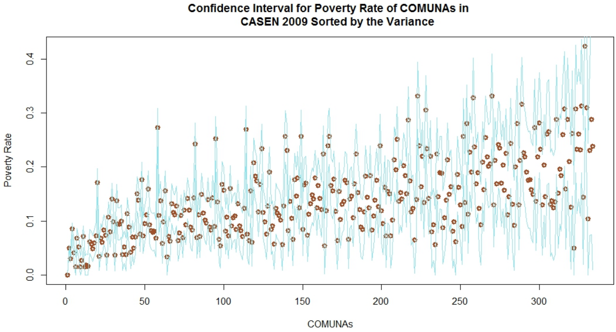


Figure: Direct poverty rate direct estimates and the associated 95% direct confidence intervals for all comunas in CASEN 2009 (sorted by the direct variance estimates)



Datta-Lahiri-Maiti Model

Ref: Datta, Lahiri, Maiti (2002)

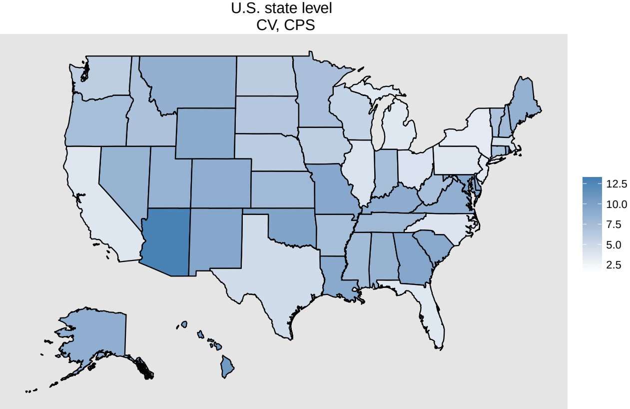
For $i = 1, \cdots, m; t = 1, \cdots, T$,

$$\begin{aligned} \textit{Level 1:} &: \; y_{it} = \theta_{it} + e_{it}; \\ \textit{Level 2:} &: \; \theta_{it} = x_{it}'\beta + v_i + u_{it} \\ \textit{Level 3:} &: \; u_{it} = u_{it-1} + \epsilon_{it} \end{aligned}$$

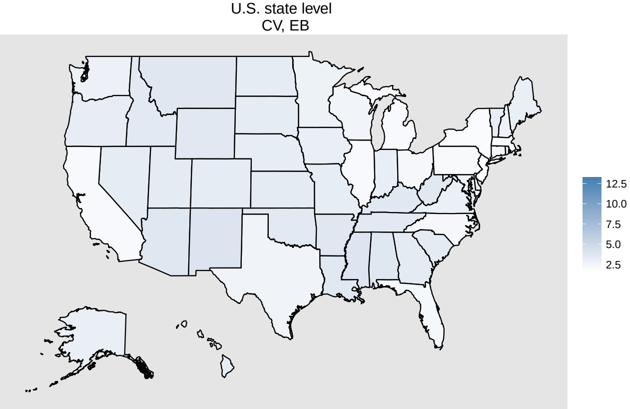
where

- 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

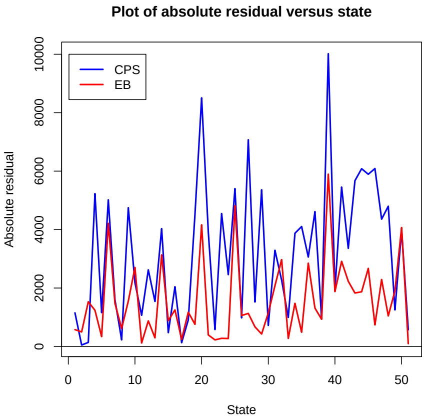


Estimates of Coefficient of Variations of EB estimates of Median Income of 4-person Families in the US States: Year 1989



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



Estimation of FGT Indices

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

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

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.

A Unified Method for Multipurpose Inferences

Partha Lahiri ,

University of Maryland, College Park, USA

&

Jiraphan Suntornchost

Chulalongkorn University, Thailand

Chilean SAE: Notations

- U_c: number of urbanicity statuses for comuna c; since for the urbanicity status, we use urban and rural statuses only, U_c is either 1 or 2 for a given comuna.
- M_{cu}: total number of PSU's in the universe of the uth urbanicity of comuna c.
- N_{cup}: total number of households in the universe of the pth PSU belonging to the uth urbanicity of the cth comuna.
- k_u is the fixed poverty line for unbanicity u ($u = 1$ and $u = 2$ for urban and rural, respectively);
- y_{cuph}: per-capita income of household h (that is, total income of the household divided by the number of household members) in PSU p, urban-rural classification u, comuna c.

The class of FGT indices

$$Q_{c,\alpha} = \frac{1}{N_c} \sum_{u=1}^{U_c} \sum_{p=1}^{M_{cu}} \sum_{h=1}^{N_{cup}} g_\alpha(y_{cuph}),$$

where

- $g_\alpha(y_{cuph}) = (\frac{k_u - y_{cuph}}{k_u})^\alpha \mathcal{I}(y_{cuph} < k_u);$
- α is a "sensitivity" parameter ($\alpha = 0, 1, 2$ correspond to poverty ratio, poverty gap, and poverty severity, respectively).

A hierarchical model

$T_{cuph} = T(y_{cuph})$: a given transformation on the study variable y_{cuph}.

$$
\begin{aligned}
T_{cuph}|\theta_{cup}, \sigma_T &\overset{ind}{\sim} N\left(\theta_{cup}, \sigma_T^2\right) \\
\theta_{cup}|\mu_{cu}, \sigma_\theta &\overset{ind}{\sim} N\left(\mu_{cu}, \sigma_\theta^2\right) \\
\mu_{cu}|\xi_{cu}, \sigma_\mu &\overset{ind}{\sim} N\left(\mathbf{x}_c^T \beta_u, \sigma_\mu^2\right)
\end{aligned}
$$

Inferential Approach

We first note that the estimation of $Q_{c,\alpha}$ is equivalent to that of

$$Q_{c,\alpha} = \frac{1}{N_c} \sum_{u=1}^{U_c} \sum_{p=1}^{M_{cu}} \sum_{h=1}^{N_{cup}} g_\alpha \left(T^{-1}(T_{cuph}) \right),$$

where T is a monotonic function (e.g., logarithm).

Following the theory of Jiang and Lahiri (JASA 2006), we target estimation of:

$$\tilde{Q}_{c;\alpha} \equiv \tilde{Q}_{c;\alpha}(\theta_c, \sigma_T) = \sum_{u=1}^{U_c} \sum_{p=1}^{m_{cu}} \sum_{h=1}^{n_{cup}} w_{cuph} E\{ g_\alpha \left(T^{-1}(T_{cuph}) \right) | \theta_{cup}, \sigma_T \},$$

- w_{cuph}: the survey weight for the hth household in the pth PSU within urbanacity u in comuna c;
- $\theta_c = \text{col}_{u,p}\theta_{cup}$;
- $g_\alpha\left(T^{-1}(T_{cuph})\right) = \left\{\frac{k_u - \left(T^{-1}(T_{cuph})\right)}{k_u}\right\}^\alpha I\left(T_{cuph} \leq l_u\right)$;
- $l_u = \ln(k_u + 1)$, the poverty line of the u^{th} urbanicity in the transformed scale.

The weights are scaled within each comuna so that sum of the weights for all households equals 1.

MCMC

C : number of comunas covered by the model
R : number of MCMC samples after burn-in
$\theta_{c;r}(\sigma_{T;r})$: rth MCMC draw of $\theta_c(\sigma_T)$, $r = 1, \cdots, R$
We define the $C \times R$, matrix $\tilde{Q}^s_\alpha = ((\tilde{Q}^s_{cr;\alpha}))$, where

$$\underset{\sim}{Q}^s_{cr;\alpha} \equiv \tilde{Q}^s_{c;\alpha}(\theta_{c;r}, \sigma_{T;r}).$$

This matrix $\tilde{Q}^s_\alpha$ provides samples generated from the posterior distribution of $\{\tilde{Q}_c, c = 1, \cdots, C\}$ and so adequate for solving a variety of inferential problems in a Bayesian way.

Point Estimation & the Associated Measure of Uncertainty

This is the focus of current poverty mapping research in both classical and Bayesian approaches.

Under SEL function, the Bayes estimate of $Q_{c;\alpha}$ for comuna c and the associated measure of uncertainty are the posterior mean and posterior standard deviation of $\tilde{Q}_{c;\alpha} \equiv \tilde{Q}_{c;\alpha}(\theta_c\,, \sigma_T\,)$, respectively.

These can be approximated by the average and standard deviation across columns of $\tilde{Q}^{s}_{\alpha}$, respectively, for the row c, which corresponds to the comuna c.

Identification of non-compliant comunas

- We would like to flag a comuna for which the true poverty indicator exceeds a pre-specified standard, say a.
- Point estimates whether direct estimates or posterior means do not give any idea about the quality of flagging a comuna for non-compliance.
- A Bayesian solution: Flag comuna c for non-compliance if the posterior probability $P(\tilde{Q}_c > a|\text{data})$ is greater than a specified cutoff
- An Approximation: proportion of columns of $\tilde{Q}^{s}_{c,\alpha}$ exceeding the threshold for row c.

Identification of hot and cold spots

- A common solution: Identify the area with the maximum (minimum) point estimate of the indicator.
- The use of direct point estimates would be quite misleading
- The Bayesian point estimates (posterior means) tend to select areas with more samples.
- No natural quality measure associated with the identification of the hot or cold spots.

A Bayesian Solution

- Select area c as the hot (cold) spot for which $P(\tilde{Q}_c \geq \tilde{Q}_k \forall k|$ data) is the maximum (minimum). Thus, along with the identification of the hot (cold) spot, we also obtain these posterior probabilities suggesting quality of the identification of the hot (cold) spot.

  $\overset{s}{\alpha}$

- We can use $\tilde{Q}$ matrix to approximate these posterior probabilities.

- For row c and column r of $\tilde{Q}$ corresponding to area c and MCMC replicate r, respectively, we can create a binary variable indicating if the area is the hot (cold) spot. Then $P(\tilde{Q}_c \geq \tilde{Q}_k \forall k|$ data) can then be approximated by the average of these binary observations across R columns.

The Chilean Case: The posterior probabilities that poverty rate for a comuna exceeds three different thresholds; $Q_{r,0}$ is direct estimate of regional poverty rate.

	$P(\widetilde{Q}_{c,0} > 1.10Q_{r,0}\vert\text{data})$	$P(\widetilde{Q}_{c,0} > 1.25Q_{r,0}\vert\text{data})$	$P(\widetilde{Q}_{c,0} > 1.50Q_{r,0}\vert\text{data})$
33	1.0000	0.9995	0.6172
13	1.0000	0.9988	0.5636
22	0.9952	0.7962	0.0314
18	0.9904	0.6996	0.0100
2	0.9834	0.4939	0.0005
36	0.6404	0.0731	0.0000
41	0.6142	0.0591	0.0001
37	0.6041	0.0775	0.0000
7	0.5705	0.0386	0.0000
47	0.5179	0.0417	0.0000

The Chilean Case: Posterior probabilities that poverty gap for a given comuna exceeds three different thresholds; $Q_{r,1}$ is direct estimate of regional poverty gap.

	$P(\widetilde{Q}_{c,1} > 1.10Q_{r,1}\|\text{data})$	$P(\widetilde{Q}_{c,1} > 1.25Q_{r,1}\|\text{data})$	$P(\widetilde{Q}_{c,1} > 1.50Q_{r,1}\|\text{data})$
33	1.0000	0.9998	0.9266
13	1.0000	0.9994	0.9060
22	0.9966	0.9143	0.2635
18	0.9918	0.8327	0.1195
2	0.9893	0.7516	0.0395
36	0.6671	0.1631	0.0006
37	0.6399	0.1772	0.0021
7	0.6376	0.1243	0.0002
41	0.6355	0.1365	0.0003
47	0.5586	0.1095	0.0003

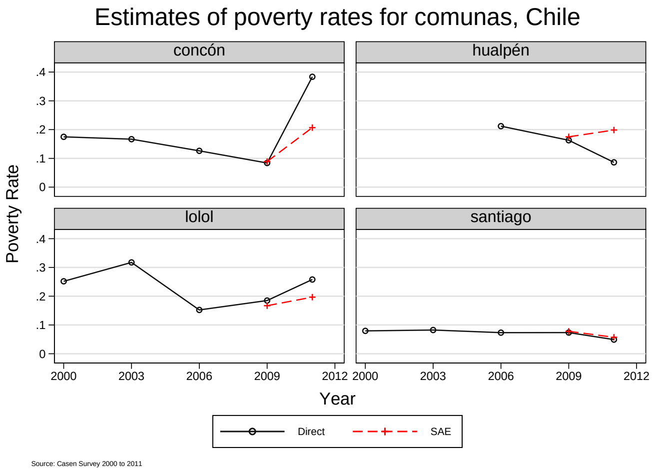
The Chilean Case: Posterior probability that poverty rate or poverty gap for a given comuna is the maximum (Prob.Max) or the minimum (Prob.Min)

COMUNA	Poverty Rate		Poverty Gap	
	Prob.Max	Prob.Min	Prob.Max.Gap	Prob.Min.Gap
33	0.5126	0.0000	0.5246	0.0000
13	0.4496	0.0000	0.4301	0.0000
22	0.0169	0.0000	0.0215	0.0000
18	0.0051	0.0000	0.0044	0.0000
45	0.0025	0.0000	0.0031	0.0000
12	0.0000	0.0121	0.0000	0.0139
48	0.0000	0.0186	0.0000	0.0237
42	0.0000	0.0240	0.0000	0.0268
1	0.0000	0.3929	0.0000	0.3945
8	0.0000	0.5310	0.0000	0.5161

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?

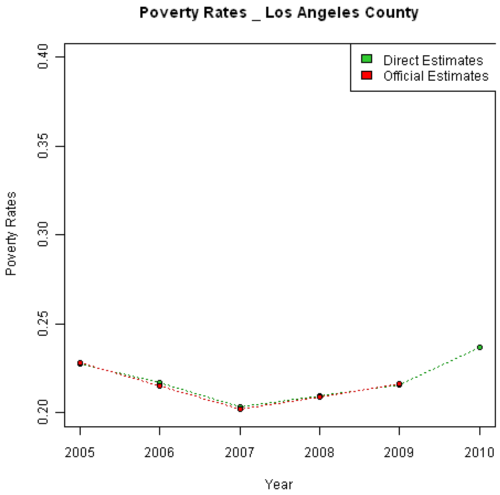
Plots of Survey-Weighted Poverty Rates and SAE for Selected Comunas (drawn by Carolina Casas-Cordero)



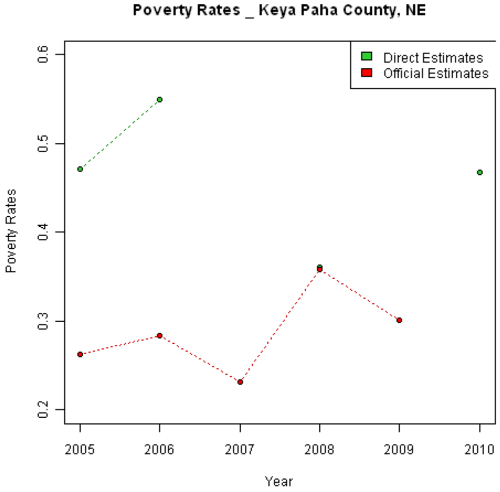
Example: Small Area Income and Poverty Estimates (SAIPE)

- 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).

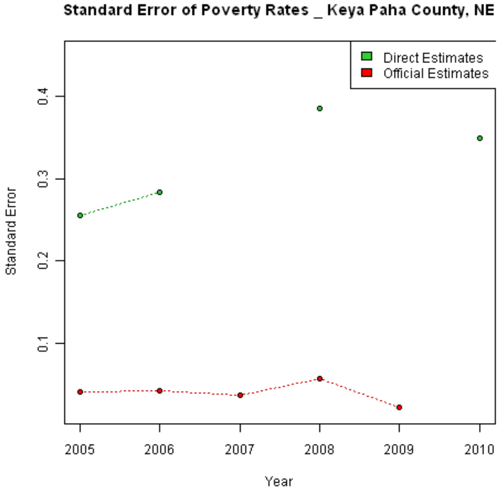
Plots of Survey-Weighted Poverty Rates and SAE for a Small County (drawn by Sam Hawala)



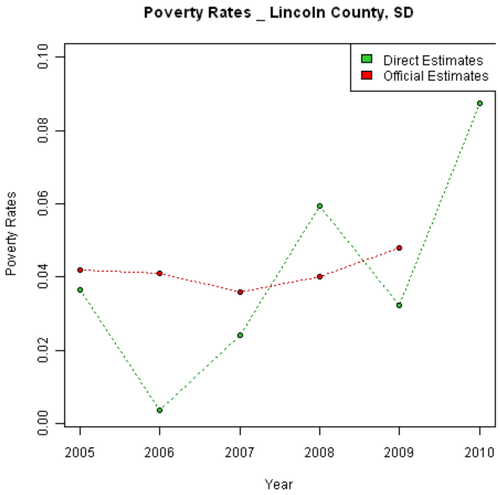
Plots of Survey-Weighted Poverty Rates and SAE for a Small County (drawn by Sam Hawala)



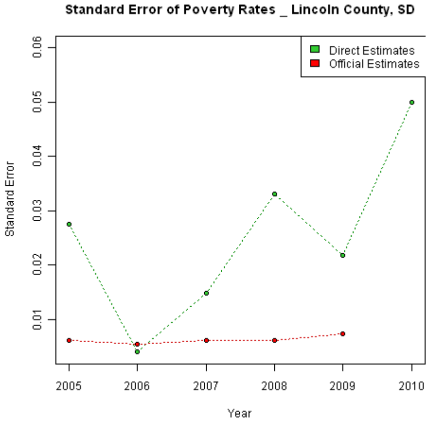
Plots of Survey-Weighted Poverty Rates and SAE for a Small County (drawn by Sam Hawala)



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)



SAE Conferences

- SAE 2015: First Latin American ISI Satellite Conference on Small Area Estimation, Santiago, Chile (`http://www.encuestas.uc.cl/sae2015/program_sae.html`)
- SAE 2014: Small Area Estimation Conference (Poznan, Poland, 2014)
- SAE 2013: The First Asian ISI Satellite Meeting on Small Area Estimation (Bangkok, Thailand, 2013)
- SAE 2011: Conference on Small Area Statistics (Trier, Germany, 2011)
- SAE 2009: Rhine River Cruise Conference 2009 on Recent Advances in Small Area Estimation (Germany, 2009)
- SAE 2009: SAE 2009 Conference on Small Area Estimation (Elche, Spain, 2009)
- SAE 2007: IASS Satellite Conference on SAE (Pisa, Italy, 2007)
- SAE 2001: International Conference on SAE and Related Topics (Maryland, USA, 2001)