Small Area Estimation: Part I

Partha Lahiri

JPSM, Univ. of Maryland, College Park, USA

May 18, 2011

Examples of Small Areas or Domains

Definition: 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).**

Examples of Small Areas or Domains: Contd.

- **US NCHS: Estimation of health variables using the NHANSE III - a majority of US states (small areas) do not have sample**
- **US Census Bureau: Poverty estimation for US counties and school districts using the American Community Survey**
- **NASS-USDA: Estimating crop acres, production and yields for counties**

Small Area Maps

- **A convenient way to display spatial variations of different socio-economic and health related estimates**
  - **Disease mapping**
  - **Poverty Mapping**
- **Reliable maps are useful to public policymakers in planning intervention and allocation of government resources.**

Design-based Direct Estimation (Cochran 1977; Rao 2003)

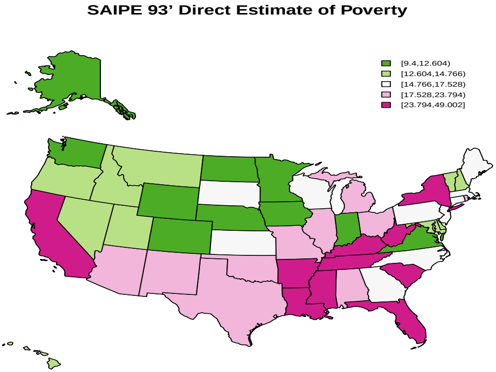
Example: Estimation of poverty rates for over 300 comunas in Chile is of great interest to Chilean government.

Main data source: CASEN, a multipurpose sample survey targeting the civilian non-institutionalized population that resides in housing units throughout the Chilean territory.

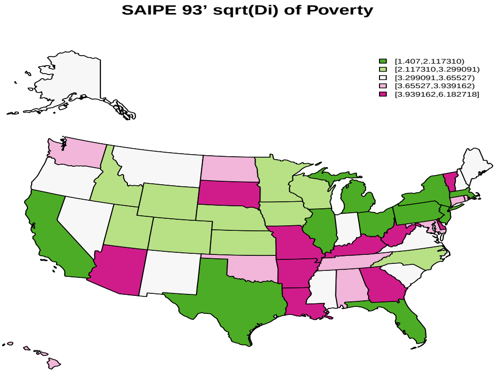
Design-based Direct Estimation Contd.

Design-based direct estimates: **survey weighted proportions that gives differential weights to individuals depending of their inclusion probability into the sample.**

Caution: **The direct estimators are highly unreliable due to small sample sizes in the areas. They have high variability and could be highly biased, depending on the situation.**



Map of Direct Survey Estimates of Poverty Rates



Map of Standard Error Estimates of Direct Survey estimates of Poverty

A Historical Note

- **11th century England and 17th century Canada - based on census or administrative records**

- **There is an 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**

Design Issues

- **Stratification - Use a large number of smaller strata**
- **Degree of Clustering - Minimize clustering**
- **Sample Allocation - Reallocate sample from large planned domains to smaller planned domains**
- **Rolling samples (ACS), multiple frames**

Sample Allocation: Planned Domains

Reallocates 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.**

Example: Canadian Labor Force Survey

Two-Step Allocation: 42,000 Households for national and province level estimates, 17,000 for UIR (SAE) level estimates.

Effects of Reallocation on Areas:

- **Canada:** $E(CV)$ **from 1.3 to 1.5**
- **Ontario:** $E(CV)$ **from 2.8 to 3.2**
- **UI region:** $E(CV)$ **from 17.7 to 9.4**

Borrowing Strength

”..the client will always require more than specified at the design stage” (Fuller, 1999)

Relevant Source of Information

- Census/Administrative information
- Related surveys

Method of Combining Information

- Choices of good small area models
- Use of a good statistical methodology

Synthetic Estimation: A historical note

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

Ref: Hansen et al. (1953)

Two different survey data used

Synthetic Estimation Contd.

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**

Synthetic Estimation Contd.

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

Synthetic Estimation Contd.

- **Using (y_i, x_i) for the 85 sampled counties, the following fitted line (synthetic estimator) was obtained:**

$$\hat{Y}_i^{Syn} = 0.52 + 0.74x_i$$

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

Synthetic Estimation: Implicit Modeling

- N_{ig}= **Female population size for the** g**th race x age-group for the** i**th state. Data source: hospital registration system.**
- $p_{.g}$= **national level direct estimate of the proportion of jaundiced infants whose mother is in the** g**th group. Data source:1980 National Natality Survey.**

Synthetic Estimation: 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	5493	0.173	950
	20-24	7657	0.167	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=**$N_{i.}p_i^s = 156,799 \times 0.216 = 33,869$**.**

Model-Based Methods

- **Flexible**
- **Borrows strength from different relevant sources**
- **Uses appropriate multi-level model that captures different sources of variations**
- **Improves on both direct and synthetic methods**

Example: U.S. Small Area Income and Poverty Estimates (SAIPE) Program

Parameters of interest: true proportions of 5-17 year old children in poverty for the fifty states and the District of Columbia.

Direct estimator: The survey-weighted proportions are obtained using the American Community Survey (ACS) data.

SAIPE Contd.

Auxiliary Variables

- **proportion of child exemptions reported by families in poverty on tax returns**
- **proportion of people under age 65 not included in an income tax return**
- **proportion of people receiving food stamps**

SAIPE Contd.

Two-Level Model

- **Level 1:** **Describes the sampling distribution of ACS survey-weighted poverty rates for the states**
- **Level 2:** **Links the true state poverty rates to state level auxiliary variables**

Estimation Method: **Empirical/Hierarchical Bayes**

Example: County Level Estimation of Crop Yield

- **Yield=production/harvested acreage**
- **Data from multiple surveys are pooled for county estimation**
- **Survey weights are not available. Direct estimates are based on county specific data using a simple county specific regression model.**

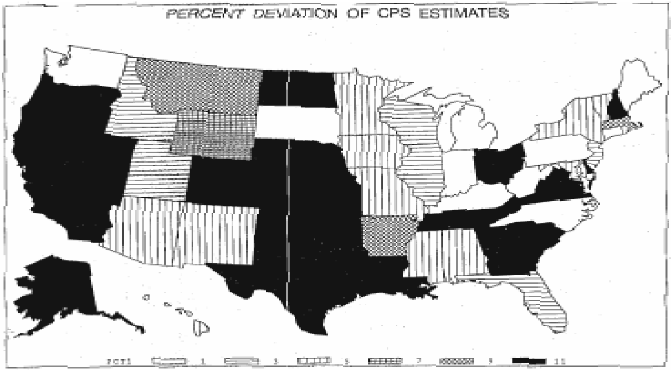
Example: County Level Estimation of Crop Yield

- **Empirical Bayes estimates use pooled survey data plus county level administrative and satellite data**
- **Evaluation criteria (AAD, etc.) are computed by measuring different distances from the census yield and then averaging over all the counties in the state. Smaller the better.**

Example: Estimation of Median Income of Four-Person Families for the states

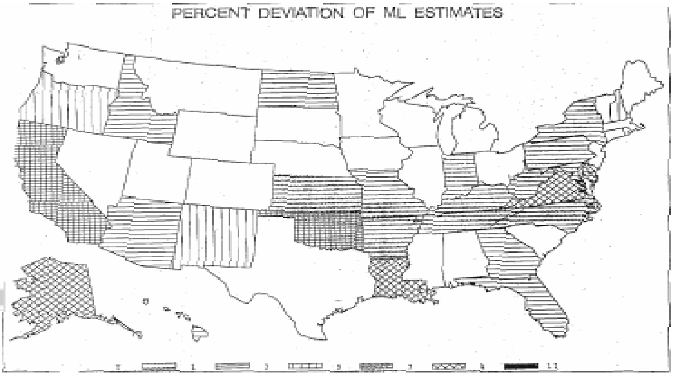
- **Primary Data: Current Population Survey (CPS)**
- **Auxiliary data: administrative/census data**
- **Model: Cross-sectional and time series multi-level model**
- **Evaluation: Map of relative errors:**

  $|est - census|/census$.



Map of Relative Errors of Direct Survey Estimates of Median

Income of 4-Person Families



Map of Relative Errors of Empirical Bayes Estimates of Median

Income of 4-Person Families