HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Small Area Estimation

Spring 2015

Topic 2: Basic concepts and approaches

Risto Lehtonen, University of Helsinki

Lecture topic 2

- **Basic concepts and approaches**
  - Design-based and model-based SAE
  - Planned and unplanned domains
  - Direct and indirect estimators
  - “Borrowing strength”
  - Examples of estimators

Approaches for domain estimation and SAE

- **Design-based methods**
- **Model-based methods**
- ----------
- Bayesian methods (not discussed here)
  - Empirical Bayes, Hierarchical Bayes
- Poverty mapping (not discussed here)
  - World Bank, Peter Lanjouw, Chris Elbers,…
  - PovMap Software

  http://econ.worldbank.org/

Design-based methods

- Estimation approach where the randomness is introduced by the sampling design
- Statistical properties of estimators are evaluated under the sampling design
- Estimators are constructed so that the complexities of the sampling design are accounted for
- Design weight variable (inverse of inclusion probability) is usually incorporated in the estimation procedure

- Key property of design-based estimators
  - (Nearly) design unbiased (by construction principle)
  - Variance can be large if domain sample size is small

Model-based methods

- Estimation approach where the randomness is introduced by an assumed superpopulation model
- Statistical properties of estimators are evaluated under the model
- Sampling design and design weights are not necessarily an issue
  - However, sampling design properties can be accounted for (to some extent, to be discussed!) if desired
  - EXAMPLE: “Pseudo” model-based methods
    e.g. Pseudo EBLUP (Jon Rao 2003)
- Key property of model-based estimators
  - Design biased (by construction principle)
  - Variance can be small even in small domains

Main methods for domain estimation and SAE (for this course)

- **Design-based methods**
  - Horvitz-Thompson (HT) estimators
  - Generalized regression (GREG) estimators
  - Model-free calibration estimators
  - Model calibration estimators
- **Model-based methods**
  - Synthetic SYN estimators
  - Empirical best linear unbiased predictor EBLUP estimators
  - Empirical best predictor EBP type estimators

The role of models

- **GREG estimators** use models as assisting tools
  - GREG estimators are **model-assisted**

  SYN estimators rely exclusively on the model
  - SYN estimators are **model-based (model dependent)**

- NOTE: The same model can be used both in the construction of a GREG estimator and an synthetic or EBLUP / EBP estimator!

Key properties of the methods

Source: Lehtonen and Veijanen (2009)

Table 1
Design-based properties of model-assisted and model-dependent estimators for domains and small areas

	Design-based model-assisted methods	Model-dependent methods
	GREG and calibration estimators	Synthetic and EBLUP estimators
Bias	Design unbiased (approximately) by the construction principle	Design biased Bias does not necessarily approach zero with increasing domain sample size
Precision (Variance)	Variance may be large for small domains Variance tends to decrease with increasing domain sample size	Variance can be small even for small domains Variance tends to decrease with increasing domain sample size
Accuracy (MSE)	MSE = Variance (or nearly so)	MSE = Variance + squared bias Accuracy can be poor if the bias is substantial
Confidence intervals	Valid design-based intervals can be constructed	Valid design-based intervals not necessarily obtained

Natural application areas of estimation approaches by domain sample size

	DOMAIN SAMPLE SIZE		
ESTIMATION APPROACH	Minor	Medium	Major
Model-based			
Synthetic SYN	++	+	0
EBLUP, EBP	+++	++	++
Design-based			
Horvitz-Thompson HT	0	+	++
GREG, MC	+	++	+++

Applicability
0 Not at all + Low ++ Medium +++ High

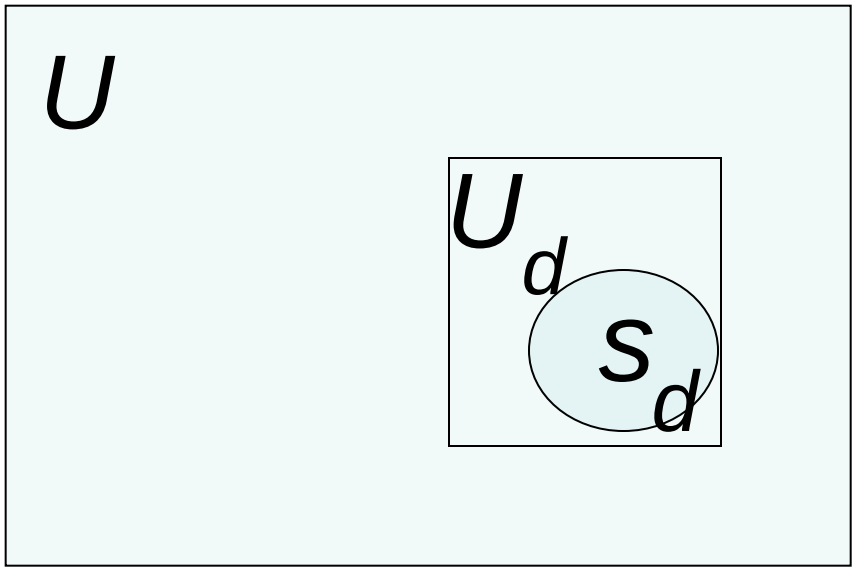
EBLUP: Empirical best linear unbiased predictor
EBP: Empirical best predictor
GREG: Generalized regression estimator
MC: Model calibration

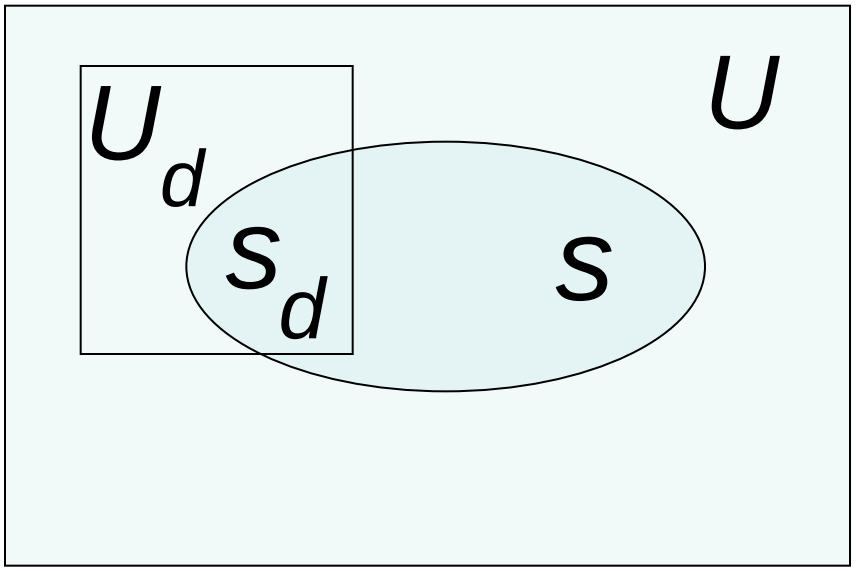
Important aspects and concepts

- **Type of domains of interest**
  - Planned domains / Unplanned domains
- **Type of domain estimator**
  - Direct / Indirect
- **Availability of auxiliary (population) data**
  - Unit-level / Aggregate-level (area-level)
- **Type of model**
  - Linear model / Non-linear model
  - Fixed-effects model / Mixed model
- **Accuracy measures**
  - Variance estimators / MSE estimators

Two main domain structures

- **Planned domains**
  - The most important domains are defined as strata in the sampling design
  - Domain sample sizes are fixed in advance
  - Domain sample sizes are controlled by allocation scheme
  - Small sample sizes can be avoided if desired
- **Unplanned domains**
  - Domain sample sizes are not fixed but are random
  - Small domain sample sizes can occur
  - Most common case in SAE practice





Planned domains

U Population

U_d Population domain d, $d = 1,...,D$

Domains = Strata

$s_d \subset U_d$ Sample drawn in domain d

Sample size n_d is **fixed** by sampling design

Unplanned domains

U Population

s Sample

U_d Population domain d, $d = 1,...,D$

$s_d = s \cap U_d$ Sample falling in domain d

Sample size n_d in domain d is **random**

Domain structures & estimation

- Let us discuss the consequences of the two domain types (planned / unplanned) to the estimation of domain (small area) parameters:
  - Estimation of domain totals, means, ratios, medians,...
  - Estimation of accuracy (variance, MSE) of domain estimates
  - Other points?

Direct and indirect estimation

- **Direct estimation**
  - *Direct* domain estimator uses values of the variable of interest *y* only from the time period of interest and only from units in the domain of interest
    (Federal Committee on Statistical Methodology, 1993)
  - Often in connection to planned domain structures
- **Indirect estimation**
  - *Indirect* domain estimator uses values of the variable of interest *y* from a domain and/or time period other than the domain and time period of interest
  - Often in connection to unplanned domain structures

Domain type and estimator type 1

Domain type	Estimator type	
	Direct	**Indirect**
Planned	Typical set-up	More rarely
Unplanned	More rarely	Typical set-up

Domain type and estimator type 2

- Let us discuss the relationship of domain type (planned / unplanned) and estimator type (direct / indirect) in more detail!

- Why the share of labour between the four combinations can be seen as presented in the table?

Example: Direct HT estimator

Design-based Horvitz-Thompson (HT) estimators

of domain total $t_d = \sum_{k \in U_d} y_k$ and mean $\bar{y}_d = t_d / N_d$

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k \text{ and } \hat{\bar{y}}_{dHT} = \hat{t}_{dHT} / N_d$$

where y_k are measurements of study variable y

$a_k = 1/\pi_k$ are design weights for element $k \in s_d$

$\hat{t}_{dHT}$ and $\hat{\bar{y}}_{dHT}$ only use y-values from s_d

Therefore these HT estimators are **direct**

Example: Direct GREG estimator

Design-based generalized regression GREG estimators of domain total t_d and mean $\bar{y}_d$

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \text{ and } \hat{\bar{y}}_{dGREG} = \hat{t}_{dGREG} / N_d$$

use linear models **fitted separately in each domain**:

$$y_k = \beta_{0d} + \beta_{1d} x_k + \varepsilon_k, \quad k \in U_d, \quad d = 1,...,D$$

where β_{0d} and β_{1d} are domain-specific intercept and slope

$\hat{y}_k = \hat{\beta}_{0d} + \hat{\beta}_{1d} x_k$ are fitted y-values calculated for every $k \in U_d$

Therefore, these GREG estimators are **direct**

Example: Indirect GREG estimator

Design-based generalized regression GREG estimators

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \text{ and } \quad \hat{\bar{y}}_{dGREG} = \hat{t}_{dGREG} / N_d$$

use a linear model **fitted for the entire sample**:

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k, \quad k \in U$$

where β_0 and β_1 are common for all domains, and

$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k$ are fitted y-values calculated for every $k \in U$

Therefore, thse GREG estimators are **indirect**

Example: Indirect SYN estimator

Model-based synthetic SYN estimators

$$\hat{t}_{dSYN} = \sum_{k \in U_d} \hat{y}_k \text{ and } \hat{\bar{y}}_{dSYN} = \hat{t}_{dSYN} / N_d$$

use a linear model **fitted for the entire sample**:

$$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k, \; k \in U$$

where the betas are again common for all domains

Fitted values $\hat{y}_k$ are calculated for every $k \in U$

Therefore, thse synthetic estimators are **indirect**

“Borrow strength”

- In general, indirect estimators are attempting to “borrow strength” from other domains and/or in a temporal dimension
- For domains with small sample size, this is a well justified goal
- The concept of “borrowing strength” is often used in model-based small area estimation

Statistical Science
1994, Vol. 9, No, 1, 55–93

Small Area Estimation: An Appraisal

M. Ghosh and J. N. K. Rao

Abstract. Small area estimation is becoming important in survey sampling due to a growing demand for reliable small area statistics from both public and private sectors. It is now widely recognized that direct survey estimates for small areas are likely to yield unacceptably large standard errors due to the smallness of sample sizes in the areas. This makes it necessary to "borrow strength" from related areas to find more accurate estimates for a given area or, simultaneously, for several areas.

EXAMPLE: Which of the estimators of domain totals discussed this far aim at borrowing strength?

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k, \text{ where } a_k = 1/\pi_k \text{ is design weight}$$

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \text{ with } \hat{y}_k = \hat{\beta}_{0d} + \hat{\beta}_{1d} x_k,\ k \in U_d$$

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \text{ with } \hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k,\ k \in U$$

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) \text{ with } \hat{y}_k = \hat{\beta}_0,\ k \in U$$

- **Let us discuss this point in some more detail!**