NATSEM

Methodologies, Tools and Techniques in Small Area Estimation: An Overview

Azizur Rahman

Presentation to the ARCRNSISS MTT forum workshop, Newcastle
June 5-6, 2008

PhD Candidate, University of Canberra

Supervisors: Professor Ann Harding
Dr. Shuangzhe Liu

http://www.natsem.canberra.edu.au

Outline

- Why is small area estimation necessary?
- Concept and Problems in SAE
- Summary of Methodologies, Tools and Techniques

- Statistical Approaches
  - Small area models
  - Methodologies to model based estimation

- Economic Approaches
  - Microsimulation modelling
  - Techniques to generate Spatial Micropopulation
  - Comparison between reweighting techniques

- Concluding Remarks

Why small area estimation?

- Regional level planning are more effective and getting popular in developed nations
- Policy makers need sufficient information at small area level for intelligible decision making
- An increasing demand for reliable small area statistics from both public and private sectors
- Beneficial for business organizations, policy makers and researchers who are interested in estimates for regional small domains
- Who lack adequate funds for a large-scale survey that could produce precise, direct survey estimates for the small domains

Concept and Problems in SAE

- Concept of Small Area:
  - **The term *small area* typically refers to a small geographical area or a statistical population unit for which reliable statistics of interest cannot be produced due to certain limitations of the available data**
    - A geographical region such as *CBD, suburb, SLA* etc.
    - A demographic group such as a specific *age X sex X education X income* group or a demographic subgroup (ethnic groups, Women-headed households, pensioners etc) within a geographic region

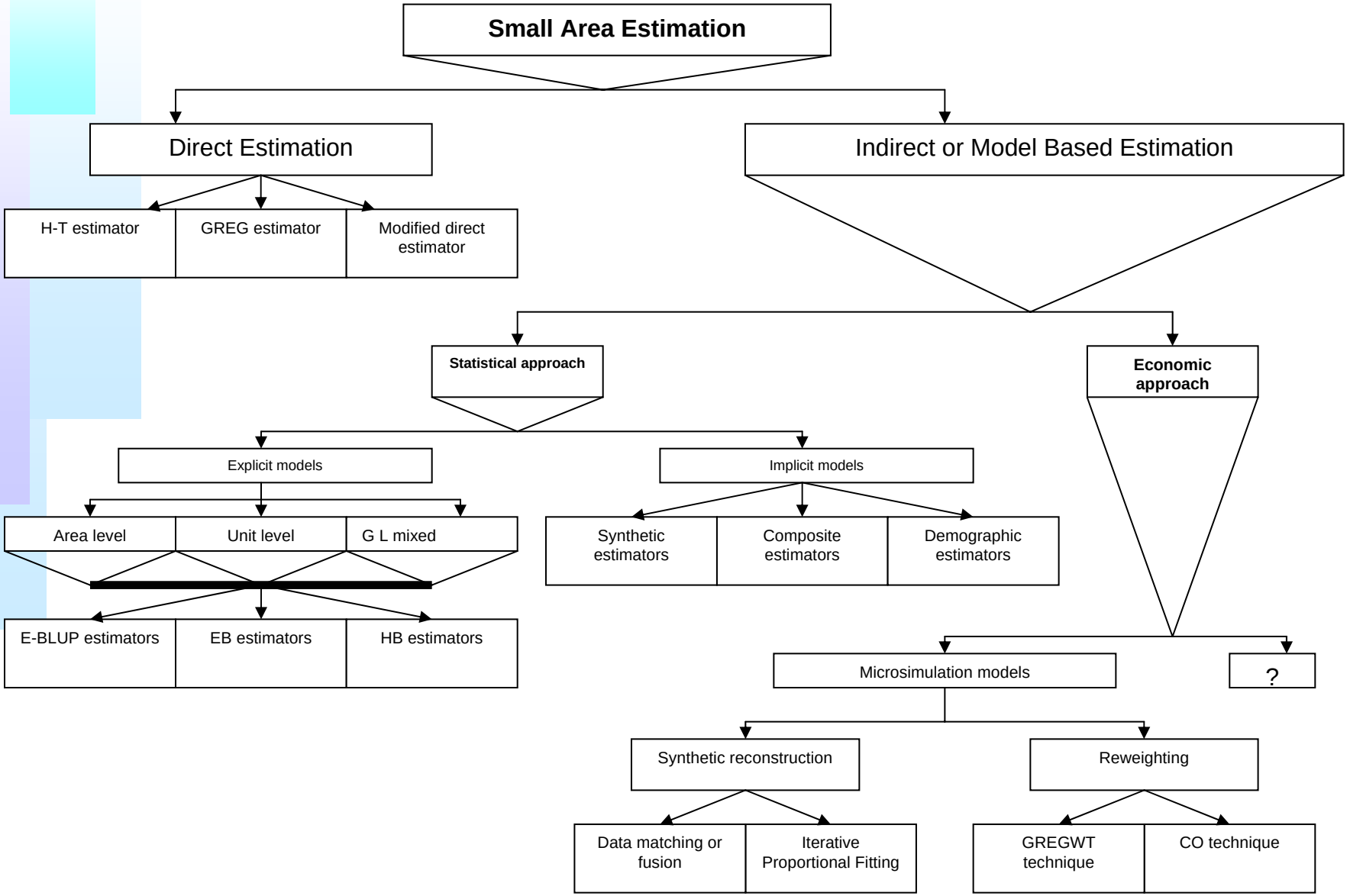
- Small Area Statistics:
  - **Estimates of a variable of interest at small area levels**
    - Population of small areas
    - Population are in housing stress
    - Poverty incidence in ethnic minority communities
    - Proportion of single mothers currently are in workforce
    - Proportion of retirees currently need specific cares in a suburb at Newcastle

Concept and Problems in SAE

Problems

- In reality domain-specific sample data are not large enough for all small areas (even zero for some small areas) to provide adequate statistical precision of their estimates
- The basic problem with national or state level surveys is that they are not designed for efficient estimation for small areas
- Depending on the type of the study and concerning about time and money it can be impossible to conduct a comprehensive sample survey to obtain adequate sample from every small area we are interested in
- Between census years, direct population counts are often not available for many small areas

A summary of Methodologies, tools & techniques in SAE



NATSEM

Statistical Approaches

(see, Ghosh and Rao 1994; Pfeffermann 2002; Rao 2003; Rahman 2008)

Small area models

- *The Fay-Herriot model can be expressed as:*

1. *Linking model:* $\theta_i = x_i'\beta + \varepsilon_i$ => $\theta_i \overset{iid}{\sim} N(x_i'\beta, \sigma_\varepsilon^2),$
2. *Matching model:* $\hat{\theta}_i = \theta_i + e_i$ => $\hat{\theta}_i \mid \theta_i \overset{iid}{\sim} N(\theta_i, \omega_i^2)$

A. Basic Area level model: $\hat{\theta}_i = x_i'\beta + \varepsilon_i + e_i$

B. Basic Unit level model: $y_{ij} = x_{ij}'\beta + \varepsilon_i + e_{ij}$

Where $i = 1,2,\ldots, n;$ *and* $j = 1,2,\ldots, N_i$

NATSEM

Simple ratio estimates

- Ratio estimates at small area i can be defined as

$$\overline{R}_{i[SR]} = \frac{\hat{D}_{.j}}{P} \times P_i$$

Where

$\hat{D}_{.j}$ = direct estimate of the characteristic of interest in large area j

P_i = population of area i

P = total population in large area j

Statistical Approaches

Methodologies to model based estimation

- E-BLUP approach

For the basic area level model defined in (A), the BLUP estimator is given as

$$\tilde{\theta}_i = \gamma_i \hat{\theta}_i + (1-\gamma_i) x_i' \hat{\beta}$$

Where,

$\gamma_i = \sigma_\varepsilon^2 / (\sigma_\varepsilon^2 + \omega_i^2)$; x_i = auxiliary vector at small area i

$\hat{\theta}_i$ = estimate of a function of i^{th} small area population estimates

$\hat{\beta}$ = WLS estimator of β with weight $(\sigma_\varepsilon^2 + \omega_i^2)^{-1}$

Note: In BLUP σ_ε^2 is considered as known, but in practice we require to calculate it from sample observations.

Hence, the BLUP estimator is known as E-BLUP estimator

Statistical Approaches

Methodologies to model based estimation

- EB approach
  - EB approach can be described by the following points:

1. Find the posterior density of μ $(\mu = I'\beta + q'\varepsilon)$,

   $f(\mu | y, \kappa)$ given the data and model parameters vector

2. Estimate κ from $f(y | \kappa)$

3. Find $f(\mu | y, \hat{\kappa})$, the estimated posterior density

4. Use the estimated posterior density for statistical inferences about the small area parameters of interest, μ

Statistical Approaches

Methodologies to model based estimation

- EB estimator:

$$\hat{\theta}_i^{EB} = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) x_i' \hat{\beta}$$

Where,

$$\hat{\gamma}_i = \hat{\sigma}_i^2 / (\hat{\sigma}_i^2 + \omega_i^2), \quad \text{with}$$

$\hat{\sigma}_i^2$ is the estimate of the model parameter (of variance) related to random effects

Statistical Approaches

Methodologies to model based estimation

- HB approach

  1. Essential to consider a prior density of model parameters, $f(\kappa)$

  2. Like the EB method, obtain the posterior density, $f(\mu|y)$

Then inferences on a parameter of interest in small areas are fully based on the posterior density.

- HB estimator:

$$\bar{\theta}_i^{HB} = E(\theta_i \mid \hat{\theta}) = E_{[\sigma_\varepsilon^2|\hat{\theta}]}(\hat{\theta}_i^{HB})$$

If σ_ε^2 is known and $f(\beta) \propto 1$, the HB estimator is identical to the BLUP estimator with a same variability

Require MCMC techniques to EVALUATE $\bar{\theta}_i^{HB}$ & MSE

Economic Approaches

(see, Williamson et al. 1998; Ballas et al. 2003; Brown and Harding 2005; Ballas et al. 2006; Chin and Harding 2006, 2007; Tanton 2007; Rahman 2008)

- Microsimulation
  - Microsimulation is a modelling technique that operates at the level of individual units such as persons, households, or firms.
- Spatial Microsimulation Model
  - Related with small geographic area or small domain
- STINMOD, CareMOD, HouseMOD etc of NATSEM
- Possibility to develop new model for housing

Economic Approaches

- Process of microsimulation modelling
  - Choose optimal Benchmarks
  - Create synthetic population for Small Area
  - Building spatial microsimulation model
  - Validate statistical reliability of small area estimations

Economic Approaches

Techniques to generate Spatial Micropopulation

- Synthetic reconstructions
  - Data fusion or matching
  - Iterative Proportional Fitting (IPF)

- Reweighing techniques
  - GREGWT
  - Combinatorial Optimisation

NATSEM

Economic Approaches

Techniques to generate Spatial Micropopulation

- GREGWT
  - It is an iterative generalized algorithm written in SAS macro to calibrate survey estimates to benchmarks
  - The GREGWT algorithm used a constrained distance function known as the truncated Chi-square distance function that is minimized subject to the calibration equations $\sum_{k \in s} w_k x_k = T_x$ for each small area

$$D_{\chi^2} = \sum_{k \in s} \frac{(w_k - d_k)^2}{d_k} \quad \text{for} \quad d_k L_k \leq w_k \leq d_k U_k$$

Where, T_x is the true population total of the auxiliary information

w_k and d_k are new and sampling weights respectively

L_k and U_k are pre specified lower and upper bounds respectively for each unit $k \in s$

Economic Approaches

Techniques to generate Spatial Micropopulation

- Combinatorial Optimisation (CO)
  - The overall process involves five steps:
    1. **collect a sample survey microdata file (such as CURFs in Australia) and small area benchmark constraints (e.g. from census or administrative records)**
    2. **select a set of households randomly from the survey sample which will act as an initial combination of households at small area**
    3. **tabulate selected households and calculate total absolute error or difference from the known small area constraints,**
       i.e., our Attempt is to minimize $TAE = \sum_{ij} | E_{ij} - O_{ij} | \rightarrow 0$
    4. **choose one of the selected household randomly and change it with a new household drown at random from the survey sample, and then follow step 3 for the new set of households combination**
    5. **repeat step 4 until no further reduction in total absolute difference is possible**

Economic Approaches

Comparison between reweighting techniques

GREGWT

- An iterative process
- Use the Newton-Raphson method of iteration
- Based on a distance function
- Attempt is to minimize the distance function subject to the known benchmarks
- Use the Lagrange multipliers as minimization tools for minimizing the distance function
- Weights are in fractions
- Boundary condition are applied to new weights to achieve a solution
- The benchmark constraints at small area levels are fixed for the algorithm

Combinatorial Optimization

- An iterative process
- Use a stochastic approach of iteration
- Based on a combination of households
- Attempt to select an appropriate combination that best fits the known benchmarks
- Use different combinatorial optimization techniques as intelligent searching tools in optimizing combinations of households
- Weights are in integers
- There is no boundary condition to new weights
- As the algorithm is designed to optimize fit to a selected group of tables, which may or may not be the most appropriate ones.

  Hence, there may be a choice of benchmark constraints

Economic Approaches

Comparison between reweighting techniques

GREGWT

- **Typically, focus on point estimates at small area levels and may be aggregation is possible at larger domains**
- **All estimates have their own standard errors obtained by a group jackknife approach**
- **In some cases convergent do not exist, and require to readjust the boundary limits or a proxy indicator for this non-convergence**
- **There is no standard index to check the statistical reliability of the estimates**

Combinatorial Optimization

- **Offers a flexibility and collective coherence of microdata, making it possible to perform mutually consistent analysis at any level of aggregation or sophistication**
- **There is no information about it in literatures. May be possible in theory, but nothing available in practice yet.**
- **There are no convergent issues. However, finally selected households combination may still fail to fit an user specified benchmark constraints**
- **There is no standard index to check the statistical reliability of the estimates**

Concluding Remarks

- Small area estimation is important for many reasons and widely used in many countries
- Methodologies for indirect estimation play key roles in SAE, and generally there are two approaches for indirect SAE
- Statistical approaches are based on different statistical models such as Area and Unit level models, and each of these models has been studied by different statistical tools & techniques
- Statistical approaches are widely using in USA and Canada. In many cases they have complex computational and unreliability issues and it can not generate small area Microdata/base population

Concluding Remarks

- Economic approaches are based on spatial microsimulation models and they are robust in the sense that further aggregation or disaggregation is possible on the basis of choice of domains, and commonly using in UK & Australia.
- To create reliable spatial micropopulation is the key challenge of these approaches. Two reweighting techniques – GREGWT and combinatorial optimization (CO) are playing vital role to produce small area micro-population.
- A comparison between GREGWT and CO reveals that they are using quite different iterative algorithms and their properties are also different. However, their performances are fairly similar according to the advantages of spatial microsimulation modelling.

References

- Ballas, D., Clarke, G.P. and Turton, I. 2003, 'A spatial microsimulation model for social policy evaluation' in B. Boots and R. Thomas, (eds), *Modelling Geographical Systems*.Kluwer, Netherlands, vol. pp. 143-168.
- Ballas, D., Clarke, G. and Dewhurst, j. 2006, 'Modelling the socio-economic impacts of major job loss or gain at the local level: a spatial microsimulation framework', *Spatial Economic Analysis*, vol. 1, no. 1, pp. 127-146.
- Brown, L. and Harding, A. 2005, 'The new frontier of health and aged care: using microsimulation to assess policy options', Tools for Microeconomic Policy Analysis, Productivity Commission, Canberra.
- Chin, S.-F. and Harding, A. 2006, *Regional Dimensions: Creating Synthetic Small-area Microdata and Spatial Microsimulation Models*, Online Technical Paper - TP33, NATSEM, University of Canberra.
- Chin, S.-F. and Harding, A. 2007, 'SpatialMSM' in A. Gupta and A. Harding, (eds), *Modelling our future: population ageing, health and aged care. Amsterdam*, North-Holland.
- Ghosh, M. and Rao, J.N.K. 1994, 'Small area estimation: an appraisal', *Statistical Science*, vol. 9, no. 1, pp. 55-93.
- Pfeffermann, D. 2002, 'Small area estimation - new developments and directions', *International Statistical Review*, vol. 70, no. 1, pp. 125-143.
- Rao, J.N.K. 2003, *Small Area Estimation*, New Jersey, John Wiley & Sons, Inc.
- Rahman, A. 2008, '*A review of small area estimation problems and methodological developments*, Online Discussion Paper – (forthcoming), NATSEM, University of Canberra.
- Tanton, R. 2007, 'SPATIALMSM: The Australian spatial microsimulation model', 1st General Conference of the International Microsimulation Association, Vienna, 20-21 August, IMA.
- Williamson, P., Birkin, M. and Rees, P. 1998, 'The estimation of population microdata by using data from small area statistics and sample of Anonymised records', *Environment and Planning Analysis*, vol. 30, pp.785-816.

NATSEM