Small Area Estimation Through Spatial Microsimulation Models: Some methodological issues

Azizur Rahman

Presentation to the 2nd General Conference of the IMA: June 8-10, 2009

Government Conference Centre, Ottawa, Canada

Azizur.Rahman@natsem.canberra.edu.au
Outline

● Small area estimation: A quick view

● Methodological issues in SMM
  ● Creation of spatial microdata
  ● Reweighting: GREGWT and CO
  ● Validation

● Some new possibilities in Methodologies
  ● Bayesian prediction
  ● Test statistic; CI estimation

● Concluding remarks
Small area estimation: A quick view

- **SAE – Why?**
  - For sufficient information to intelligible decision
  - For effective and functional regional level planning
  - For business organisations, policy makers and researchers who are interested in spatial estimates
  - For who are in lack of adequate funds to conduct a large-scale survey for all small areas
Estimate of a variable of interest related with issues at small area level.

- Population of small areas
- Households are in housing stress
- Poverty incidence in ethnic minority communities
- Single mothers currently are not in workforce
- Proportion of retirees need specific care at a suburb in Ottawa
Methodological issues in SMM

- Spatial microsimulation is used to create a simulated spatial microdata (e.g. detailed unit record file for SLA)
- Find a CURF data for small area level (from ABS)
- Reweight CURF file to Census benchmarks
- Benchmarks chosen to be relevant to final variable of interest
- But how the process work?
  - Use ‘reweighting’ techniques’
    - GREGWT
    - Combinatorial Optimisation

(see Tanton 2007; Chin and Harding 2006, 2007; Rahman 2008a)
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

\[
D^2 = \sum_{k \in s} \left( \frac{w_k - d_k}{2d_k} \right)^2 \quad \text{for} \quad L_k \leq \frac{w_k}{d_k} \leq U_k
\]

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 \( k \in s \).
Combinatorial optimisation

● The overall process involves five steps:

1. Collect a survey microdata (CURFs in Australia) and small area benchmark (e.g. from census or administrative records) files

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 Distance from the known small area constraints, 
   i.e., our Attempt is to minimize 
   \[ TAD = \sum_{ij} | E_{ij} - O_{ij} | \rightarrow 0 \]

4. Choose one of the selected household randomly and change it with a new household drawn 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 TAD is possible
A comparison of absolute distance and Chi-squared distance measures
NATSEMs’s method

- Reweighting tool is GREGWT (a deterministic method)
- Constrained optimisation process is based on generalised regression
- Convergence achieved by Newton-Raphson method of iteration either all conditions met, or when no improvement in weights under specified convergence criteria
- GREGWT is written in SAS macro
- GREGWT and CO are using quite different iterative algorithms and their properties are also different
Importance of benchmarks and auxiliary data

- Selection of a right benchmark is very important

- A representative auxiliary data should be used

- Better auxiliary data will provide more accurate sample based population estimates

- Differences between sample based estimates and the selected benchmarks have large effect on New Weights, and then finally on our ultimate estimates
Plots of sampling design weights and new weights for specific cases

- $d_k$
- $w_{(C1)}$
- $w_{k(C2)}$
- $w_{k(C3)}$
- $w_{k(C4)}$
Validation

- Validation is an important issue in SMM

- A synthetic spatial microdata is simulated using reweighting techniques that typically does not exist

- Different researchers use their own ways to validate the model outputs

- There is no well accepted statistical means to deal with validation issue
Some new possibilities

- **Bayesian Prediction**

For a variable of interest \( y_{ij} \) at \( i^{th} \) small area, we always have

\[
t_{y_i} = \sum_{j \in S_i} y_{ij} + \sum_{j \in \bar{S}_i} y_{ij}
\]
Bayesian prediction theory

- But how can we relate the observed data to the unobserved?
- ‘Bayesian prediction theory’ can be an answer
  1. Obtain a suitable joint prior distribution:
     \[ p(E_i); \forall i \]
  2. Find the conditional distribution:
     \[ p(y_{ij} : j \in s_i \mid y_{ij} : j \in s_i); \forall i \]
  3. Derive the posterior distribution using Bayes theorem:
     \[ p(\theta | s, X); E_i \subseteq \theta \]
  4. Get simulated copies of the entire population from the posterior

- Benefits: reliable estimates, variance estimation, Bayes CB or CI
- Not very easy to do

(see, Ericson 1969; Lo 1986; Rahman 2008b)
Statistical significance test

● Hypotheses

\[ H_0 : \text{SMM estimates are same as the true values at small areas} \]
\[ H_A : \text{SMM estimates are different from the true values} \]

● In this regard researchers need an effective “TEST SATISTIC”

● We propose two test statistic(s) for small area housing stress estimates in Australia

● Confidence interval estimation

● Essentially need *margin of error* estimate – which is based on the critical value and standard error measures

● Our next manuscript on housing will also address this issue
Concluding remarks

- To generate reliable spatial microdata is the key challenge for small area estimation through SMM
- GREGWT and CO are two common reweighting tools used in SMM
- These reweighting tools are based on different distance measures and using different iterative techniques
- There are possibility of using Bayesian prediction theory as a reweighting tools in SMM
- New way of validation for SMM estimates can be done by statistical test
- CI estimation of SMM estimates may be possible and our next manuscript should address such an issue

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,


Rahman, A. 2008a, A review of small area estimation problems and methodological developments, Online Discussion Paper (DP) - 66 NATSEM, University of Canberra, Canberra.

Rahman, A. 2008b, Bayesian predictive inference for some linear models under Student-t errors, VDM Verlag, Saarbrucken.
