Small area estimation methods and their links to linear and linear mixed models

Stephen Haslett

Massey University, New Zealand

s.j.haslett@massey.ac.nz

Abstract

Small area estimation (sae) is a set of techniques used to provide sufficiently accurate estimates at a finer level than is possible using the data for each small area alone. Indeed small area estimates can be obtained even in areas in which there is no sample. Models for sae come in a variety of forms that extend beyond linear and mixed linear models, but this talk (rather than attempting to be encyclopaedic) will instead consider common themes and differences between sae using linear and linear mixed models, and the usual framework used for linear and linear mixed models. The intent is to provide context for the other talks in this session. Common themes will include estimation of fixed and prediction of random parameters. Differences will include use of sample survey data in sae, and the interest in aggregates of combinations of fixed and random parameter estimates, rather than in prediction of individual observations or estimation and prediction of individual parameters.

Small Area Estimation models

Ghosh and Rao (1994) classify small area models into two broad categories, area level and unit level models. Area level models refer to sets of models that can be considered when only area-specific auxiliary variables are available. Unit level models, on the other hand, refer to models that can be considered when there are unit-specific auxiliary variables and unit level values of the variable under study can be used. All such models are special cases of a general linear or generalized linear mixed model, and usually involve both fixed and random effects.

Area-level models

For area level models, it is assumed that the population mean $(\bar{Y}_a)$ of the a^{th} small area or some suitable function $\theta_a = g(\bar{Y}_a)$ is related to the area-specific auxiliary variables $\mathbf{x}_a = (x_{a1}, \ldots, x_{ap})'$ through a linear model

$$\theta_a = \mathbf{x}_a' \boldsymbol{\beta} + c_a v_a \qquad (1)$$

Area-level models

$$\theta_a = \mathbf{x}_a' \boldsymbol{\beta} + c_a v_a \qquad (1)$$

where $a = 1, \ldots, k$, $v_a \sim \text{iid}(0, \sigma_v^2)$, $\boldsymbol{\beta}$ is a vector of regression parameters, c_a are known or estimated positive constants to allow for heteroscedasticity, k is the total number of small areas under study and p is the number of auxiliary variables. It is assumed that a direct design-based estimator, $\hat{\bar{Y}}_a$, of the population mean $\bar{Y}_a$ is available whenever the area sample size $n_a \geq 1$, and that

$$\hat{\theta}_a = \theta_a + e_a \qquad (2)$$

Area-level models

$$\hat{\theta}_a = \theta_a + e_a \qquad (2)$$

where $\hat{\theta}_a = g(\hat{\bar{Y}}_a)$ and the sampling errors e_a are independent $N(0, V_a)$ with known variance V_a. Combining equation (1) and (2) gives the area level linear mixed model:

$$\hat{\theta}_a = \mathbf{x}_a'\boldsymbol{\beta} + c_a v_a + e_a. \qquad (3)$$

Area-level models

We note that (3) involves both design-based random variables e_a and model-based random variables v_a (Rao 1999), where design-based variables are due to the sample selection mechanism, and model-based ones to the super-population structure in which the model is embedded.

Area level models have various extensions so they can for example handle correlated sampling errors, spatial dependence of random small area effects, time series and cross-sectional data (see Rao 2003, 1999 and Ghosh and Rao 1994).

Unit-level models

The unit level model assumes that the variable of interest Y_{ah} for the h^{th} unit in the a^{th} small area is related to the element-specific auxiliary data $\mathbf{x}_{ah} = (x_{ah1}, \ldots, x_{ahp})'$ through a nested error regression model:

$$Y_{ah} = \mathbf{x}'_{ah}\boldsymbol{\beta} + v_a + e_{ah} \qquad (4)$$

where $a = 1, \ldots, k$, $h = 1, \ldots, N_a$, $\boldsymbol{\beta} = (\beta_0, \ldots \beta_{p-1})$ is $p \times 1$ vector of regression parameters and N_a is the number of population units or households in the a^{th} small area. It is also assumed that the random effects v_a, are iid $N(0, \sigma_v^2)$ and are independent of the unit errors e_{ah} which are assumed to be iid $N(0, \sigma_e^2)$. Extensions that allow errors to be heteroscedastic, with known scaling constant(s) are also possible.

Unit level models - ELL method

Modeling per capita income or expenditure of households instead of poverty measures themselves (such as poverty incidence and gap) is one of the distinctive features of the ELL method. As mentioned in the previous section, the ELL method involves fitting the income or expenditure model to the survey data and applying it to the census data prior to the generation of the small area estimates of poverty measures. The income/expenditure model is as follows:

$$Y_{bh} = \mathbf{x}'_{bh}\boldsymbol{\beta} + u_{bh} \qquad (5)$$

where $b = 1, ..., M$, $h = 1, ..., N_b$; Y_{bh} is the log-transformed per capita income or expenditure of the h^{th} unit or household in the b^{th} cluster, M is the total number of clusters in the population and N_b is the total number of households in the b^{th} cluster in the population. $\mathbf{x}_{bh}$ is a set of the auxiliary variables available in both the survey and the census, which generally need to be contemporaneous;

Unit level models - ELL method

Model (5) can be written as

$$Y_{bh} = \mathbf{x}'_{bh}\boldsymbol{\beta} + v_b + e_{bh} \qquad (7)$$

which is similar in form to the unit level model or nested error regression model mentioned in the previous section. However while the form of the model is similar, the group being referred to is different, *e.g.*, Y_{ah} refers to the h^{th} household in the a^{th} small area, while Y_{bh} refers to the h^{th} household in the b^{th} cluster. Clusters, based on the survey design, will typically be much smaller than the areas for which small area estimates are sought, and generally (unlike almost all the small areas) not all clusters are sampled. For example in the Philippines, estimates are sought at the municipal level which is composed of barangays or clusters.

Fitting methods

- There are a range of possible methods of fitting small area estimation models
- The methods have obvious links to fitting of linear mixed (or generalized linear) models
- The additional complication however is that the data are generally from a sample survey with complex design, for example including stratification, clustering, and unequal selection probabilities.

Pseudo-empirical BLUP

You and Rao (2002) proposed an estimator of the small area mean by deriving an estimator of $\boldsymbol{\beta}$ based on the unit level model (4). The process of deriving the estimator of $\boldsymbol{\beta}$ starts with the computation of the best linear unbiased predictor (BLUP) of v_a given the parameters $\boldsymbol{\beta}$, σ_e^2 and σ_v^2 from the aggregated (survey-weighted) area level model:

$$\bar{Y}_{aw} = \bar{\mathbf{x}}'_{aw}\boldsymbol{\beta} + v_a + \bar{e}_{aw} \qquad (14)$$

Pseudo-empirical BLUP

which proceeds as follows:

$$\hat{v}_{aw}(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2) = \gamma_{aw}(\bar{y}_{aw} - \bar{\mathbf{x}}'_{aw}\boldsymbol{\beta}) \qquad (15)$$

where $\bar{\mathbf{x}}_{aw} = \sum_{h=1}^{n_a} w_{ah}\mathbf{x}_{ah}$, $\bar{y}_{aw} = \sum_{h=1}^{n_a} w_{ah}y_{ah}$, $\gamma_{aw} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2\,\delta_a^2)$, $w_{ah} = \tilde{w}_{ah}/\sum_{h=1}^{n_a}\tilde{w}_{ah}$, $\delta_a^2 = \sum_{h=1}^{n_a} w_{ah}^2$, and $\tilde{w}_{ah}$ are the unit level survey weights; then solving for the survey-weighted estimating equation for $\boldsymbol{\beta}$:

$$\sum_{a=1}^{k}\sum_{h=1}^{n_a} \tilde{w}_{ah}\mathbf{x}_{ah}[y_{ah} - \mathbf{x}'_{ah}\boldsymbol{\beta} - \hat{v}_{aw}(\boldsymbol{\beta}, \sigma_e^2, \sigma_v^2)] = 0 \qquad (16)$$

Pseudo-empirical BLUP

from which the estimator of $\boldsymbol{\beta}$ is obtained as

$$\hat{\boldsymbol{\beta}}_w = \left\{ \sum_{a=1}^{k} \sum_{h=1}^{n_a} \mathbf{x}_{ah} \mathbf{z}'_{ah} \right\}^{-1} \left\{ \sum_{a=1}^{k} \sum_{h=1}^{n_a} \mathbf{z}_{ah} y_{ah} \right\} \qquad (17)$$

where $\mathbf{z}_{ah} = \tilde{w}_{ah}(\mathbf{x}_{ah} - \gamma_{aw}\bar{\mathbf{x}}_{ah})$. The corresponding covariance matrix is then as follows:

$$\begin{aligned} \boldsymbol{\Phi}_w = {} & \sigma_e^2 \left(\sum_{a=1}^{k} \sum_{h=1}^{n_a} \mathbf{x}_{ah} \mathbf{z}'_{ah} \right)^{-1} \\ & \left(\sum_{a=1}^{k} \sum_{h=1}^{n_a} \mathbf{z}_{ah} \mathbf{z}'_{ah} \right) \left(\sum_{a=1}^{k} \sum_{h=1}^{n_a} \mathbf{x}_{ah} \mathbf{z}'_{ah} \right)^{-1} \\ & + \sigma_v^2 \left(\sum_{a=1}^{k} \sum_{h=1}^{n_a} \mathbf{x}_{ah} \mathbf{z}'_{ah} \right)^{-1} \\ & \left\{ \sum_{a=1}^{k} \left(\sum_{h=1}^{n_a} \mathbf{z}_{ah} \right) \left(\sum_{a=1}^{n_a} \mathbf{z}_{ah} \right)' \right\} \left\{ \left(\sum_{a=1}^{k} \sum_{h=1}^{n_a} \mathbf{x}_{ah} \mathbf{z}'_{ah} \right)^{-1} \right\}'. \end{aligned} \qquad (18)$$

Pseudo-empirical BLUP

The variance components are estimated using Henderson's Method 3 (Henderson 1953), to generate unbiased estimates even in the presence of correlated elements in the model.

Iterative weighted estimating equation method

The estimator proposed by You *et al.* (2003) is similar to the Pseudo-EBLUP estimator, except that it incorporates the sampling weights in the computation of the variance components, and it generates the parameter estimate $\boldsymbol{\beta}$ and the variance components by using an iterative weighted estimating equation (IWEE) approach. The authors derived the estimator of σ_e^2 and σ_v^2 as follows:

$$\hat{\sigma}_{ew}^{2(t)} = \frac{\sum_{a=1}^{k}\sum_{h=1}^{n_a} \tilde{w}_{ah}\left[y_{ah} - \bar{y}_{aw} - (\mathbf{x}_{ah} - \bar{\mathbf{x}}_{aw})'\hat{\boldsymbol{\beta}}^{(t-1)}\right]^2}{\sum_{a=1}^{k}\left[(1-\delta_a^2)\sum_{h=1}^{n_a}\tilde{w}_{ah}\right]}$$

$$\equiv \tilde{\sigma}_{ew}^{2(t)}(\boldsymbol{\beta}) \qquad (21)$$

Iterative weighted estimating equation method

and

$$\tilde{\sigma}_{vw}^{2(t)} = \frac{1}{k}\sum_{a=1}^{k} \tilde{v}_{aw}^2 + \frac{\tilde{\sigma}_{vw}^{2(t-1)}}{k}\sum_{a=1}^{k}(\gamma_{aw}-1)^2 + \frac{\tilde{\sigma}_{ew}^{2(t)}}{k}\sum_{a=1}^{k}\delta_a^2\gamma_{aw}^2$$
$$\equiv \tilde{\sigma}_{vw}^{2(t)}(\tilde{v}_w, \sigma_e^2, \sigma_v^2). \qquad (22)$$

The survey weighted estimates of $\boldsymbol{\beta}$, σ_e^2, σ_v^2 are obtained simultaneously by following iterative updating steps, t in the equation above stands for the t^{th} iteration. Since the variance components σ_v^2 and σ_e^2 are unknown, initial estimates for the iterative steps are generated by Henderson's method. Again, as for Pseudo-EBLUP, for the ELL regression model formulation (7), the subscript a is replaced by b.

Iterative weighted estimating equation method

This approach is similar to the probability-weighted iterative generalized least squares (PIWGLS) method proposed by Pfeffermann *et al.* (1998) for fitting multilevel models where the estimation process considered the unequal selection probabilities at each stage of sampling and involves iterating between the parameter $\boldsymbol{\beta}$ and the variance components until convergence. A model-based approach is also proposed by Pfeffermann, Moura and Silva (2006), which involves deriving the hierarchical model for given sample data as a function of the population model and the selection probabilities, and then fitting the sample model using Bayesian approach by use of Markov Chain Monte Carlo algorithm.

Generalised survey regression method

Another approach to generate the estimator of the parameter β and its variance is the design-based methodology for fitting regression models (Lohr 1999). This technique is currently used in the Stata, Sudaan, and WesVar package, for example. The estimator of $\boldsymbol{\beta}$ given below is the sample weighted regression estimator for a model with homoscedastic variance structure and uncorrelated observations in the population.

$$\hat{\boldsymbol{\beta}}_S = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}. \qquad (23)$$

Generalised survey regression method

This estimator is not derived under the model specified by (7) even under the homoscedastic variances for household errors. The linearized/robust variance estimate for $\hat{\boldsymbol{\beta}}_S$ is based on the design-based variance estimator for a total, given as,

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_S) = \mathbf{D}\left\{\frac{m}{m-1}\sum_{b=1}^{m}\left(\sum_{h=1}^{n_b} w_{bh}\mathbf{d}_{bh}\right)'\left(\sum_{h=1}^{n_b} w_{bh}\mathbf{d}_{bh}\right)\right\}\mathbf{D} \qquad (24)$$

where $\mathbf{d}_{bh} = \hat{e}_{bh}\mathbf{x}_{bh}$; $\hat{e}_{bh}$ is the residual from WLS regression; $\mathbf{x}_{bh}$ is a vector of the independent variables; w_{bh} is a sampling weight; $\mathbf{D} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$; and $\mathbf{W}$ is a diagonal matrix of the sampling weights.

Generalised survey regression method

The General Survey Regression method differs from the other techniques in the computation of the estimates, and generates the estimates without computing the variance components, σ_v^2 and σ_e^2. As shown above, the equations for the estimator of the parameter $\boldsymbol{\beta}$ and its corresponding estimated covariance matrix only involve the sampling weights matrix $\mathbf{W}$. The estimated covariance matrix in (24) is often referred to as a sandwich estimator.

Millennium Development Goals, Targets and Indicators

Goal 1. Eradicate extreme poverty and hunger

Target 1.
Halve, between 1990 and 2015, the proportion of people whose income is less than one dollar a day

Indicators
1. Proportion of population below $1 (1993 PPP) per day (World Bank)[a]
2. Poverty gap ratio [incidence x depth of poverty] (World Bank)
3. Share of poorest quintile in national consumption (World Bank)

Target 2.
Halve, between 1990 and 2015, the proportion of people who suffer from hunger

Indicators
4. Prevalence of underweight children under five years of age (UNICEF-WHO)
5. Proportion of population below minimum level of dietary energy consumption (FAO)

Footnotes:

[a] **For monitoring country poverty trends, indicators based on national poverty lines should be used, where available.**

FGT Measures

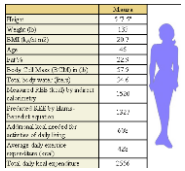
$$P_R = \frac{1}{N_R} \sum_{ij \in R} \left(\frac{Z - Y_{ij}}{Z} \right)^{\alpha} Ind(Y_{ij} < Z)$$

- Poverty Incidence – proportion below the poverty line
- Poverty Gap – average amount below the poverty line
- Poverty Severity – gives more weight to the very poor

FGT = Foster-Greer-Thorbecke

Food Poverty Line

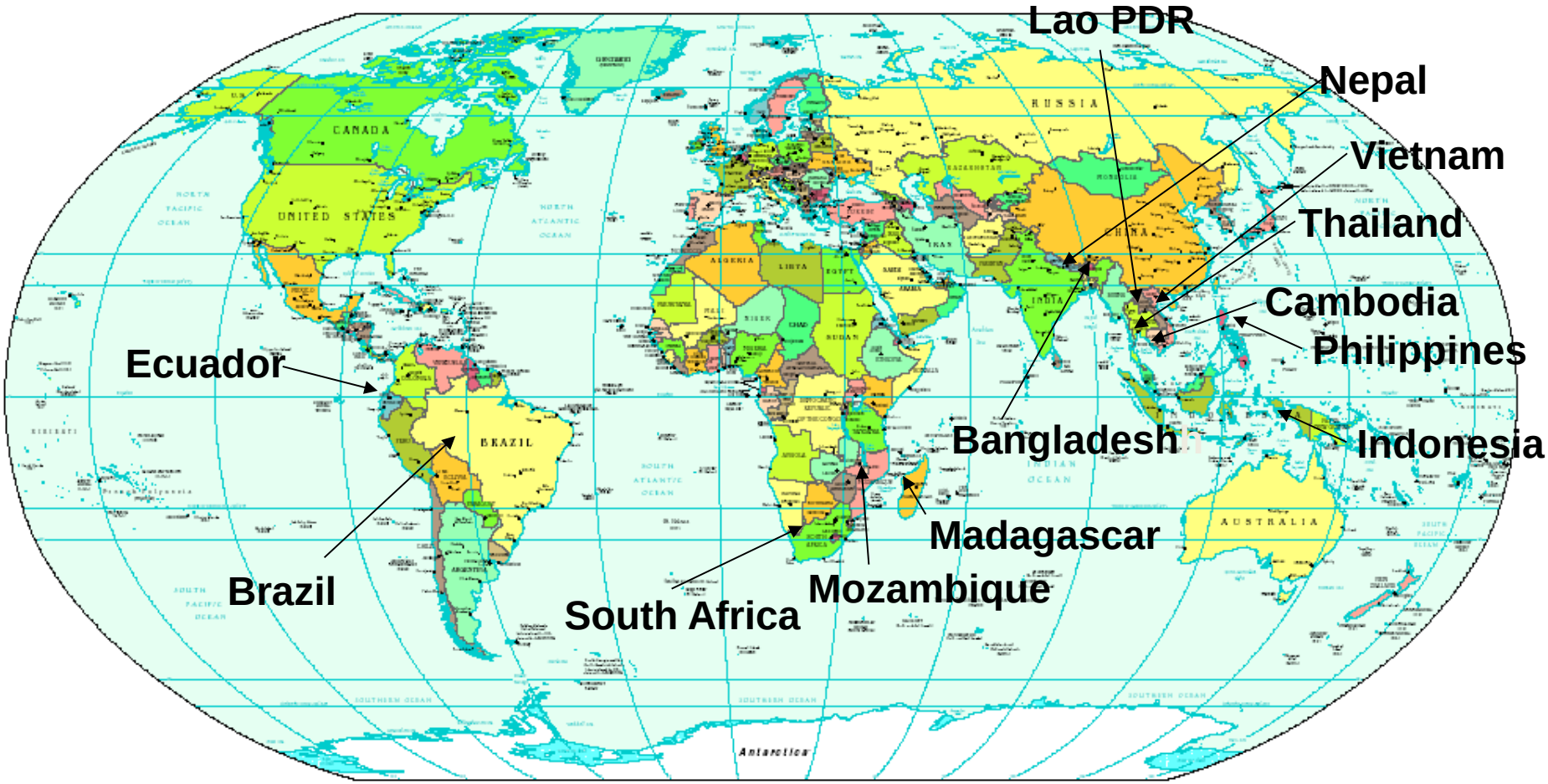
- **Food requirements (kcal/day)**
- **Basket of typical food items**
- **Kcals from "basket"**
- **Cost of suitably-scaled basket**
- **Cash required for a person to be able to eat enough**



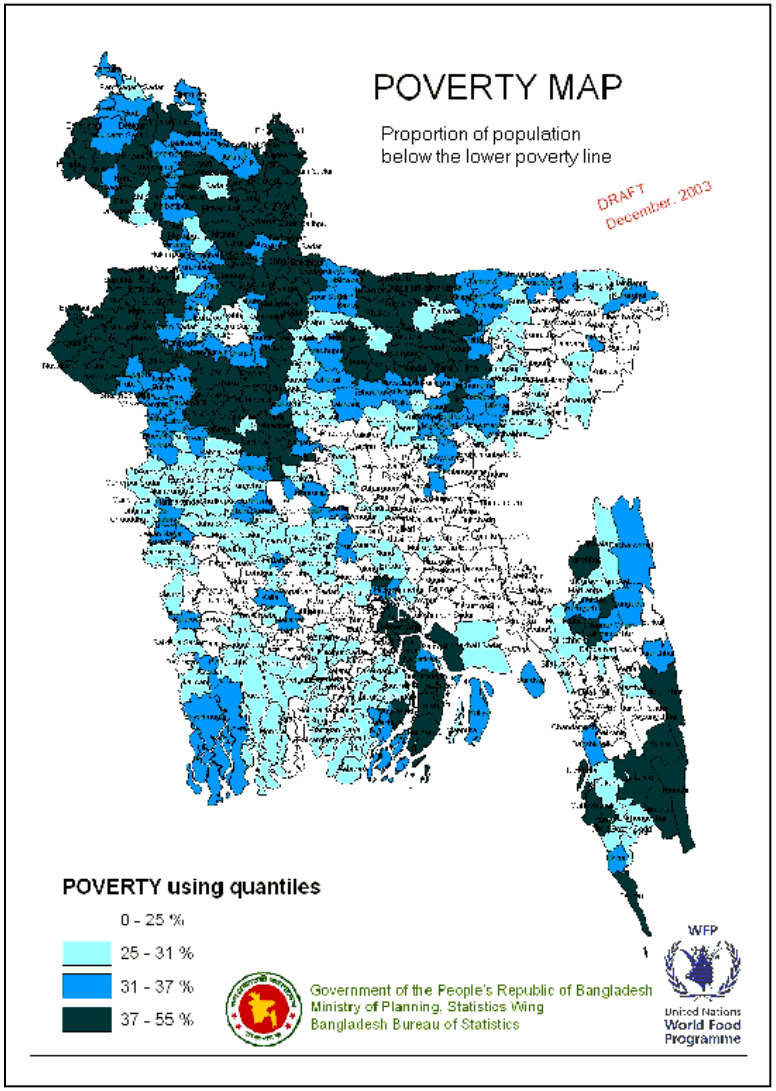
3.60 kcal/g

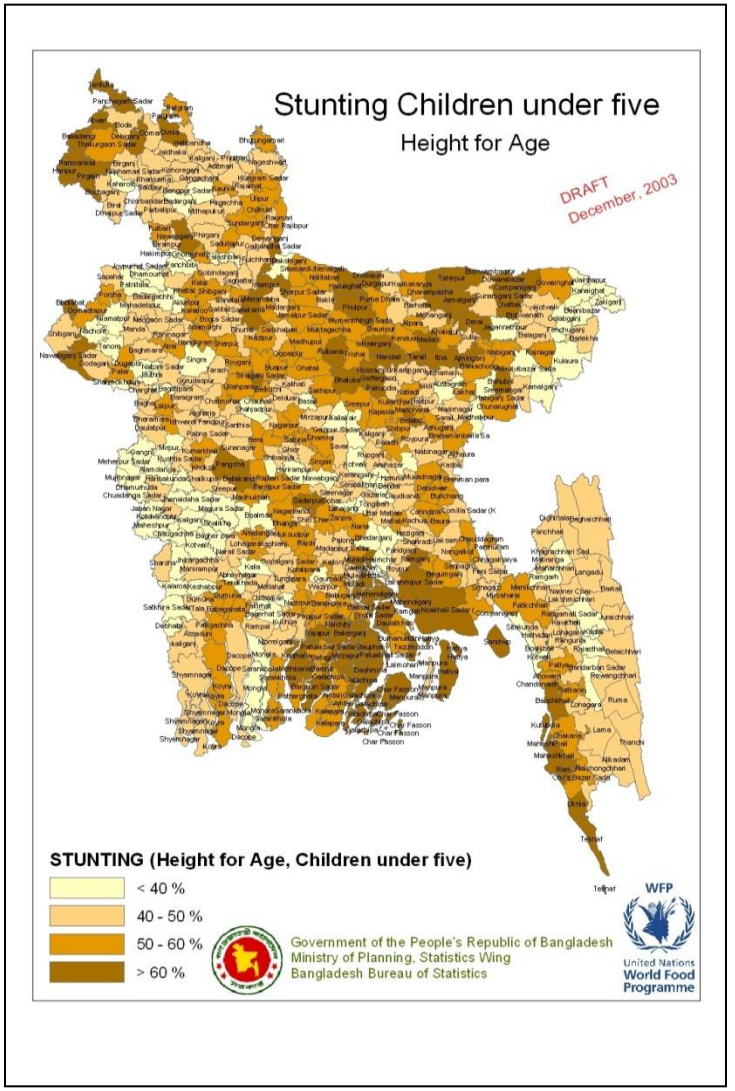
$1.25

Where?

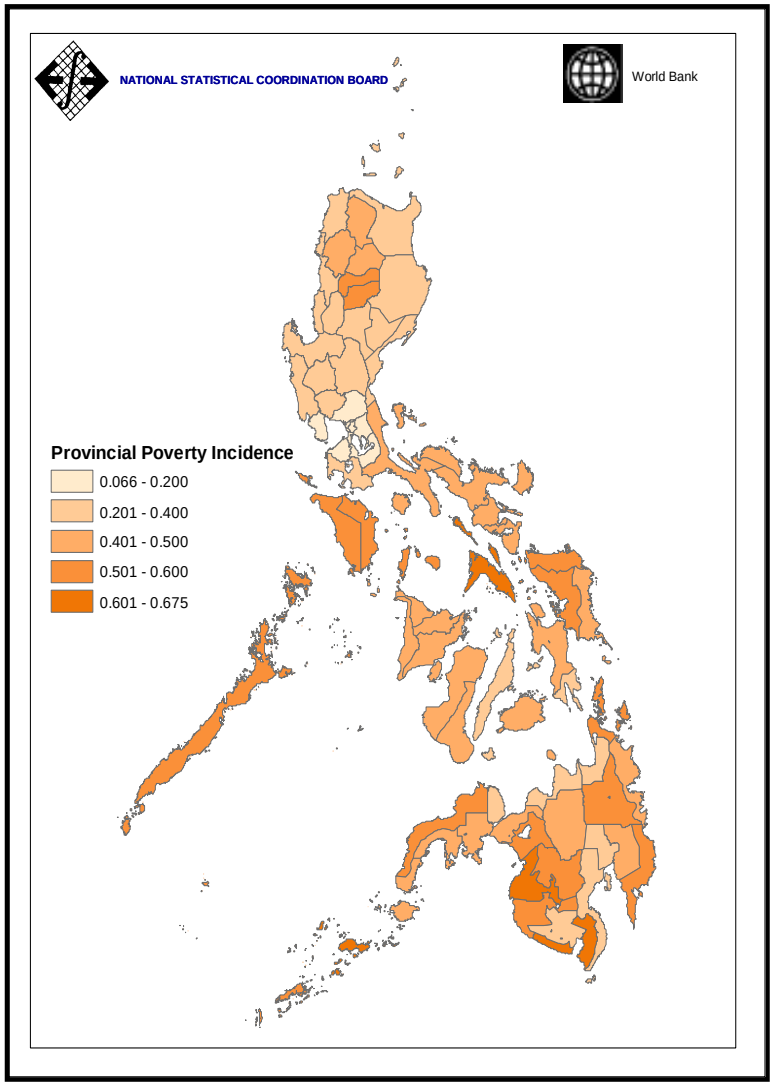

Lao PDR
Nepal
Vietnam
Thailand
Cambodia
Philippines
Indonesia
Bangladesh
Madagascar
Mozambique
South Africa
Brazil
Ecuador
CANADA
UNITED STATES
BRAZIL
RUSSIA
CHINA
INDIA
AUSTRALIA
Antarctica

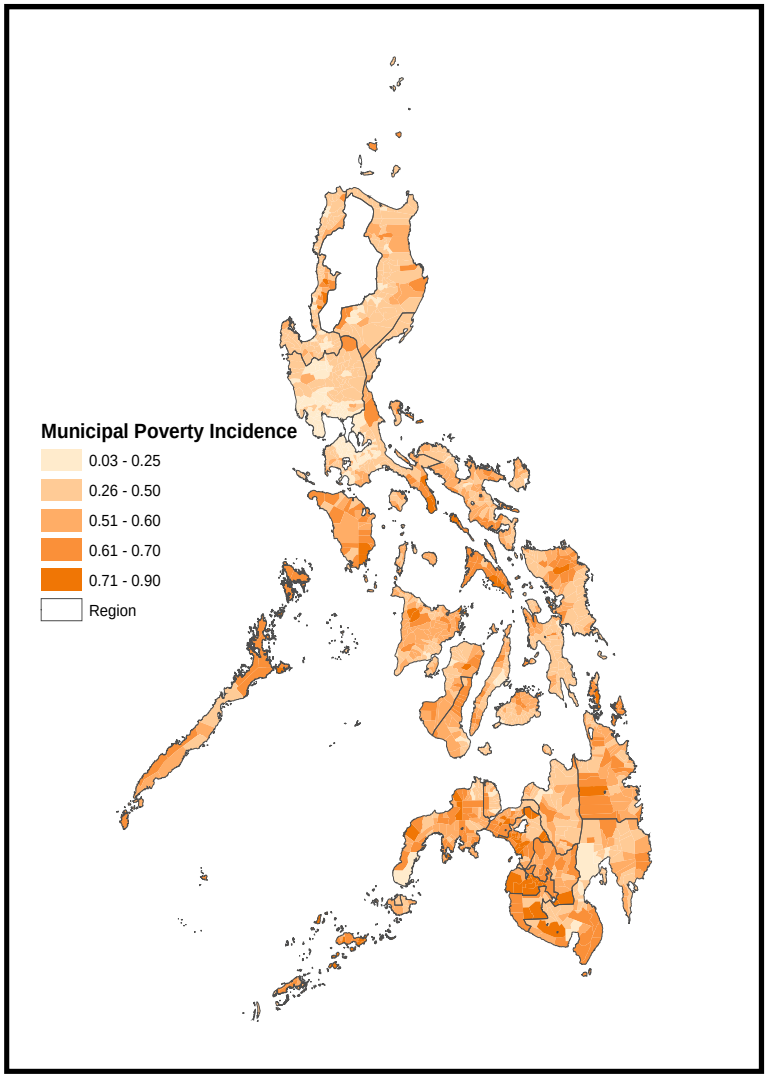
SAE - Bangladesh



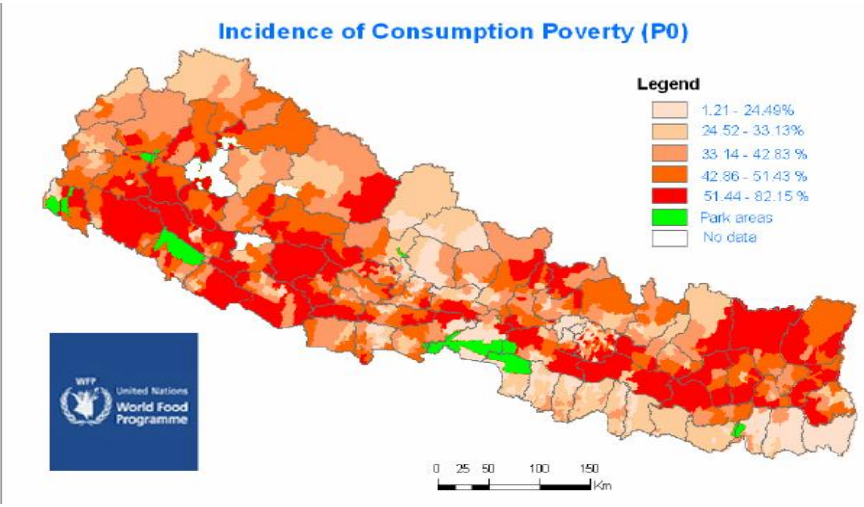


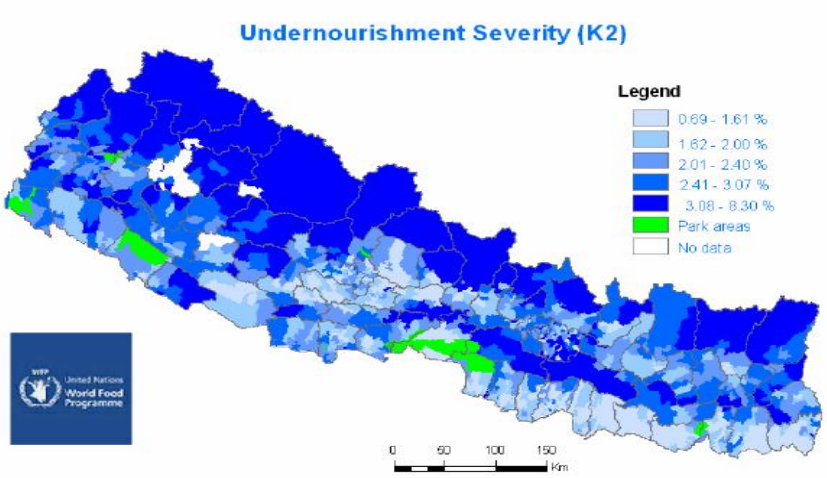
SAE - Philippines

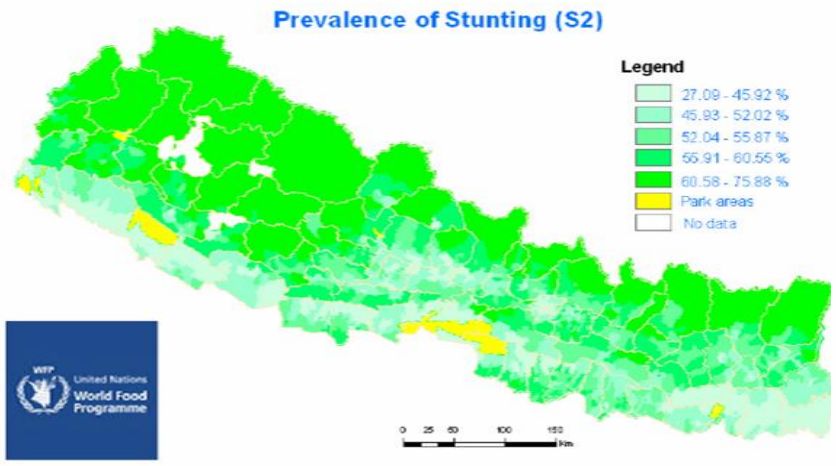


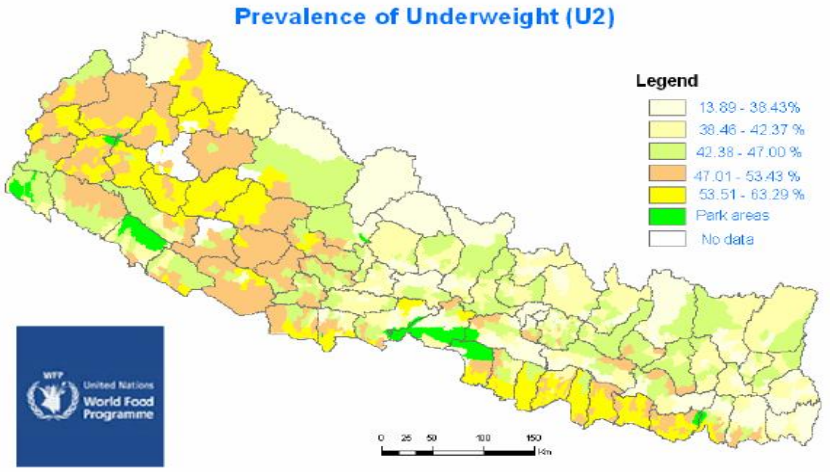


SAE - Nepal

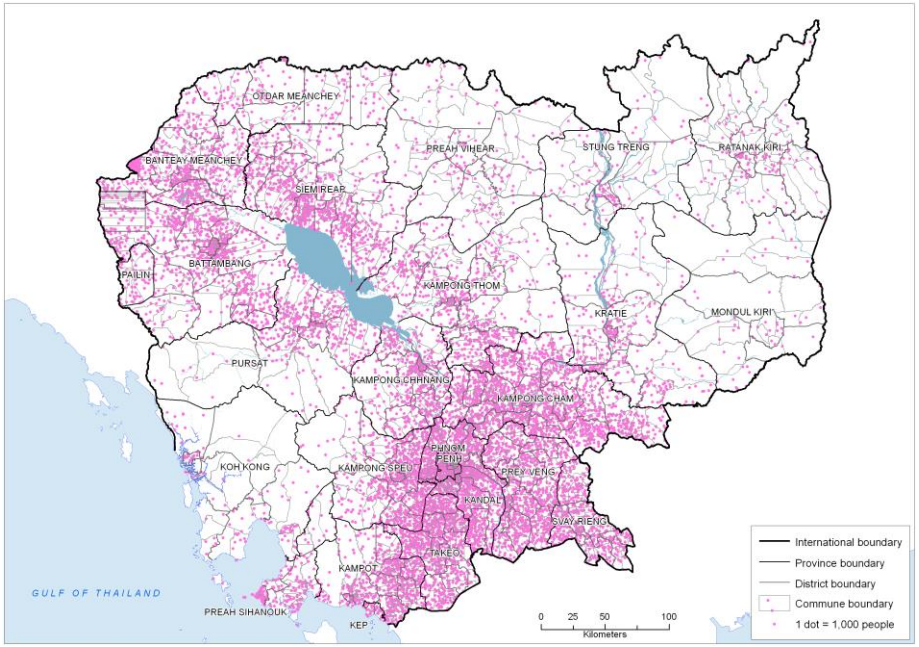

Incidence of Consumption Poverty (P0)
Legend
1.21 - 24.49%
24.52 - 33.13%
33.14 - 42.83 %
42.86 - 51.43 %
51.44 - 82.15 %
Park areas
No data
United Nations World Food Programme
0 25 50 100 150 Km

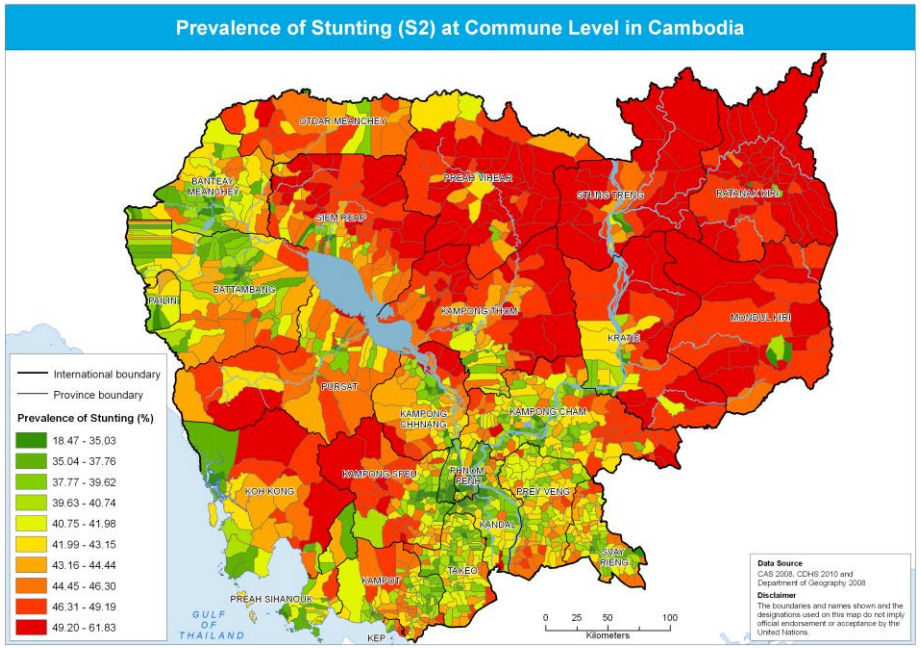

Undernourishment Severity (K2)
Legend
0.69 - 1.61 %
1.62 - 2.00 %
2.01 - 2.40 %
2.41 - 3.07 %
3.08 - 8.30 %
Park areas
No data
United Nations World Food Programme
0 50 100 150 Km


Prevalence of Stunting (S2)
Legend
27.09 - 45.92 %
45.93 - 52.02 %
52.04 - 55.87 %
55.91 - 60.55 %
60.58 - 75.88 %
Park areas
No data
United Nations World Food Programme


Prevalence of Underweight (U2)
Legend
13.89 - 38.43%
38.46 - 42.37 %
42.38 - 47.00 %
47.01 - 53.43 %
53.51 - 63.29 %
Park areas
No data
United Nations World Food Programme
0 25 50 100 150 Km

SAE - Cambodia





Small Area Estimation - Extensions

- Quantile regression
  - Different models are fitted to quantiles of the dependent variable, but note effect of subsetting data.
- Multivariate models
  - Dependent variables are modelled together. Note however Cambodia (where stunting and underweight are positively correlated) c.f. Nepal (stunting and wasting negatively correlated).
- Spatial models
  - Care needed as smoothing spatially can mask model inadequacies.
- Non-linear and generalized linear models
  - Extensions include modelling of proportions via classification trees and regression trees .

References

Ghosh, M., and Rao J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Haslett, S. J., Isidro, M.C. and Jones, G. (2010) Comparison of survey regression techniques in the context of small area estimation of poverty, *Survey Methodology*, 36, 1, 157-170.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics*, 9, 226-252.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Brooks/Cole Publishing Company.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society* B, 60, 23-40.

Pfeffermann, D., Moura, F.A. and Silva, P.L. (2006). Multi-level modelling under informative sampling. *Biometrika*, 93, 949-959.

Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.

Rao, J.N.K. (2003). *Small Area Estimation*, Wiley Series in Survey Methodology. Wiley-Interscience, John Wiley & Sons, Inc.

You, Y., and Rao, J.N.K. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *The Canadian Journal of Statistics*, 30, 431-439.

You, Y., Rao, J.N.K. and Kovačević, M. (2003). Estimating fixed effects and variance components in a random intercept model using survey data. *Proceedings: Symposium 2003, Challenges in Survey Taking for the Next Decade*. Statistics Canada.