

Bayesian Statistics, Small Area Estimation and why no one is poor in Sweden

V. Gómez-Rubio¹, N. Best¹, S. Richardson¹ and P. Clarke²

¹Epidemiology and Public Health, Imperial College London

²Neighbourhood Statistics, Office for National Statistics

Royal Statistical Society Conference
York, 18 July 2007

Outline

- Small Area Estimation
- Example: Average *Equivalised* Income per Household
- Direct Estimation
 - Survey Sampling
- Model-based Estimation
 - Model Comparison and Selection
 - Policy Making and Ranking of areas
- Models for Missing Data
 - Full Data vs. Missing Data
- Summary of results

Small Area Estimation

Objectives

Provide estimates of the variables of interest at different geographical levels.

Data Available

- Official data sets: Census, Labour Force Survey, Health Records
- Aggregated (area level) data (from statistical bureaus such as ONS)
- Surveys conducted *ad hoc*

Statistical Models

- Direct estimators
- Model-assisted estimators
- Model-based estimators

Motivation Example: Av. Equivalised Income per Household

Average *Equivalised* Income per Household (AEIH) in Sweden

Measures the average income *per capita* and takes into account whether the household members are children/adults

LOUISE Population Register in Sweden

Contains a detailed record of every household in the country, including:

- Av. Eq. Income
- Number of persons in hh.
- Head of hh: gender, age, education, employment status

How would we estimate AEIH?

- Conduct survey to record AEIH and related covariates.
- Rely on other information to estimate AEIH: area level data

Direct Estimation

Survey Sampling

- A (significant) sample of the population is taken from areas of interest
- Random sampling without replacement

Direct Estimator

Sample of area i : $\{(y_{ij}, x_{ij}) : j = 1, \dots, n_i\}$

Weights survey design: $w_{ij} = N_i/n_i$

$$\hat{Y}_{D,i} = \frac{\sum_j w_{ij} y_{ij}}{\sum_j w_{ij}} = \frac{\sum_j y_{ij}}{n_i} = \bar{y}_i; \quad \text{var}[\hat{Y}_{D,i}] = (1 - n_i/N_i)S_i^2$$

Problems of Direct Estimation

- Too many areas to estimate
- Sampling becomes very expensive and unfeasible for all areas
- Ignore complex data structure (spatial effects, etc.)

Model-based Estimators

Introduction

- Direct estimator cannot provide estimates in non-sampled areas
- Model-based estimators rely on a fitted model to predict values in non-sampled areas

Main effects

- Covariates (unit/area level)
- Unstructured random effects
- Spatial random effects
- Temporal random effects

Combination of different sources of information

- Survey data
- Area level data (from *official* sources)

Area Level Models

Fay-Herriott Estimator

$$\begin{aligned}\hat{Y}_{D,i} &= \mu_i + e_i \\ e_i &\sim N(0, \hat{\sigma}_{e_i}^2)\end{aligned}$$

$$\begin{aligned}\mu_i &= \alpha + \beta \bar{X}_i + u_i + v_i \\ u_i &\sim N(0, \sigma_u^2)\end{aligned}$$

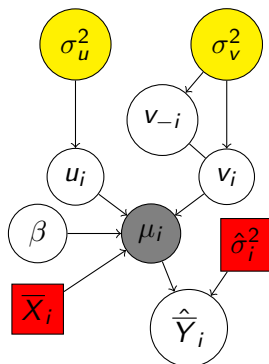
$$v_i | v_{-i} \sim N\left(\sum_{j \in \delta_i} \frac{v_j}{|\delta_i|}, \frac{\sigma_v^2}{|\delta_i|}\right)$$

$$\sigma_u^2, \sigma_v^2 \sim Ga^{-1}(0.001, 0.001)$$

Small Area Estimation

$$\hat{Y}_{A,i} = \hat{\mu}_i$$

Graphical Model



Unit Level Models

Model description

$$y_{ij} = \mu_{ij} + e_{ij}$$

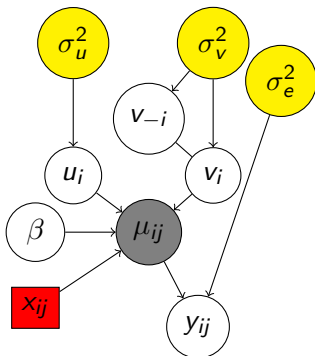
$$e_{ij} \sim N(0, \sigma_e^2)$$

$$\mu_{ij} = \alpha + \beta x_{ij} + u_i + v_i$$

Small Area Estimation

$$\hat{Y}_{u,i} = \hat{\alpha} + \hat{\beta} \bar{X}_i + \hat{u}_i + \hat{v}_i$$

Graphical Model



Unit Level Models

Model description

$$y_{ij} = \mu_{ij} + e_{ij}$$

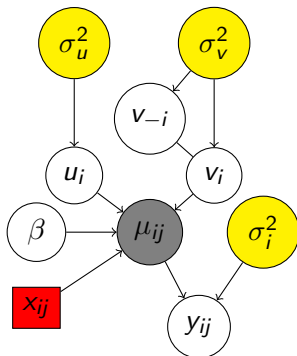
$$e_{ij} \sim N(0, \sigma_*^2)$$

$$\mu_{ij} = \alpha + \beta x_{ij} + u_i + v_i$$

Small Area Estimation

$$\hat{Y}_{u,i} = \hat{\alpha} + \hat{\beta} \bar{X}_i + \hat{u}_i + \hat{v}_i$$

Graphical Model



Unit Level Models

Model description

$$y_{ij} = \mu_{ij} + e_{ij}$$

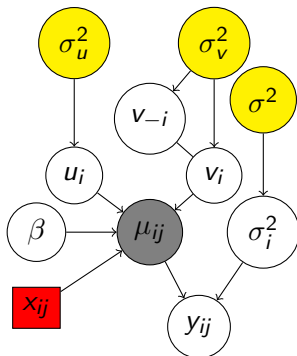
$$e_{ij} \sim N(0, \sigma_*^2)$$

$$\mu_{ij} = \alpha + \beta x_{ij} + u_i + v_i$$

Small Area Estimation

$$\hat{Y}_{u,i} = \hat{\alpha} + \hat{\beta} \bar{X}_i + \hat{u}_i + \hat{v}_i$$

Graphical Model



Average Equivalised Income per Household in Sweden

Data

- 20 different *surveys* from the LOUISE Pop. Reg.
- 284 municipalities in Sweden in 1992
- Sample size: 1% of total number of households
- True area values are known
- Covariates:
 - Number of persons in hh.
 - Head of hh: gender, age, education, employment status

Models compared

- Models with different random effects are compared: u_i , v_i , $u_i + v_i$
- Area and unit levels
- Which areas have the lowest income?

Model Comparison and Model Selection

Average (Relative) Empirical Mean Square Error

$$AEMSE = \sum_{k=1}^{20} \frac{1}{20 \cdot 284} \sum_{i=1}^{284} (\hat{Y}_i^{(k)} - \bar{Y}_i)^2$$

$$AREMSE = \sum_{k=1}^{20} \frac{1}{20 \cdot 284} \sum_{i=1}^{284} \frac{(\hat{Y}_i^{(k)} - \bar{Y}_i)^2}{\bar{Y}_i}$$

Deviance Information Criterion (DIC)

$$DIC = D(\hat{\theta}) + 2p_D$$

Aims

- Select the *best* model in terms of prediction of the area level values
- AMESE is more appropriate but DIC can be computed in practice

Results (Small Area Estimation)

Summary

- Area level models seem to work better (effect of survey design?)
- Model with unstructured (u_i) and spatially correlated (v_i) are better

| | | AEMSE | | | AREMSE | |
|----------|---------|--------------|-----------------|----------|--------|-------|
| | | | Mean | s.d. | Mean | s.d. |
| A. Level | Model | ui | 1949.320 | 189.830 | 1.526 | 0.136 |
| | | vi | 1671.908 | 160.956 | 1.290 | 0.115 |
| | | ui+vi | 1600.953 | 162.346 | 1.250 | 0.119 |
| U. Level | Model 1 | ui | 3649.421 | 1778.944 | 2.970 | 1.445 |
| | | vi | 2871.242 | 1093.657 | 2.350 | 0.905 |
| | | ui+vi | 2824.710 | 1060.653 | 2.311 | 0.878 |
| U. Level | Model 2 | ui | 2960.006 | 269.001 | 2.188 | 0.183 |
| | | vi | 2118.649 | 196.699 | 1.616 | 0.146 |
| | | ui+vi | 2096.845 | 190.188 | 1.590 | 0.141 |
| U. Level | Model 3 | ui | 2959.718 | 268.957 | 2.189 | 0.183 |
| | | vi | 2106.200 | 195.023 | 1.607 | 0.145 |
| | | ui+vi | 2099.994 | 191.782 | 1.593 | 0.142 |

Results (Small Area Estimation)

Summary

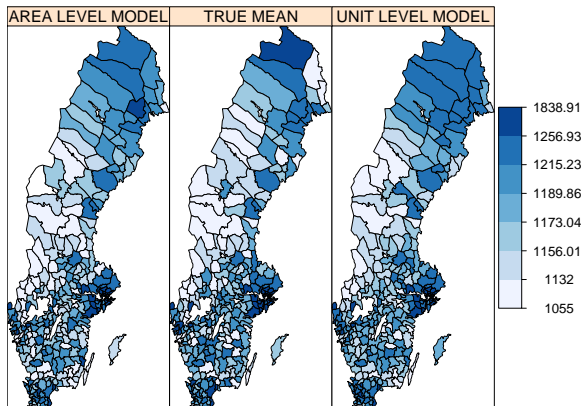
- Area level models seem to work better (effect of survey design?)
- Model with unstructured (u_i) and spatially correlated (v_i) are better

| | Variable | Model with u_i | | Model with v_i | | Model with $u_i + v_i$ | |
|------------|----------|------------------|----------|------------------|----------|------------------------|----------|
| | | Mean | s.d. | Mean | s.d. | Mean | s.d. |
| Area Level | DIC | 3253.15 | 15.58 | 3279.75 | 26.31 | 3230.95 | 18.44 |
| Model | pD | 134.41 | 11.19 | 112.34 | 14.12 | 115.54 | 13.78 |
| Unit Level | DIC | 497847.89 | 30837.81 | 497804.93 | 30850.78 | 497804.48 | 30850.78 |
| Model 1 | pD | 126.16 | 52.36 | 83.29 | 34.57 | 85.46 | 35.31 |
| Unit Level | DIC | 474723.70 | 5063.78 | 474689.21 | 5065.26 | 474683.91 | 5064.01 |
| Model 2 | pD | 463.11 | 7.98 | 417.20 | 10.16 | 427.26 | 11.57 |
| Unit Level | DIC | 474715.34 | 5063.86 | 474678.98 | 5065.28 | 474678.54 | 5063.76 |
| Model 3 | pD | 456.41 | 7.91 | 411.25 | 10.00 | 420.58 | 11.43 |

Results (Small Area Estimation)

Summary

- Area level models seem to work better (effect of survey design?)
- Model with unstructured (u_i) and spatially correlated (v_i) are better



Results (Coefficients of covariates)

Problems

- Aim is on the Small Area Estimation, not focused on the effects of the covariates
- Estimates of the coefficients may differ between area and unit level models

| Covariate | Area Level Model | | Unit Level Model 2 | |
|----------------|------------------|-------|--------------------|-------|
| | Post. Mean | S.D. | Post. Mean | S.D. |
| Interception | 1197.00 | 8.28 | 476.45 | 13.66 |
| Age | 28.60 | 12.57 | 12.07 | 0.28 |
| Education | 43.83 | 11.93 | 190.66 | 8.54 |
| Employment | 46.20 | 17.10 | 261.75 | 6.27 |
| Persons in hh. | -22.71 | 22.45 | -127.71 | 6.27 |
| Gender | -6.67 | 15.97 | 106.62 | 7.35 |

Ranking of areas and Policy Making

Why ranking areas?

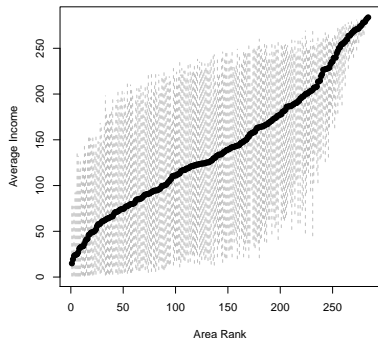
- League tables are useful to compare areas
- Ranking the areas is useful to detect areas that need special attention

How can we rank areas?

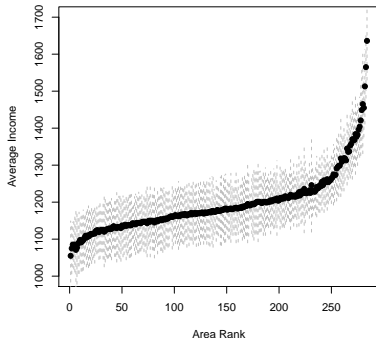
- Estimate of the AEIH
- Relative ranking
- Prob. of being among the 10%,20% areas with the lowest income
- Poverty line (60% national median AEIH: 693.695)

Ranking of areas and Policy Making

Ranking of areas



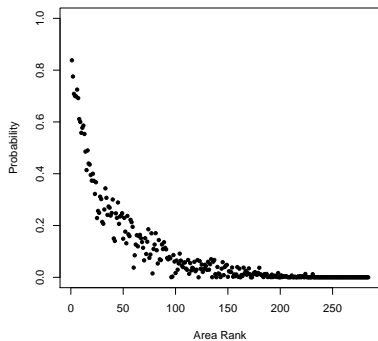
Ranking of areas – Average Income



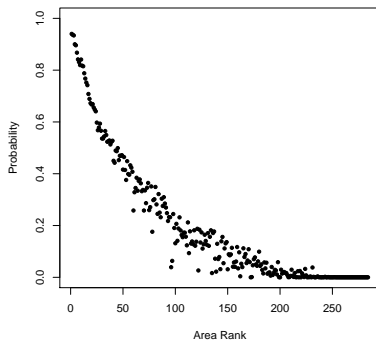
The probability of being above the poverty line is 1 for all municipalities!!

Ranking of areas and Policy Making

Prob. in 10% most deprived areas



Prob. in 20% most deprived areas



Missing Data

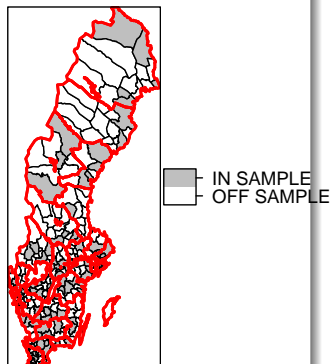
Why do missing data appear?

- Surveys can seldom cover all areas
- Two-stage sampling is often used
- Our missing-data comprises the sample from certain areas

Multiple Imputation

- Area level estimates are obtained by relying on the fitted model and the covariates
- Spatially correlated random effects can be used to borrow information from nearby areas

Primary Sampling Units

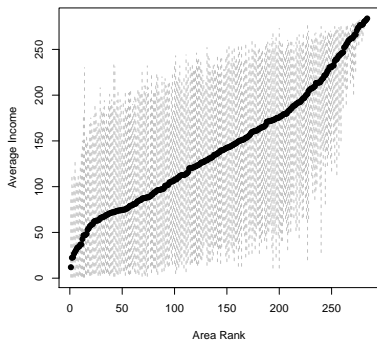


Results (Models with Missing Data)

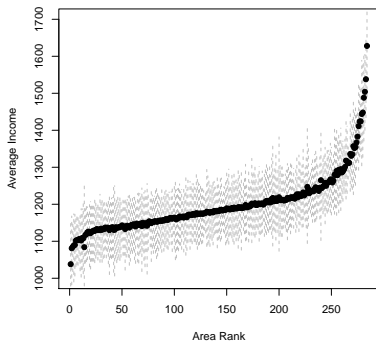
Main Results

- Performance systematically worse than previous models
- However, results are still reliable

Ranking of areas



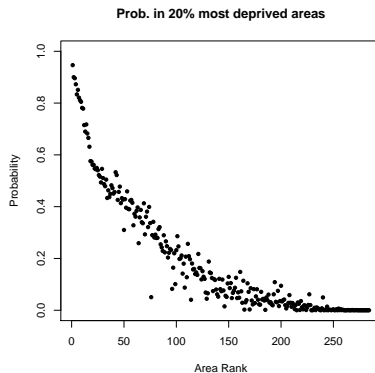
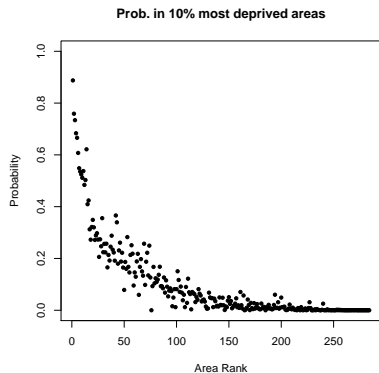
Ranking of areas – Average Income



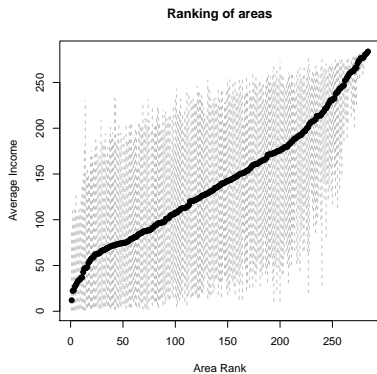
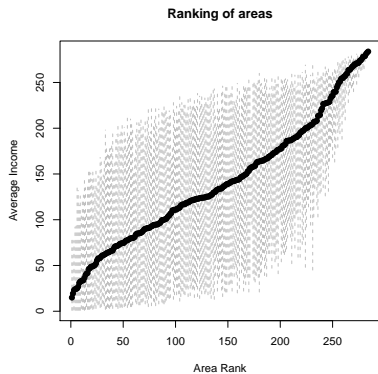
Results (Models with Missing Data)

Main Results

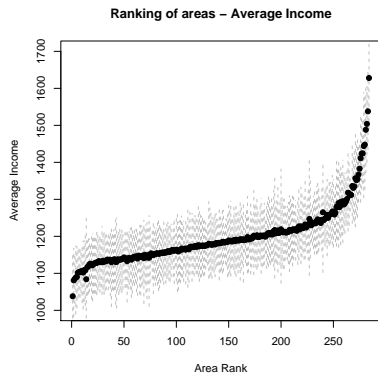
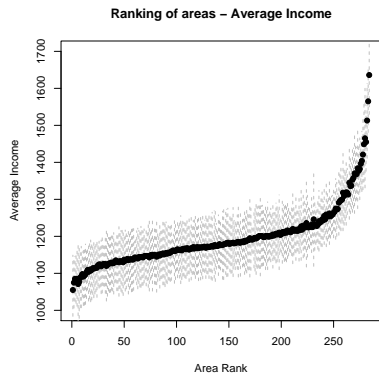
- Performance systematically worse than previous models
- However, results are still reliable



Results (Full Data vs. Missing Data)



Results (Full Data vs. Missing Data)



Summary of results

Small Area Estimation

- SAE can be used efficiently to estimate different variables of interest
- Different types of response variables can be considered

Area or unit level models?

- Area Level Models seem to provide better estimates
- However, when the sample size is very small unit level model perform better

Missing Data

- Missing data occur naturally because of the way data are collected
- Bayesian Inference provides a convenient way of handling missing data
- Spatial correlation can help to improve the results

Future Work

Statistical Models

- Include other types of spatial effects (for example, krigging)
- Include time as well (to improve estimation)
- Consider non-Normal response (unemployment, # persons househ.)
- Alternative priors for variances (for example, half-t)

Model Selection

- How can we compare Unit and Area level models properly?
- *Area level DIC* for unit level models

Policy Making/Policy Assessment

- Alternatives ways of ranking areas
- Reduce uncertainty about the ranking

Acknowledgements

National Centre for Research Methods/ESRC

For funding the BIAS Project

Office for National Statistics

For support and helping to get our hands into the Swedish data

Statistics Sweden

For providing the data of the LOUISE Population Register

References

- BIAS Project. <http://www.bias-project.org.uk>
- EURAREA Consortium (2004). Project reference volume. Technical report, EURAREA Consortium.
- Ghosh, M. and J. N. K. Rao (1994). Small area estimation: An appraisal. *Statistical Science* 9(1), 55–76.
- Goldstein, H. and D. J. Spiegelhalter (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society, Series A* 159(3), 385–443.
- Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, New Jersey.