Bayesian Small Area Estimation for policy making and policy assessment

V. Gómez-Rubio[1], N. Best[1], S. Richardson[1] and P.Clarke[2]

[1]Epidemiology and Public Health, Imperial College London
[2]Office for National Statistics

Research Methods Festival
Oxford, 3 July 2008

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
- Methods for assessing the impact of policies
- Summary of results

Small Area Estimation

Objectives

Provide estimates of the variables of interest at different geographical levels

Data Available

- Official Statistics: 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

Motivating Example

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:

- Average Equivalised Income
- Number of persons in household
- 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: $\{(y_{ij}, x_{ij}) : j = 1, \ldots, n_i\}$
Survey design weights: $w_{ij} = N_i/n_i$

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

Problems of Direct Estimation

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

Model-based Estimators

Motivation

- 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)

Bayesian Hierarchical Models

Introduction

- BHM are Multilevel Models
- All unknown quantities and parameters of the model θ are considered as random variables
- Inference is based on the distribution of θ given the observed data
- Complex models must be fitted using computational procedures (Markov Chain Monte Carlo methods) to obtain a sample from the posterior distribution of θ

Some benefits of Bayesian Inference

- Probability statements about the parameters can be made, i.e., $P(\theta_L < \text{Av. Income} < \theta_U)$.
- Results can be summarised as posterior probabilities: What is the probability of having an income higher than £1000/week?

Area Level Models

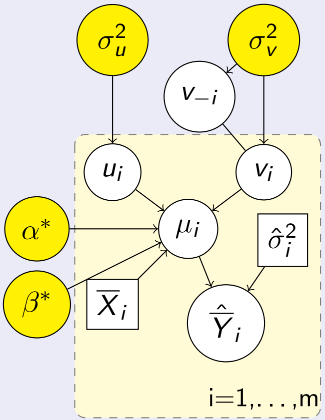
Fay-Herriott Estimator

$$
\begin{aligned}
\hat{\overline{Y}}_{D,i} &= \mu_i + e_i \\
e_i &\sim N(0, \hat{\sigma}^2_{e_i}) \\
\\
\mu_i &= \alpha + \beta \overline{X}_i + u_i + v_i \\
u_i &\sim N(0, \sigma^2_u) \\
v_i | v_{-i} &\sim N(\textstyle\sum_{j \in \delta_i} \frac{v_i}{|\delta_i|}, \frac{\sigma^2_v}{|\delta_i|}) \\
\\
\sigma^2_u, \sigma^2_v &\sim Ga^{-1}(0.001, 0.001)
\end{aligned}
$$

Small Area Estimation

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

Graphical Model



Unit Level Models

Model description

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

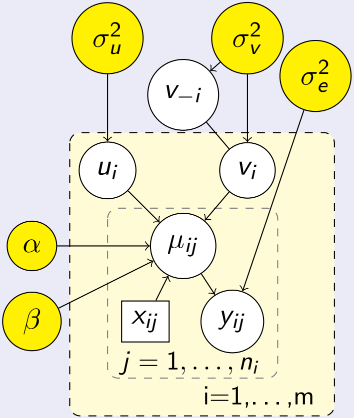
$$e_{ij} \sim N(0, \sigma_e^2)$$
$$\sigma_e^2 \sim Ga^{-1}(0.001, 0.001)$$

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

Small Area Estimation

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

Graphical Model



Unit Level Models

Model description

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

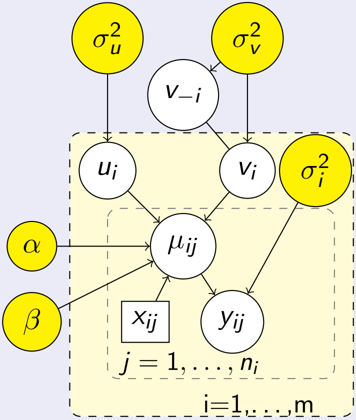
$$e_{ij} \sim N(0, \sigma_i^2)$$
$$\sigma_i^2 \sim Ga^{-1}(0.001, 0.001)$$

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

Small Area Estimation

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

Graphical Model



Unit Level Models

Model description

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

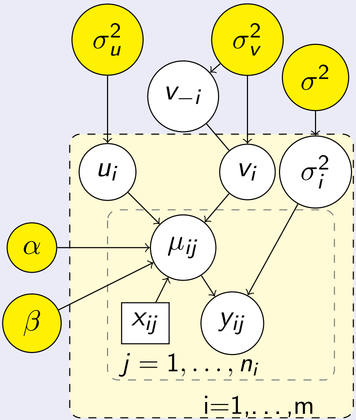
$$e_{ij} \sim N(0, \sigma_i^2)$$
$$\log(\sigma_i^2) \sim N(0, \sigma_i^2)$$

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

Small Area Estimation

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

Graphical Model



Average Equivalised Income per Household in Sweden

Data

- 20 different *surveys* from the LOUISE Population Register
- 284 municipalities in Sweden in 1992
- Sample size: 1% of total number of households
- True area values are known (so can be used for model evaluation)
- 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

Model Comparison and Model Selection

Average (Relative) Empirical Mean Square Error

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

Deviance Information Criterion (DIC)

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

$D(\hat{\theta})$ is the deviance of the model evaluated at the posterior estimates
p_D is the effective number of parameters

Aims

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

Results (Small Area Estimation)

Summary

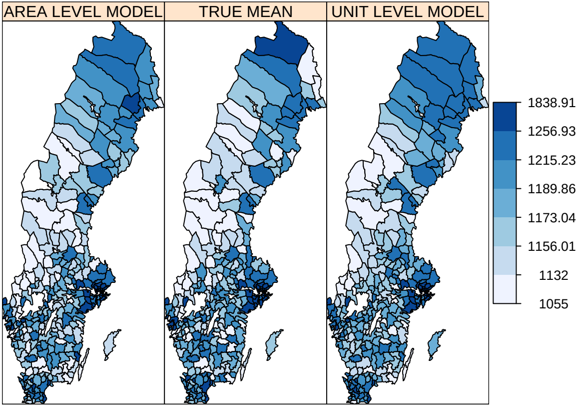
- Area level models seem to work better (effect of survey design?)
- Model with unstructured (u_i) and spatially correlated (v_i) are better

			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)

			DIC	
			Mean	s.d.
A. Level	Model	u_i	3253.15	15.58
		v_i	3279.75	26.31
		$\mathbf{u_i + v_i}$	**3230.95**	18.44
U. Level	Model 1	u_i	497847.89	30837.81
		$\mathbf{v_i}$	**497804.93**	30850.78
		$\mathbf{u_i + v_i}$	**497804.48**	30850.78
U. Level	Model 2	u_i	474723.70	5063.78
		v_i	474689.21	5065.26
		$\mathbf{u_i + v_i}$	**474683.91**	5064.01
U. Level	Model 3	u_i	474715.34	5063.86
		$\mathbf{v_i}$	**474678.98**	5065.28
		$\mathbf{u_i + v_i}$	**474678.54**	5063.76

Results (Small Area Estimation)



Ranking of areas and Policy Making

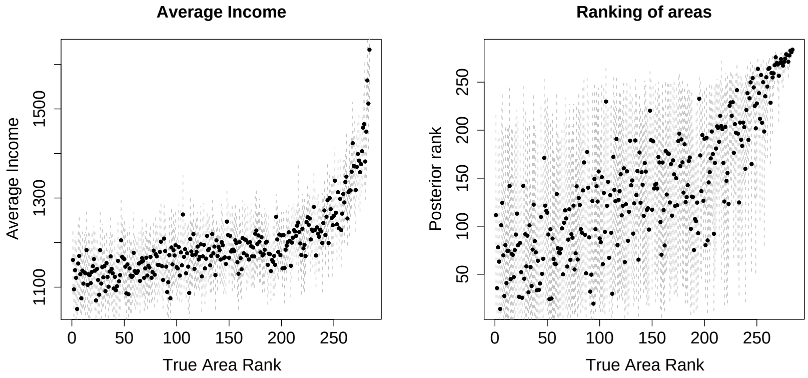
Why rank areas?

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

How can we rank areas?

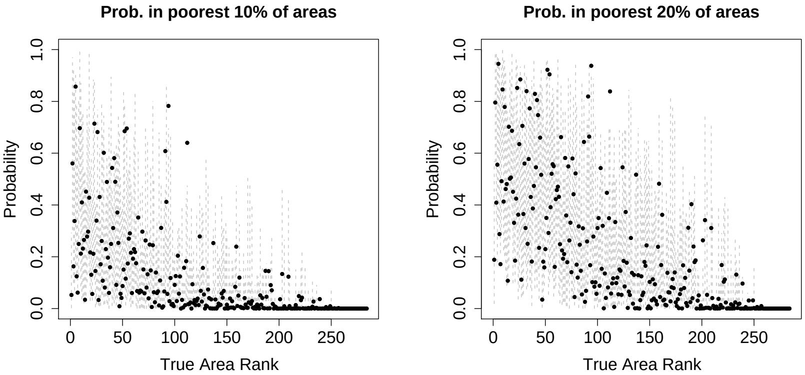
- Rank the point estimate of AEIH
- Relative ranking
- Prob. of being among the 10%,20% areas with the lowest income
- Poverty line (60% national median AEIH: 693.695)

Ranking of areas and Policy Making



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

Ranking of areas and Policy Making



The intervals are **sampling intervals** that measure the variation of the posterior probabilities for 20 different survey data.

Missing Data

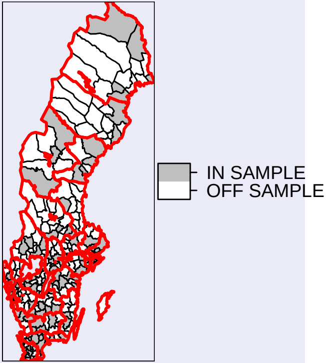
Why do missing data appear?

- Surveys can seldom cover all areas
- Two-stage sampling is often used
- Our observed data comprises the sample from a few 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

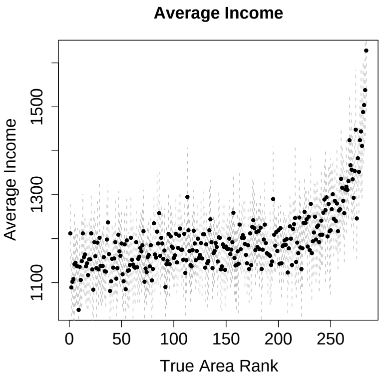
Primary Sampling Units

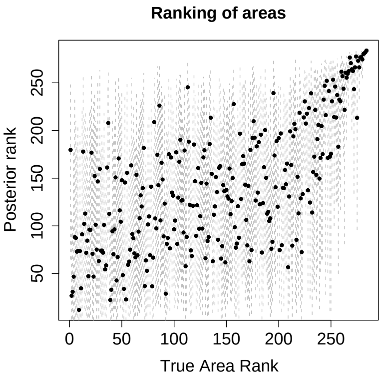


Results (Models with Missing Data)

Main Results

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

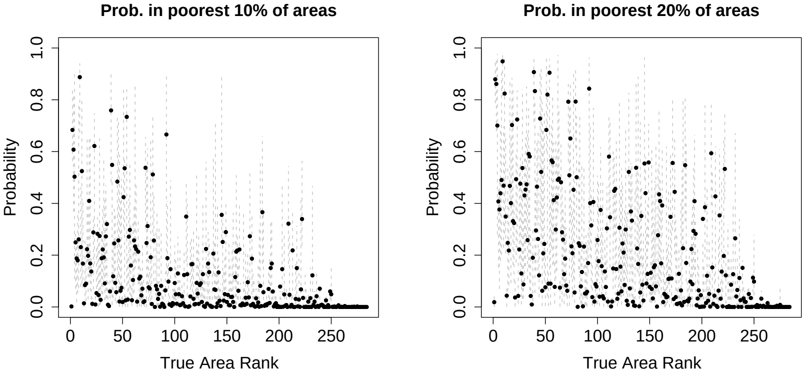




Results (Models with Missing Data)

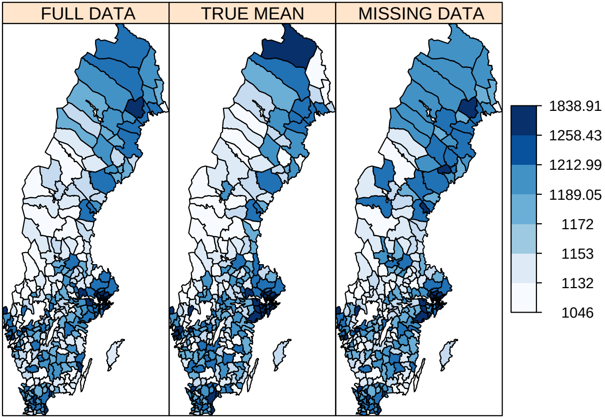
Main Results

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



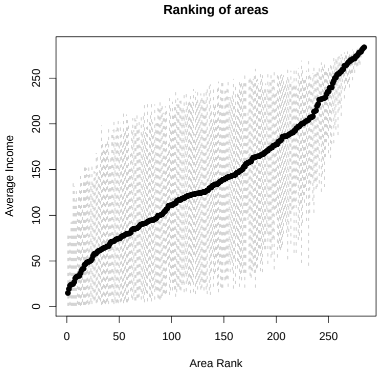
Results (Full Data vs. Missing Data)

Results of area level models with both random effects

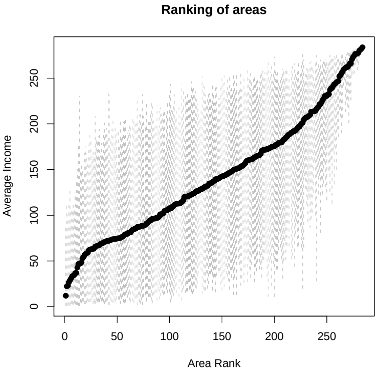


Results (Full Data vs. Missing Data)

Full data



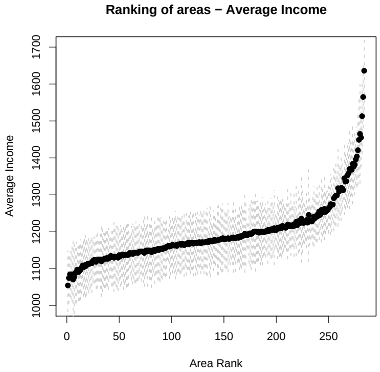
Missing data



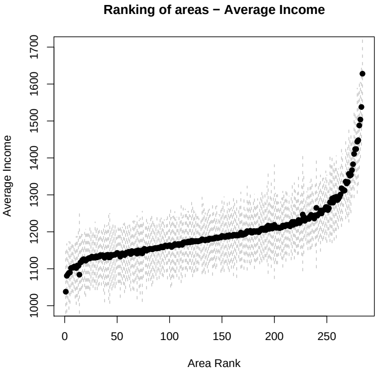
Ranking is now based on the posterior ranks of the model with full data in both plots to make comparisons easier

Results (Full Data vs. Missing Data)

Full data



Missing data



Ranking is now based on the posterior ranks of the model with full data in both plots to make comparisons easier

Family Resources Survey

Survey description

- The survey covers England and Wales
- Carried out in 2001
- Includes a number of socioeconomic covariates
- Primary sampling unit: Postcode level
- Level of interest: Local Authority Districts

Average Income per Household

- Response: Income per household
- Covariates: 25 socio-economic covariates (LAD level)
- Spatial models developed at LAD level

FRS: Results

Main results

- Several unit level models have been compared
- Best model has been chosen according to the DIC:

Unit Model 3	DIC	p_D
$\mathbf{u_i}$	**51494.900**	363.760
v_i	51502.100	353.597
$u_i + v_i$	51502.200	377.413

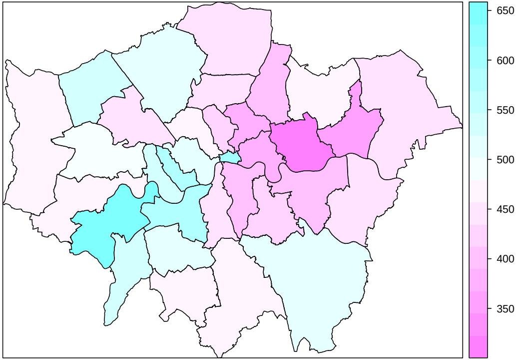
- The *best* model is unit model 3 with non-spatial random effects

Aims of the study

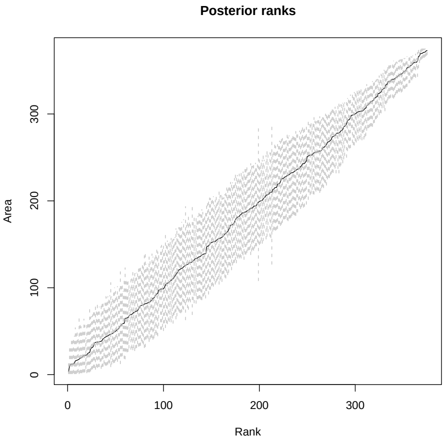
- Provide estimates of the average income per household at LAD level
- Rank areas according to income
- Provide maps of the small area estimates

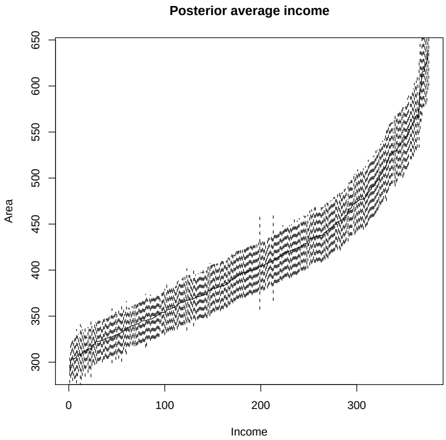
FRS: Results

Average Income per Household in London

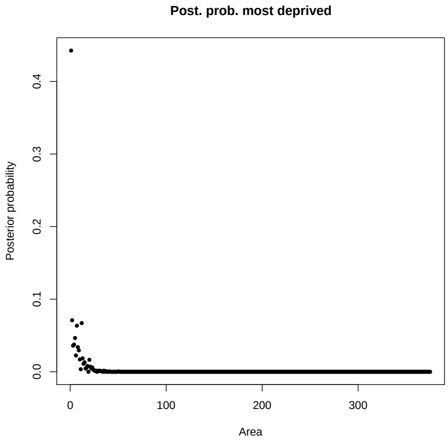


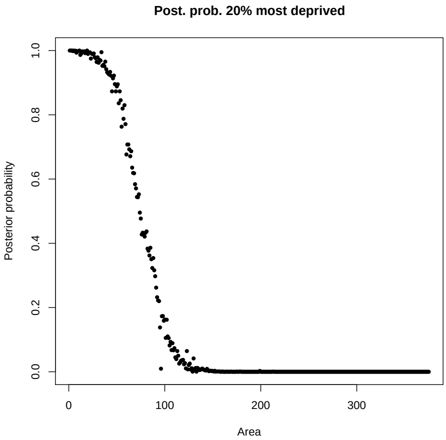
FRS: Results


Posterior ranks
Area
Rank
0
100
200
300


Posterior average income
Area
Income
300
350
400
450
500
550
600
650
0
100
200
300

FRS: Results


Post. prob. most deprived
Posterior probability
0.0
0.1
0.2
0.3
0.4
0
100
200
300
Area


Post. prob. 20% most deprived
Posterior probability
0.0
0.2
0.4
0.6
0.8
1.0
0
100
200
300
Area

Methods for policy assessment

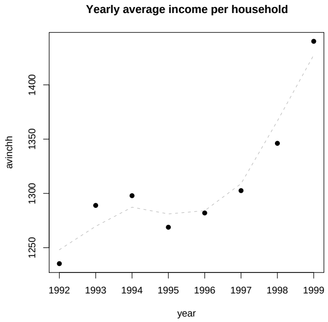
Motivation

- How can we know if a policy had a positive impact?
- Did the areas affected by the policy suffer any change over time?
- If we have data prior to the implementation of the policy and the following years then it is possible to measure the effect of the policy.
- In addition to policy assessment, we may be able to monitor abrupt changes in time
- It may be difficult to detect the origin of the change

Methods

- Space-time models
- Time: We want to model the overall temporal trend and changes
- Space: We still need to account for the variability between areas

Some ideas...



Statistical methods

- Non-parametric smoothing of the global temporal trend and look for abrupt changes
- Compare predicted trend (using *pre-policy* data) to observed data
- Use methods to find change-point in time

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 time as well (to improve estimation)
- Consider non-Normal response (unemployment, # persons househ.)

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
- Follow-up of specific areas to identify changes in time

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.