Approximate Bayesian Inference for Small Area Estimation

V. Gómez-Rubio

Dep. of Mathematics, U. of Castilla-La Mancha, Spain
Dep. of Epidemiology and Public Health, Imperial College London

Trondheim, 14 May 2009

Joint work with Nicky Best, Sylvia Richardson, Philip Li and Philip Clarke (ONS)

Outline

- Small Area Estimation
- Example: Average Income per Household
- Direct Estimation
  - Survey Design
- Model-based estimation
  - Model selection
  - Classification, ranking and policy making
- Models with missing observations
- How can INLA be used for SAE?
- Discussion

Small Area Estimation

Aims

Provide estimates of the target variables at different administative levels.

Data

- Official statistics: Census, Family Resources Survy, Cancer Registers, etc.
- Aggretate Data (at different levels) can be obtained from National Statistics Bureaus
- *Ad hoc* surveys

Statistical Models

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

Example

Average Income per Household (AIH) in Sweden

Average income *per capita* accounting for the number of adults and children in the household

LOUISE Population Register in Sweden

Detailed register of every household in Sweden:

- Income
- Number of persons in the household
- Head of household: gender, age, education level, employment status

How can AIH be estimated?

- Survey data to measure AIH and other covariates of interest
- Use additional information to estimate the AIH: aggregate data

Direct Estimation

Survey

- Significant sample of the population of interest
- Simple random sampling without replacement (but there are others...)

Direct Estimator

Sample from area i: $\{(y_{ij}, x_{ij}) : j = 1, \ldots, n_i\}$
Survey 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 from all areas is too expensive
- Ignores complexity of the data (spatial effects, etc.)

Model-based estimators

Motivation

- Direct Estimator cannot be used in areas with no data
- Model-based estimators are based on a model that can be used to predict the target variables in the areas with no data

Main effects

- Covarites (individual and area levels)
- Random effects
- Spatial random effects
- Temporal random effects

Combining different sources

- Sample
- Aggregate data (from official sources)

Bayesian Hierarchical Models

Introduction

- BHM are multilevel models
- All unknown quantities and parameters of interest θ of the model are considered as random variables
- Inference is based on the posterior distribution of θ given the observed data
- Complex models can be fitted using simulation techniques (Markov Chain Monte Carlo) or approximate methods (INLA!) to obtain an approximation to the posterior distribution of θ

Some benefits of Bayesian Inference

- Probability estatements about the parameters of the model can be made: $P(\theta_L < \mathrm{IMH} < \theta_U)$.
- Results can be summarised as posterior probabilities: Probability of having an income higher than 500EUR/week.

Area level data

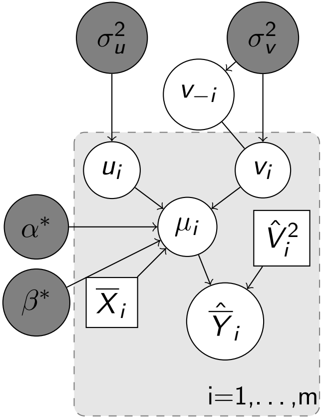
Fay-Herriott Estimator

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

Small Area Estimation

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

Graphical Model



Unit level models

Model

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

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

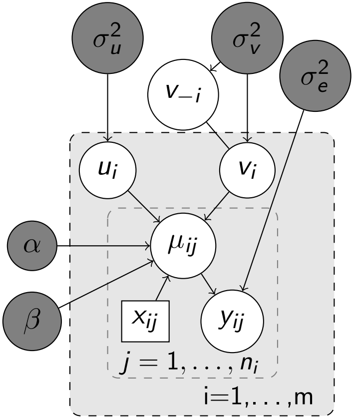
$$\sigma^2_e \sim Ga^{-1}(0.001, 0.001)$$

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

Small Area Estimation

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

Modelo Gráfico



Unit level models

Model

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

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

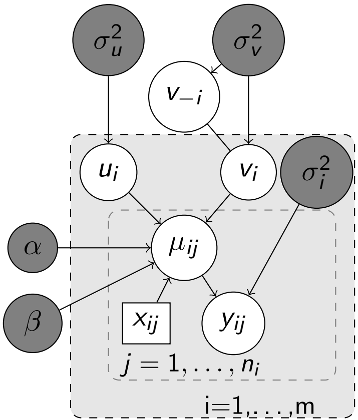
$$\sigma_i^2 \sim Ga^{-1}(0.001, 0.001)$$

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

Small Area Estimation

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

Modelo Gráfico



Unit level models

Model

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

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

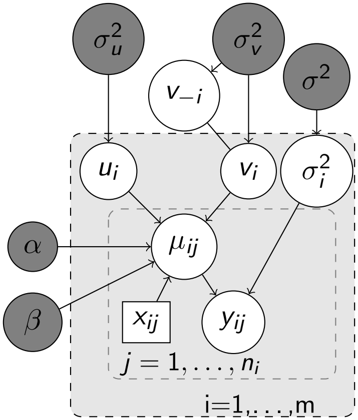
$$\log(\sigma_i^2) \sim N(0, \sigma^2)$$

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

Small Area Estimation

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

Modelo Gráfico



Average Income per Household in Sweden

Data

- Different *surveys* from the LOUISE Register
- 284 municipalities in Sweden in 1992
- Sample size: 1% of total number of households
- Actual values are known (and they can be use to validate the models)
- Covariates:
  - Number of people in household
  - Head of household: gender, age, education level, employment

Models compared

- Different models have been compared: u_i, v_i, $u_i + v_i$
- Area and unit level models

Model comparisson

Average (Relative) Empirical Mean Square Error

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

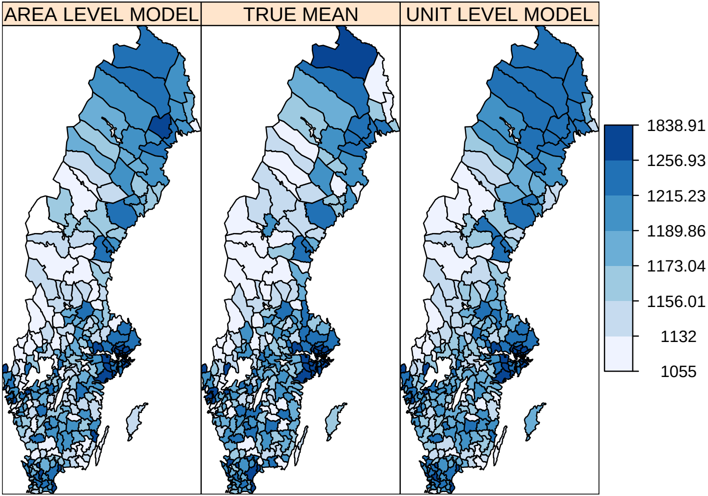
Deviance Information Criterion (DIC)

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

Aims

- Select the *best* model in terms of prediction of the values in the Small Areas
- AEMSE is more appropriate but in practice we can only compute the DIC

Small Area Estimates



Classification of areas for policy making

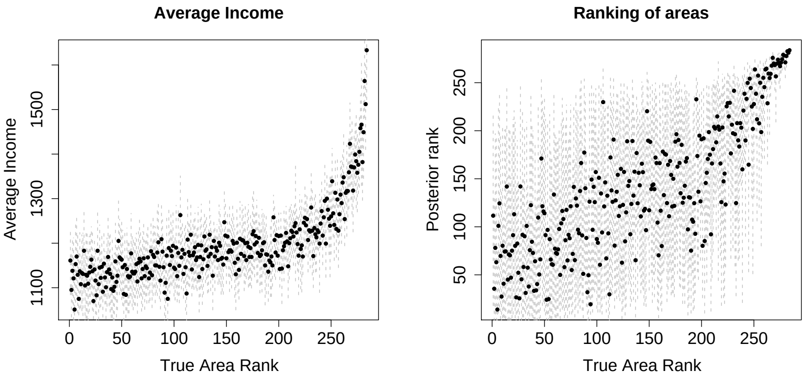
Why rank areas?

- To compare them
- Detect areas with special needs (i.e., high unemployment, low income, etc.)

How can we rank areas?

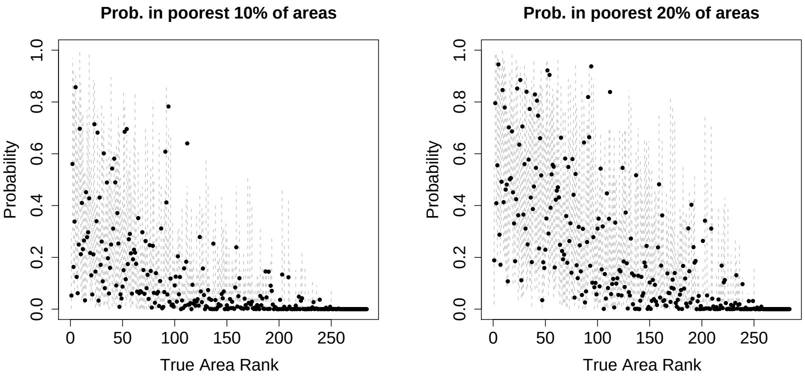
- Point estimates (i.e., posterior means)
- Ranks
- Posterior probabilities
- Poverty line (60% of national average income)

Classification with real data (Area level models)



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

Classification with real data (Area level models)



Intervals are **sampling intervals** which show sample-to-sample variation of the posterior probabilities.

Using INLA for Small Area Estimation

Why?

- Many reasons but to tell you the truth...
- We submitted these results for publication
- The referees asked to increase the number of samples from 20 to 100
- More than six months later we are still running some of the models!!!

Other reasons

- Statistical offices and policy makers need to provide results in a reasonable time
- Area level models are usually fast, but Unit level models usually take longer, especially if the sample size is large
- Random effects models take even longer
- Exploiting the full posterior is usually very expensive with MCMC (for example, for spatial prediction and ranking)

General problems in Small Area Estimation

- Provide Small Area estimates from the marginals
- Produce ranking of the areas
- Deal with designs that include (many) areas with no survey data (other data may be available)
- Triple-goal estimation: SA estimates, histograms and ranking
- Benchmarking and raking: producing SA estimates that are consistent when aggregated over higher administrative levels
- Useful for poverty mapping, i.e., estimate the proportion or number of households below the poverty line.

Exploiting the marginal distributions

Motivation

- INLA provides an approximation to the marginal distribution of several parameters and quantities
- Where is the limit when we make inference with the marginals only?

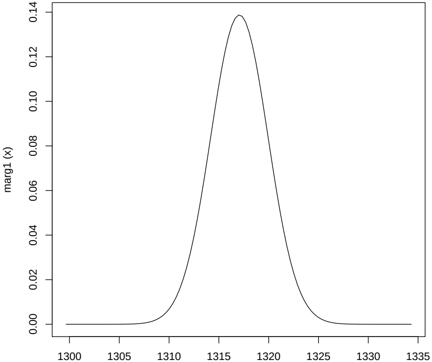
First step: Operations with the marginals

- Fit spline to the marginal: `inla.spline`, `fitmarginalsp`
- Distribution function: `dmarginal`
- Sample from the marginal: `rmarginal`
- Compute probabilities: `pmarginal`
- Compute quantiles: `qmarginal`

Marginal distribution

```
xy<-res$marginals.linear.predictor[[1]]
marg1<-fitmarginalsp(xy)

curve(marg1, from=min(xy[,1]), to=max(xy[,1]) )
```



Computing probabilities

Compute probability of average income being higher than 1350

```
a<-1350
inlaprob<-lapply(res$marginals.linear.predictor, function(X){
        marg<-fitmarginalsp(X)
        1-pmarginal(a, marg, range=range(X[,1]))
})
inlaprob<-unlist(inlaprob)
inlaprob[1:10]
     index.1      index.2      index.3      index.4      index
7.429086e-03 9.540118e-06 8.508139e-05 3.263719e-05 1.149532e-
     index.7      index.8      index.9     index.10
2.992667e-02 4.010670e-02 3.621850e-03 8.379177e-03
```

Computing quantiles

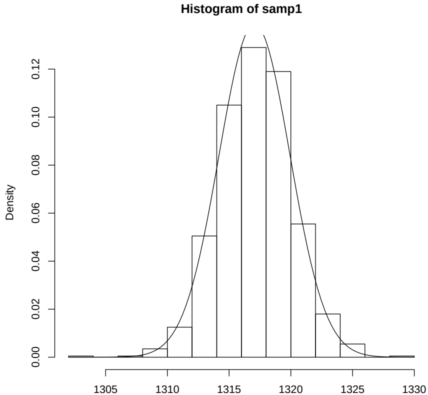
```
inlaq<-lapply(res$marginals.linear.predictor, function(X){
        marg<-fitmarginalsp(X)
        a<-qmarginal(.025, marg, range=range(X[,1]))
        b<-qmarginal(.975, marg, range=range(X[,1]))
        c(a,b)
})
```

```
#Summary statistics from INLA
283 1323.658 2.427987   1318.889   1328.419 0.000000e+00
284 1314.117 2.546993   1309.115   1319.112 7.395571e-32
#R code
> inlaq[283:284]
$index.283
[1] 1318.899 1328.416
$index.284
[1] 1309.125 1319.109
```

Sampling from the marginals

```
samp1<-rmarginal(1000, marg1, range=range(xy[,1]) )

hist(samp1, freq=FALSE)
curve(marg1, add=TRUE)
```



Ranking of areas

What can be done with INLA (so far)?

- Use SA estimates $\hat{\mu}_i$ to establish a ranking of the areas
- Compute posterior probablities: $P[\mu_i > baseline]$

What cannot be done with INLA?

- Compute probability of being the most deprived area
- Compute probability of being among the q% more deprived areas
- Compute posterior distribution of the area ranks
- Can we make use of the marginals to simulate from the full posterior?
  - Ranks based on repeated samling from $\pi(\mu|y) = \Pi_i \pi(\mu_i|y)$ will not work

Areas with no data

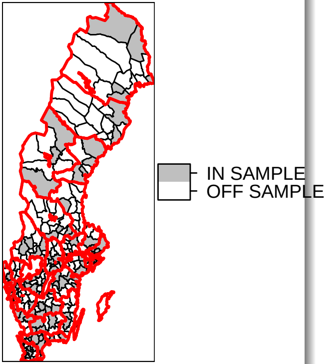
Why do they appear?

- Sampling seldom covers all areas
- Two-stage sampling
- Observed data only cover a few areas

Estimation in off-sample areas

- Small Area estimates are obtained by means of the fitted model and the area level covariates
- Spatially structured random effects can be used to borrow information from neighbouring areas

Áreas en el muestreo



INLA and missing data

- INLA can handle missing observations in the response
- This is fine with area level models because:
  - What we usually lack is the area level direct estimator
  - We have covariates that can be used in the preduction
  - Can INLA borrow information from neighbouring areas when spatially structured random effects are used?

Unit level models

- More complex issues: non-response, etc.
- Sometimes we have detailed covariates on the households, so we can just predict the income using the marginals of the model parameters
- Usually we have indivual data from a survey and aggregate covariates and we need to combine both data sources

Triple-goal estimation

Motivation

When producing Small Area estimates we may want to obtain good estimates, in the following sense:

- They are no over-shrunk (i.e., too much towards the average value of the SA estimates)
- Good histogram, i.e., the distribution function of the estimates is similar to that of the ensemble of μ_i's
- Good ranks (useful to detect areas with extremes μ_i)

See Rao (2003, pages 211-214) and Shen and Louis (1998) for details

Constrained estimators

Setting

We want a new set of estimates $\{t_i\}$ by minimising the posterior expected squared loss $E[\sum_i (\mu_i - t_i)^2 | y]$ subject to

$$\text{Match mean:} \quad t_. = \frac{1}{K} \sum_i t_i = \hat{\mu}_. = \frac{1}{K} \sum_i \hat{\mu}_i$$

$$\text{Match variance:} \quad \frac{1}{K-1} \sum_i (t_i - t_.)^2 = E[\frac{1}{K-1} \sum_i (\mu_i - \mu_.)^2 | y]$$

$$\hat{t}_i = \hat{\mu}_. + a(\hat{\mu}, \lambda)(\hat{\mu}_i - \hat{\mu}_.)$$

$$a(\hat{\mu}, \lambda) = \left[1 + \frac{(1/K) \sum_i V(\mu_i | y, \lambda)}{\{(1/(K-1)\} \sum_i (\hat{\mu}_i - \hat{\mu}_.)^2} \right]^{1/2}$$

Histogram

Motivation

The empirical distribution function on $\hat{t}_i$'s is

$$\tilde{F}_m(t) = m^{-1} \sum_i I(\hat{t}_i \leq t)$$

but this is a poor estimator of $F_m(t) = m^{-1} \sum_i I(\mu_i \leq t)$
An optimal estimator $A(t)$ is obtained by minimising the posterior expected integrated squared error loss $E[\int \{A(t) - F(m)\}^2 dt | y]$.

If $A(t)$ is constrained to be discrete, the optimal estimator $\hat{F}_m(t)$ is discrete with mass $1/K$ at

$$\hat{U}_l = \overline{F}_m^{-1}((2l-1)/(2m)) \text{ where } \overline{F}(t) = \frac{1}{m} \sum_i P(\theta_i \leq t | y)$$

Ranks

Motivation

- How good are the ranks based on $\hat{t}_i$ compared to those based on μ_i?
- $\hat{t}_i$ ranks are identical to those based on $\hat{\mu}_i$
- The true rank of area i is $R(i) = \sum_l I(\mu_i \geq m_l)$

An optimal estimator $Q(i)$ can be obtained by minisming the expected posterior squared error loss $E[\sum_i (Q(i) - R(i))^2 | y]$

$$Q_{opt}(i) = \tilde{R}(i) = E[R(i)|y] = \sum_l P(\mu_i \geq \mu_l | y)$$

In general, $\tilde{R}(i)$ are not integers, so we rank them to obtain $\hat{R}(i)$.

Shen and Louis' triple-goal estimators

$$\hat{\mu}_i^{TG} = \hat{U}_{\hat{R}(i)}$$

Summary and Discussion

Small Area Estimation

- The marginals are usually most of what we need in SAE
- INLA provides a suitable framework to obtain good estimates in a short time
- Wide range of models can be used: area level, unit level, spatial, spatio-temporal, etc.
- Marginals could be further exploited to provide some methods for ranking
- However, important problems like estimate the posterior ranks and provide triple goal cannot be tackled (?)
- Is there any way of approximating $P(\mu_i \geq \mu_l | y)$ to produce triple-goal estimates? For example, with $P(\mu_i \geq \hat{\mu}_l | y)$