Small Area Estimation with R

Unit 4: EBLUP estimators

V. Gómez-Rubio

Department of Mathematics
Universidad de Castilla-La Mancha, Spain

Mixed-effects models for SAE

Mixed-effects models

- Random effects can help in the fitting of the model by accounting for different types of hidden structures
- They can be thought of as if they measured the effects of unobserved covariates

Small Area Estimation

- Random effects at the area level
- Spatial correlation between areas
- Temporal correlation

Area level model with random effects

$$\hat{\overline{Y}}_i = \mu_i + \epsilon_i;\ \ \epsilon_i \sim N(0, \hat{\sigma}_i^2)$$

$$\mu_i = \beta_0 + \beta_1 \overline{X}_i + Z u_i;\ \ u_i \sim N(0, \sigma_u^2);\ \ i = 1, \ldots, K$$

- u_i are the random effects
- σ_u^2 is the variance of the random effects
- Z reflects the structure of the random effects
- They are specified at the area level
- This model can be fitted because $\hat{\sigma}_i^2$ is known
- Estimation can be done either by Maximum Likelihood (ML) or Restricted ML (REML)

Model structure and estimation

$$\hat{\bar{Y}}_i \sim N(\mu, V)$$

$$\mu = (\mu_1, \ldots, \mu_K)^T = X^T\beta \qquad V = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2) + \sigma_u^2 Z I_K Z^T$$

β can be estimated by

$$\hat{\beta} = (X^T V^{-1} X)^{-1}) X^T V^{-1} (\hat{\bar{Y}} - X^T\beta)$$

σ_u^2 is usually not know and it is plugged-in the previous equation
The small area estimate becomes

$$\hat{\bar{Y}}_{M,i} = \hat{\beta} X_i + \hat{u}_i$$

Estimation of the random effects

EBLUP estimator

- The random effects can be estimated in different ways
- A common way of estimating them is by means of (Empirical) Best Linear Unbiased Predictors (EBLUP)
- EBLUP estimator are a linear combination of the observed values:

$$\hat{u}_i = GZ^T V^{-1}(\hat{\bar{Y}} - X\beta)$$

- V in the unconditional variance of the response
- G is the variance of the random effects
- In addition, EBLUP estimators are taken so that they minimise the Mean Square Error

Estimation of the MSE of the estimates

When computing the variance of the estimates we have to take into account the following:

$$MSE[\hat{\bar{Y}}_{M,i}] \approx G1 + G2 + 2 \cdot G3$$

- Uncertainty about the small area estimate (G1)
- Uncertainty about $\hat{\beta}$ (G2)
- Uncertainty about $\hat{\sigma_u^2}$ (G3)
- Other terms may appear if the variance components have more parameters (like in the case of the Spatial EBLUP)

Package **nlme**

- Linear mixed-effects models: `lme`
- Generalized mixed-effects models: `nlme`
- Linear models with complex variance structure: `gls`

- The structure of the random effects can be defined by
  - the covariance structure
  - correlation between the units
  - Different structures can be defined and combined

Package **SAE**

- Implements EBLUP estimators for area level models
- Under development!!
- Area level model with independent random effects
- Area level model with spatially correlated random effectos
- Provides estimates of the MSE of the small area estimates
  - Spatial EBLUP
  - The spatial random effects have a SAR structure (that I will discuss later)
- Includes a vignette with examples on how to use different methods using a simulated data set

Example: Area level model

```
> library(SAE2)
> spam.options(eps=.0000001)
> dmm<-cbind(data.frame(REG=1:8, DIREST=destdom$RMT85,
+    DESVAR=destdom$se^2), REGCOV)
> dmmeblup<-EBLUP(DIREST~ME84, ~DESVAR, data=dmm )

> dmmeblup

Call:
EBLUP(formula = DIREST ~ ME84, varformula = ~DESVAR, data = dmm)

Coefficients:
            [,1]
[1,] 81.27455635
[2,]  0.05871661

Variance of the random effects: 4857.405

Log likelihood: -264.1687
```

Unit level

$$y_{ij} = \mu_{ij} + \varepsilon_{ij};\ \varepsilon_{ij} \sim N(0, \sigma^2)$$

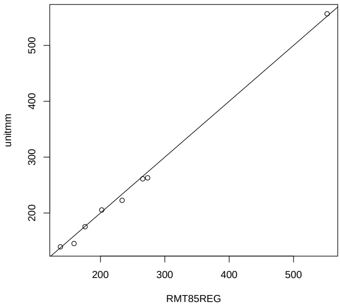
$$\mu_{ij} = \beta_0 + \beta_1 x_{ij} + u_i;\ u_i \sim N(0, \sigma_u^2)$$

Small area estimates are provided by

$$\hat{\overline{Y}}_{m,i} = \hat{\beta} X_i + \hat{u}_i$$

where $\hat{\beta}$ and $\hat{u}_i$ are computed similarly as for the area level case. Note that now matrices tend to be significantlly larger.

Example: Unit level mixed-effects model

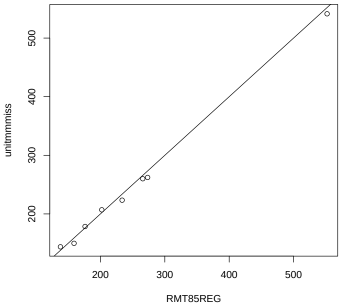



```
> library(nlme)
> mm<-lme(RMT85~ME84, random=~1|REG, data=dsmpcl)
> unitmm<-predict(mm, dmm)

Linear mixed-effects model fit by REML
  Data: dsmpcl
  Log-restricted-likelihood: -145.3564
  Fixed: RMT85 ~ ME84
(Intercept)        ME84
  -6.372505    0.138095

Random effects:
 Formula: ~1 | REG
        (Intercept) Residual
StdDev: 0.001222797 20.33021

Number of Observations: 32
Number of Groups: 8
```

Models with missing data

- Mixed-effects model can still be fitted when there are data from some areas missing by desing
- The fitting procedure is done as before, but with the reduced data set
- Small area estimates are produced as before
- However, $\hat{u}_i = 0$ for the areas with no sample at all. Hence, in this areas the EBLUP estimator is reduced to a synthetic estimator

Models with missing data




```
> mmmiss<-lme(RMT85~ME84, random=~1|REG, data=dsmpcl2)
> unitmmmiss<-predict(mmmiss, dmm)
> unitmmmiss[is.na(unitmmmiss)]<-predict(mmmiss, dmm, lev

Linear mixed-effects model fit by REML
  Data: dsmpcl2
  Log-restricted-likelihood: -144.2672
  Fixed: RMT85 ~ ME84
(Intercept)        ME84
  5.2739567   0.1315343

Random effects:
 Formula: ~1 | REG
        (Intercept) Residual
StdDev:    1.542383 20.58033

Number of Observations: 32
Number of Groups: 4
```

Spatial models

Motivation

- Sometimes the data exhibit spatial autocorrelation that should be modelled in some way
- Covariates which show spatial structure can be used
- Any remaining spatial structure can be modelled by means of random effects
- The structure of the random effects should mimic the spatial configuration of the data
- The spatial random effects can follow different structures, but in general they are based on the adjacency between the areas. For example, neighbouring regions should have a higher correlation than regions that are further apart
- The spatial random effects are modelled at the area level

Spatial structures

Pratesi & Salvati (2008) developed a Spatial EBLUP (SEBLUP) estimator based on a SAR specification for the random effects

Simultaneously Autoregressive Models (SAR)

Recalling the area level model

$$\hat{\overline{Y}}_i = \mu_i + \epsilon_i;\ \ \epsilon_i \sim N(0, \hat{\sigma}_i^2)$$

If we model the area level mean as

$$\mu_i = \alpha + \beta \overline{X}_i + v_i$$

the SAR specification is as follows

$$v = \rho W v + u; u \sim MVN(0, \sigma^2 I)$$

$$v = (I - \rho W)^{-1} u;\ \ v \sim N(0, \sigma^2 [(I - \rho W)(I - \rho W))]^{-1}$$

Example: Spatial EBLUP

```
> library(spdep)
> regnb<-poly2nb(swreg)
> W<-nb2mat(regnb, style="W")
>

> dmmseblup<-SEBLUP(DIREST~ME84, ~DESVAR, data=dmm , W=W, method="REML")
> dmmseblup

Call:
SEBLUP(formula = DIREST ~ ME84, varformula = ~DESVAR, data = dmm,
    W = W, method = "REML")

Coefficients:
            [,1]
[1,] 17.59511680
[2,]  0.05898785

Variance of the random effects: 9937.627

Log likelihood: -271.6073
```

Spatial approach and Data 'missing' by design

- LeSage and Pace (2004) and Saei and Chambers (2005) discuss the issue of using spatial correlation to improve the estimation in small areas with 'missing' data.
- When the random effects are independent of each other, the estimate of the random effect in off-sample areas is 0
- Given that spatial random effects are correlated, EBLUP estimates of the random effects can be predicted in off-sample

Other model-based estimators

- Temporal models
  The structure of the random effects can be set so that every time period has an associated random effect.
- Space-time models
  Space and time random effects can be combined to produce space-time models

References

- LeSage, J. P. and R. K. Pace (2004). Models for spatially dependent missing data. *Journal of Real Estate Finance and Economics* **29**(2), 233–254.
- Pratesi, M. and N. Salvati (2008). Small Area Estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applciations* **17**: 113–141.
- Saei, A. and R. Chambers (2005). Working paper m05/03: Empirical best linear unbiased prediction for out of sample areas. Technical report, Southampton Statistical Sciences Research Institute, University of Southampton.