Analysing Spatial Data in R
Worked examples: Small Area Estimation

Virgilio Gómez-Rubio

Department of Epidemiology and Public Heath
Imperial College London
London, UK

31 August 2007

Small Area Estimation

- Small Area Estimation provides a general framework for investigating the spatial distribution of variables at different administrative levels
- Disease Mapping is a particular case of Small Area Estimation
- Very important for governemt agencies and statistical bureaus
- Lehtonen and Pahkinen describe different direct and regression-based estimators and provide trainning materials on-line
- Rao (2003) provides a complete summary of different methods for SAE.

How do we get the data?

Statistical offices

- Different types of small area data
- Public release as yearly reports, books, atlas, etc.
- Aggregated data (usualy)
- Individual data might be available (on request)

Survey data

- Provide accurate information at individual level (person, houshold, ...)
- Difficult to obtain from public sources
- *Ad-hoc* surveys can be carried and linked to aggregated public data
- Some way of combining individual and aggregated data

Overview of R packages for SAE

- `sampling`: Sampling methods for complex surveys
- `survey`: Analysis of data from complex surveys
- `glm`: Generalised Linear Models
- `nlme`: Mixed-effect models
- `SAE`: Some EBLUP estimators for Small Area Estimation
- `spsurvey`: Spatial survey design and analysis

The MSU284 Population

The MSU284 Population (Särndal et al., 2003) describes the 284 municipalities of Sweden. It is included in package `sampling`.

- LABEL. Identifier.
- P85. Population in 1985
- RMT85. Revenues from the 1985 municipal taxation
- ME84. Number of Municipal Employees in 1984
- REG. Geographic region indicator (8 regions)
- CL. *Cluster* indicator (50 clusters)

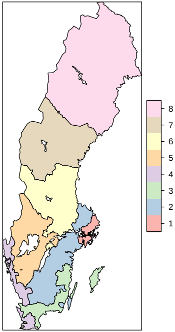
```
> library(sampling)
> data(MU284)
> MU284 <- MU284[order(MU284$REG), ]
> MU284$LABEL <- 1:284
> summary(MU284)
```

Basics of Survey Design

- Surveys are used to obtain representative data on all the population in the study region
- Ideally, the survey data would contain a small sample for each area
- In practice, surveys are clustered to reduce costs (for example, *two-stage* sampling)
- Define *sampling frame*
- Example: General Houshold Survey 2000 (ONS)
  - Primary Sampling Units (PSUs): Postcode
  - Secondary Sampling Units (SSUs): Household
- Outcome is $\{(x_{ij}, y_{ij}), j \in s_i; i = 1, \ldots, K\}$
  - y_{ij} target variable
  - x_{ij} covariates

Regions in Sweden

- Municipalities in Sweden can be grouped into 8 regions
- We will treat the municipalities as the *units*
- To estimate the regional mean we will sample from the municipalities



Survey sampling with R

Simple Random Sampling Without Replacement

- Sample is made of 32 municipalities ($\sim$11% sample)
- Equal probabilities for all municipalities

```
> N <- 284
> n <- 32
> nreg <- length(unique(MU284$REG))
> set.seed(1)
> smp <- srswor(n, N)
> dsmp <- MU284[smp == 1, ]
> table(dsmp$REG)

1 2 3 4 5 6 7 8
2 5 6 3 7 3 2 4
```

Survey sampling with R

Stratified SRS Without Replacement

- Sample is made of 32 municipalities ($\sim$11% sample)
- 4 municipalities sampled per region
- Equal probabilities for all municipalities **within** strata

```
> set.seed(1)
> smpcl <- mstage(MU284, stage = list("cluster", "cluster"),
+     varnames = list("REG", "LABEL"), size = list(8, rep(4,
+         8)), method = "srswor")
> dsmpcl <- MU284[smpcl[[2]]$LABEL, ]
> table(dsmpcl$REG)

1 2 3 4 5 6 7 8
4 4 4 4 4 4 4 4
```

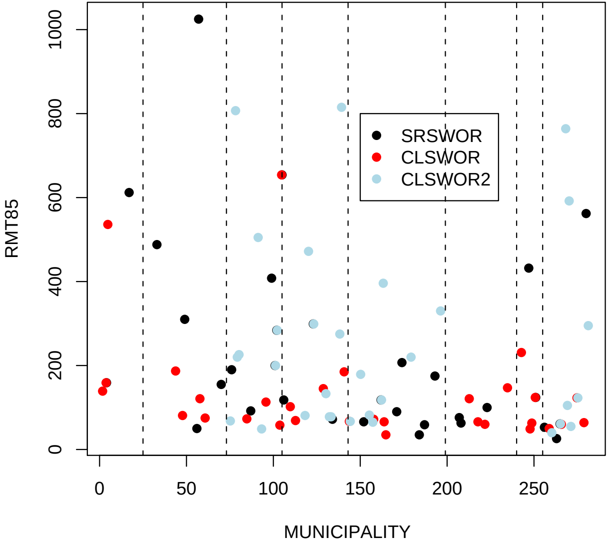
Survey sampling with R

Stratified SRS Without Replacement (Two-Stage Sampling)

- Sample is made of 32 municipalities (∼11% sample)
- 8 municipalities sampled per region
- Equal probabilities for all municipalities **within** strata
- Some regions do not contribute to the survey sample

```
> set.seed(1)
> smpcl2 <- mstage(MU284, stage = list("cluster", "cluster"),
+     varnames = list("REG", "LABEL"), size = list(4, rep(8,
+         8)), method = "srswor")
> dsmpcl2 <- MU284[smpcl2[[2]]$LABEL, ]
> table(dsmpcl2$REG)

3 4 5 8
8 8 8 8
```

Survey sampling with R



Small Area Estimators

Sample-based Estimators

Based on the survey data

- *Direct Estimator*
- *GREG Estimator*

Indirect Estimators

Based on survey data and some appropriate model

- (Generalised) *Linear Regression*
- Mixed-Effects Models
- EBLUP Estimation
- Models with Spatially Correlated Effects

Direct Estimation

- Direct estimators rely on the survey sample to provide small area estimates
- Not appropriate if there are out-of-sample areas

Horvitz-Thomson estimator:

$$\hat{Y}_{direct} = \sum_{i \in s} \frac{1}{\pi_i} y_i \qquad \hat{\bar{Y}}_{direct} = \sum_{i \in s} \frac{\frac{1}{\pi_i} y_i}{\sum_{i \in s} \frac{1}{\pi_i}}$$

For SRS without replacement: $\pi_i = \frac{n}{N}$

```
> library(survey)
> RMT85 <- sum(MU284$RMT85)
> RMT85REG <- as.numeric(by(MU284$RMT85, MU284$REG, sum))
```

Direct Estimation

- Direct estimators rely on the survey sample to provide small area estimates
- Not appropriate if there are out-of-sample areas

$$Y_{direct} = \sum_{i \in s} \frac{1}{\pi_i} y_i$$

For SRS without replacement: $\pi_{ij} = \frac{n_i}{N_i}$

```
> library(survey)
> svy <- svydesign(~1, data = dsmp, fpc = rep(284, n))
> dest <- svytotal(~RMT85, svy)
```

Direct Estimation

A domain refers to a subpopulation of the area of interest
In the example, we may estimate the revenues for each region

$$Y_{direct,i} = \sum_{j \in s_i} \frac{1}{\pi_{ij}} y_{ij}$$

```
> fpc <- lreg[dsmpcl$REG]
> svycl <- svydesign(id = ~1, strata = ~REG, data = dsmpcl,
+     fpc = fpc)
> destcl <- svytotal(~RMT85, svycl)
```

Direct Estimation

A domain refers to a subpopulation of the area of interest
In the example, we may estimate the revenues for each region

$$Y_{direct,i} = \sum_{j \in s_i} \frac{1}{\pi_{ij}} y_{ij}$$

```
> fpc2 <- lreg[dsmpcl2$REG]
> svycl2 <- svydesign(id = ~1, strata = ~REG, data = dsmpcl2,
+     fpc = fpc2)
> destcl2 <- svytotal(~RMT85, svycl2)
```

Direct Estimation of Domains

A domain refers to a subpopulation of the area of interest
In the example, we may estimate the revenues for each region

$$Y_{direct,i} = \sum_{j \in s_i} \frac{1}{\pi_{ij}} y_{ij}$$

```
> svyby(~RMT85, ~REG, svy, svytotal)
```

```
  REG statistics.RMT85 se.RMT85
1   1         6842.625 5244.545
2   2        17998.500 9620.438
3   3        16223.500 6874.105
4   4         4339.875 2699.869
5   5         6656.250 2505.059
6   6         2121.125 1138.299
7   7         4934.500 3725.099
8   8         6230.250 4711.205
```

Direct Estimation of Domains

A domain refers to a subpopulation of the area of interest
In the example, we may estimate the revenues for each region

$$Y_{direct,i} = \sum_{j \in s_i} \frac{1}{\pi_{ij}} y_{ij}$$

```
> svyby(~RMT85, ~REG, svycl, svytotal)

  REG statistics.RMT85   se.RMT85
1   1         44356.25 34347.1708
2   2          5568.00  1184.5134
3   3          7184.00  4299.5057
4   4          4759.50   908.4262
5   5          3360.00   455.2333
6   6          4038.50   825.9968
7   7          1751.25   532.0153
8   8          2153.25   444.6669
```

Direct Estimation of Domains

A domain refers to a subpopulation of the area of interest
In the example, we may estimate the revenues for each region

$$Y_{direct,i} = \sum_{j \in s_i} \frac{1}{\pi_{ij}} y_{ij}$$

```
> svyby(~RMT85, ~REG, svycl2, svytotal)

  REG statistics.RMT85 se.RMT85
3   3         9436.000 2450.388
4   4        10597.250 3080.939
5   5        10199.000 2299.526
8   8         7376.875 2418.904
```

Generalised Regression Estimator

Definition

- Model-assisted estimator
- Relies on survey design and (linear) regression
- It can be expressed as a direct estimator plus some correction term based on additional information (covariates)

$$\hat{Y}_{GREG} = \sum_{j \in s} \frac{1}{\pi_j} y_j + \sum_k \beta_k \left(\sum_{p=1}^{N} x_p - \sum_{j \in s} \frac{1}{\pi_j} x_j \right)$$

$$\hat{Y}_{GREG,i} = \sum_{j \in s_i} \frac{1}{\pi_{ij}} y_{ij} + \sum_k \beta_k \left(\sum_{p=1}^{N_i} x_p - \sum_{j \in s_i} \frac{1}{\pi_{ij}} x_{ij} \right)$$

Coefficients β_k are estimated using weigthed linear regression.

GREG Estimation with R

```
> pop.totals = c("(Intercept)" = N, ME84 = sum(MU284$ME84))
> svygreg <- calibrate(svy, ~ME84, calfun = "linear", population = pop.
> svytotal(~RMT85, svygreg)

      total     SE
RMT85 67473 1217.2

> svygregcl <- calibrate(svycl, ~ME84, calfun = "linear",
+      population = pop.totals)
> svytotal(~RMT85, svygregcl)

      total     SE
RMT85 68170 873.04

> svygregcl2 <- calibrate(svycl2, ~ME84, calfun = "linear",
+      population = pop.totals)
> svytotal(~RMT85, svygregcl2)

      total     SE
RMT85 68387 914.81
```

Linear Regression

- `lm` assumes that the sample comes from an *infinite* population
- `svyglm` accounts for the survey design and provides a correction for *finite population* in the estimation of the standard errors

We are trying to model the total tax revenues according to the number of municipal employees

```
> plot(MU284$ME84, MU284$RMT85)
> plot(MU284$ME84, MU284$RMT85, xlim = c(0, 10000))
> survlm <- lm(RMT85 ~ ME84, dsmp)
> survglm <- svyglm(RMT85 ~ ME84, svy)
> summary(survlm)
> summary(survglm)
```

Mixed-effects models and EBLUP estimators

- Mixed-effects models can be used to improve estimation
- Random Effects measure variation due to unmesared factors
- Spatial patterns can be accounted for by means of random effects

Fay-Herriot Area Level Model

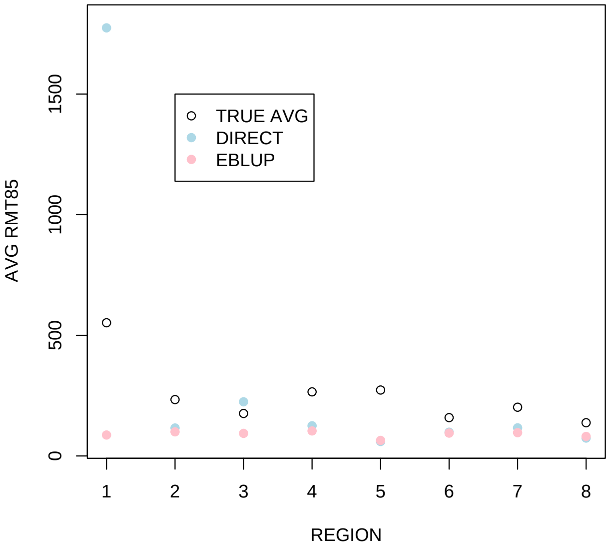
$$\begin{array}{rclcrcl} \hat{\overline{Y}}_i & = & \mu_i + e_i & \qquad & e_i & \sim & N(0, \hat{\sigma}_i^2) \\ \mu_i & = & \beta X_i + u_i & \qquad & u_i & \sim & N(0, \sigma_u^2) \end{array}$$

- $\hat{\overline{Y}}_i$ is often a direct estimator
- $\hat{\sigma}_i^2$ is the variance of the direct estimator
- $\hat{\mu}_i$ is a new (improved) small area estimator
- $\hat{u}_i$ are estimated using EBLUP estimators

EBLUP estimators with R

```
> library(SAE)
> destmean <- svyby(~RMT85, ~REG, svycl, svymean)
> Y <- matrix(destmean[, 2], ncol = 1)
> sigma2i <- matrix(destmean[, 3], ncol = 1)^2
> X <- matrix(as.numeric(by(MU284$ME84, MU284$REG, mean)),
+     ncol = 1)
> ebluparea <- EBLUP.area(Y, cbind(1, X), sigma2i, 8)
> print(sum((destmean[, 2] - (RMT85REG/lreg))^2))

[1] 1590108

> print(sum((ebluparea$EBLUP - (RMT85REG/lreg))^2))

[1] 329263.7

> print(ebluparea$randeff[, 1])

[1]   0.3319200   9.6791711   2.6907938  13.8812442 -25.4537694
[6]   3.4234902   5.9494749 -10.5023248
```

EBLUP estimators with R



Spatial EBLUP estimators

- The random effects can be used to model spatial dependence
- There are different approaches to model spatial dependence
- Petrucci and Salvati (2006) propose a Spatial EBLUP estimator based in a SAR specification

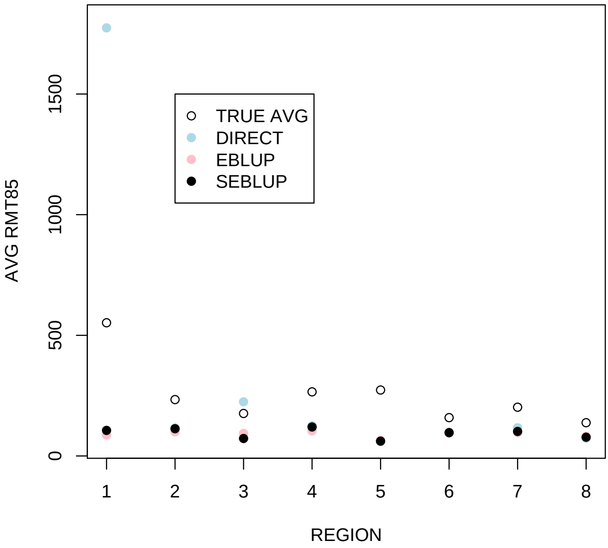
$$\begin{array}{rclrcl} \hat{\bar{Y}}_i & = & \mu_i + e_i & e_i & \sim & N(0, \hat{\sigma}_i^2) \\ \mu_i & = & \beta X_i + v_i & v & \sim & N(0, \sigma_u^2[(I - \rho W)(I - \rho W^T)]^{-1}) \end{array}$$

- ρ measures spatial correlation
- W is a *proximity* matrix which can be defined in different ways

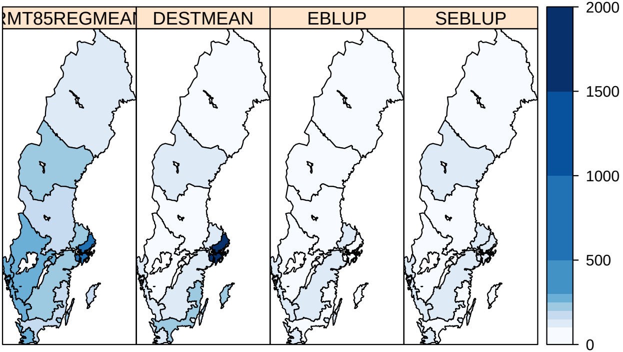
Spatial EBLUP estimators with R

```
> moran.test(Y, nb2listw(nb), alternative = "two.sided")

Moran's I test under randomisation


data:  Y
weights: nb2listw(nb)


Moran I statistic standard deviate = 1.1501, p-value = 0.2501
alternative hypothesis: two.sided
sample estimates:
Moran I statistic       Expectation          Variance
      -0.02635814       -0.14285714        0.01026137

> sebluparea <- SEBLUP.area(Y, matrix(cbind(1, X), ncol = 2),
+      sigma2i, 8, W, init = c(0, ebluparea$sigma2u))
> print(paste("Rho:", sebluparea$rho, "s.d.", sqrt(sebluparea$varsigmar
+      2]), sep = " ", collapse = " "))

[1] "Rho: -0.402461548158343 s.d. 0.120181628230132"

> print(sebluparea$randeff[, 1])

[1]  -9.097686  18.450828 -19.126460  23.199879 -35.424211   6.951748
[7]   8.234322 -11.566655
```

EBLUP estimators with R



Mapping the results



Assessment of the Estimators

$$AEMSE = \frac{1}{K}\sum_{i=1}^{K}(\hat{Y}_i - Y_i)^2$$

Estimation of the National Mean

Estimator	sqrt(AEMSE)
Direct (SRS)	4258.4
Direct (CL)	3565.8
Direct (CL2)	31996

Estimation in Domains

Estimator	sqrt(AEMSE)
Direct (CL)	157.62
EBLUP	71.727
SEBLUP	69.355

References and other sources

- Additional documentation for `survey` package: `http://faculty.washington.edu/tlumley/survey/`
- Practical Exemplars and Survey Analysis (ESRC/NCRM): `http://www.napier.ac.uk/depts/fhls/peas/`
- A. Petrucci and N. Salvati (2006). Small Area Estimation for Spatial Correlation in Watershed Erosion Assessment. *Journal of Agricultural, Biological & Environmental Statistics* **11** (2): 169-182.
- J.N.K. Rao (2003). *Small Area Estimation*. John Wiley & Sons, Inc.
- C.E. Särndall, B. Swensson and J. Wretman (2003). *Model Assisted Survey Sampling*. Springer-Verlag.