

Introduction to Small Area Estimation

SAE package developers

22nd January 2007

1 Introduction

In this vignette we will describe an example on how to produce Small Area Estimates using different types of techniques. Different direct and model based estimators will be briefly described and their computation using the Rsoftware will be illustrated with a simulated data set.

Small Area Estimation tackles the problem of providing reliable estimates of one or several variables of interest in areas where the information available on those variables is, on its own, not sufficient to provide a valid estimate. The information is usually collected by conducting a survey in some or all areas. The survey may involve the collection of information from the areas themselves or some of the individuals living in those areas, whose data are later used to provide area-based estimates.

Direct estimators provide estimates based only on the local data (i.e., the data collected from the area itself) assuming that the sample is large enough, which seldom happens in practice. This problem is usually overcome by borrowing strength from other areas, usually neighbours or observations in the same area recorded at different times. Hence, model based estimators can be used to share information between different areas.

In what follows, we will use \bar{Y}_i and \bar{X}_i to denote the area-level means of the target variable and covariate, respectively, and y_{ij} and x_{ij} to denote individual level values for subject j sampled from area i .

2 Survey design

The study described in this vignette is based on simulated data using code by N. Salvati. Nevertheless, we would like to provide a description of the survey design that we have used, for which we have tried to follow the guidelines provided in Särndal et al. [1992, Section 1.2].

The *population* is made of all the individuals in the 42 small areas that we are considering and from which we will take a sample. Hence, in this example the population is the *domain* or *target population* (or subpopulation) and for each individual we have a series of characteristics or *variables*. Some may be known a priori, but we will assume that the values of the variable of interest are only available for the sampled individuals. In our example, we have one variable of interest (Y) and a covariate (X). The latter is known for all areas (area level models) or individuals (unit level models), while the former is only available for the sampled individuals.

In this case the *sampling frame* is the list of the complete population in all those areas. Since we have simulated the data we have data for every individual in the population. Data are stored in a PC and accessed through the Rsoftware.

The *sample* has been taken choosing a sample size (at random, from a uniform between 3 and 15) for each area. The average sample size is 7.79. The individuals within each area have been sampled WITH replacement. Hence the probability of sampling of an individual is independent between areas and the same within areas (and equal to the inverse of the population size in area i).

2.1 Simulated data

The covariate is simulated taking values uniformly from an interval which is slightly different for each small area. For each individual, the variable of interest is computed as the sum of these three terms:

- **Fixed term** based on the covariate ($\beta = 0.21$)
- **Random term based on the area level effect.** These effects have been simulated using a SAR specification with variance $\sigma_u^2 = 100$ and $\rho = 0.85$. The spatial structure of the regions is based on the contiguity matrix shown in Figure 1. Note that region 25 has no neighbours at all.
- **Random term based on the individual variation.** It is the product of a normal random variable with mean 0 and variance 1.34, multiplied by the square root of the value of the covariate. This ensures that the variance of the individual effects is different for each individual.

3 Summary statistics of the simulated data set

The total number of individuals is 40236 and the mean number of individuals is 239.5 per region. Table 1 summarises different statistics computed for each region.

- Ymean: true mean of the variable Y in the region.
- Xmean: true mean of the covariate X in the region.
- N_i : number of individuals in the region.
- n_i : number of individuals in the sample.
- YSmean: mean of variable Y using only the sample.
- XSmean: mean of covariate X using only the sample.
- sigma2e: true with-in area variance of the variable Y , computed for each region separately.
- sigma2eS: with-in area variance of Y for the the sampled units.

Note that S is used above to refer to those values of the target variable or covariate in the sample.

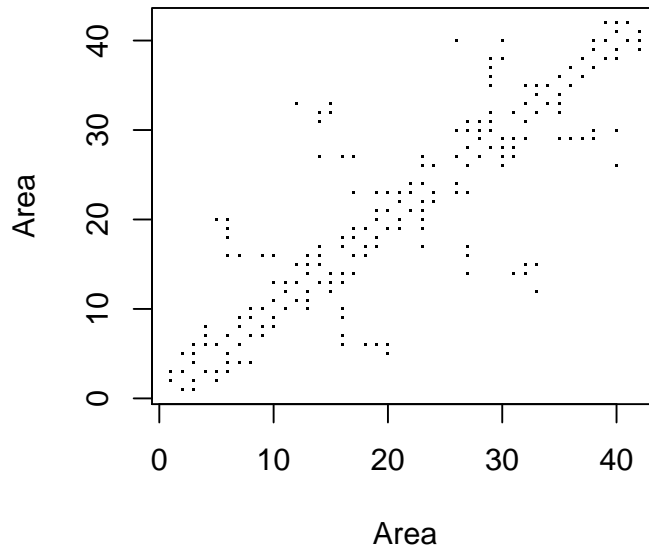


Figure 1: Matrix of contiguity used to simulate the data set.

```

> smmry <- data.frame(region = 1:m)
> smmry$Ytot <- as.vector(by(y[, 2], y[, 1], sum))
> smmry$Ymean <- as.vector(by(y[, 2], y[, 1], mean))
> smmry$Xtot <- as.vector(by(X[, 1], X[, 2], sum))
> smmry$Xmean <- as.vector(by(X[, 1], X[, 2], mean))
> smmry$N <- as.vector(by(y[, 2], y[, 1], length))
> smmry$n <- n
> smmry$YStot <- as.vector(by(S[, 2], S[, 1], sum))
> smmry$YSmean <- as.vector(by(S[, 2], S[, 1], mean))
> smmry$XStot <- as.vector(by(XS[, 1], XS[, 2], sum))
> smmry$XSmean <- as.vector(by(XS[, 1], XS[, 2], mean))
> smmry$sigma2e <- as.vector(by(y[, 2], y[, 1], var))
> smmry$sigma2eS <- as.vector(by(S[, 2], S[, 1], var))
> library(xtable)
> l <- c(1, 3, 5, 6, 7, 9, 11, 12, 13)
> xt <- xtable(smmry[, l], align = "r|c|ccc|ccc|cc", caption = "Summary of the different a
+   label = "tab:smmry")

```

	region	Ymean	Xmean	N	n	YSmean	XSmean	sigma2e	sigma2eS
1	1.00	138.54	485.26	251.00	3.00	150.45	535.11	4378.15	2413.29
2	2.00	124.00	439.80	334.00	7.00	74.98	210.47	2628.16	1608.40
3	3.00	79.41	199.32	166.00	9.00	73.83	225.99	623.11	665.42
4	4.00	84.85	220.22	195.00	10.00	88.62	208.16	726.47	1220.42
5	5.00	104.03	364.65	301.00	14.00	107.16	391.20	2153.32	1163.23
6	6.00	66.92	181.47	344.00	3.00	78.10	267.47	378.56	457.86
7	7.00	120.29	445.26	339.00	7.00	103.90	406.80	3196.33	316.76
8	8.00	92.64	317.55	290.00	13.00	89.24	323.64	1902.93	1453.25
9	9.00	92.87	337.62	227.00	5.00	71.28	224.93	1814.82	2587.66
10	10.00	93.33	296.27	116.00	3.00	78.68	245.37	1373.45	249.79
11	11.00	86.86	167.56	260.00	3.00	90.82	146.71	471.26	1315.79
12	12.00	135.44	376.21	328.00	11.00	126.97	361.48	1757.39	1192.84
13	13.00	63.98	123.84	123.00	4.00	60.30	113.04	194.66	83.40
14	14.00	30.62	96.47	173.00	9.00	41.32	118.79	222.42	300.22
15	15.00	65.70	128.80	292.00	8.00	70.94	126.49	272.32	256.84
16	16.00	104.22	380.07	163.00	5.00	138.31	472.31	2200.26	2213.43
17	17.00	71.89	140.88	229.00	5.00	61.86	114.34	382.25	517.99
18	18.00	90.89	280.65	269.00	8.00	95.59	298.32	1112.06	2013.04
19	19.00	125.96	401.57	136.00	8.00	130.10	424.25	1919.03	1306.15
20	20.00	61.93	103.18	275.00	10.00	62.13	103.77	137.24	195.03
21	21.00	119.19	427.72	339.00	7.00	95.89	339.93	1934.37	1777.94
22	22.00	101.23	395.74	307.00	10.00	102.61	415.23	2078.18	1407.72
23	23.00	109.04	358.52	129.00	7.00	90.13	279.42	1280.91	1754.82
24	24.00	56.26	191.44	160.00	6.00	35.93	118.39	692.99	510.06
25	25.00	43.28	193.52	285.00	11.00	54.05	228.80	651.85	689.21
26	26.00	64.47	206.34	178.00	4.00	49.69	134.08	523.05	194.09
27	27.00	41.83	76.75	189.00	8.00	35.88	85.87	141.68	84.27
28	28.00	15.20	110.16	328.00	3.00	6.86	87.84	269.25	17.89
29	29.00	94.43	490.65	168.00	4.00	84.89	491.92	3163.45	3301.29
30	30.00	76.28	332.60	208.00	6.00	80.64	369.88	1603.95	1017.83
31	31.00	82.73	391.13	230.00	11.00	103.52	422.80	2008.22	2587.25
32	32.00	74.95	356.48	131.00	7.00	80.77	365.60	1677.28	1399.99
33	33.00	72.36	404.14	339.00	12.00	60.12	346.70	2400.83	2453.08
34	34.00	-0.57	150.66	290.00	10.00	9.51	211.75	487.80	571.71
35	35.00	28.21	209.64	134.00	10.00	36.74	214.86	817.11	419.27
36	36.00	28.03	243.97	332.00	3.00	47.26	296.04	738.25	315.63
37	37.00	19.36	201.55	196.00	13.00	21.14	208.19	620.08	332.66
38	38.00	83.20	279.62	114.00	13.00	94.82	315.38	1147.09	823.82
39	39.00	133.30	427.19	346.00	8.00	149.38	497.52	2662.53	2898.70
40	40.00	74.32	198.95	291.00	12.00	77.61	178.34	467.58	365.24
41	41.00	126.35	380.68	274.00	4.00	107.30	285.39	2158.65	2107.52
42	42.00	113.16	289.24	280.00	13.00	103.39	222.11	1063.90	764.61

Table 1: Summary of the different areas simulated. See text for details.

4 Methods described

Currently there are many methods for Small Area Estimation. In this vignette we will constrain to the most widely known, together with the recently appeared Spatial EBLUP [Petrucci et al., 2005].

In order to assess the quality of the estimates provided by each estimator, we have considered the Average Empirical Mean Square Error (AEMSE), which is obtained by taking the mean value of all the EMSEs obtained for the 42 areas:

$$\text{AEMSE} = \frac{1}{m} \sum_{i=1}^m (\hat{Y}_i - \bar{Y}_i)^2$$

where \hat{Y}_i is an estimate of the true area mean \bar{Y}_i .

Hence, the AEMSE provides a global measure of the quality of the estimates, and the results for different estimators can be compared. Note that the lower the AEMSE the better the estimates fit to the real values.

Other measures of fit can be used as well [EURAREA Consortium, 2004].

5 Direct estimators

5.1 π -estimator

This is the most basic estimator and can only be used when all the areas have been sampled. For the area mean value it is as follows:

$$\hat{Y}_{i,DIRECT} = \sum_j w_{ij} y_{ij} / \sum_j w_{ij} \quad (1)$$

The weights w_{ij} have been taken as the inverse of the probability of an individual to be in the sample. Note that since all areas are sampled independently and with replacement, the probability of selecting individual j in area i (p_{ij}) is $1/N_i$, where N_i is the number of individuals in area i . Hence, if the sample size in region i is n_i , the probability of selecting an individual at least once is $1 - (1 - \frac{1}{N_i})^{n_i}$. This is the inclusion probability and we will use weights

$$w_{ij}^{-1} = w_i^{-1} = 1 - (1 - \frac{1}{N_i})^{n_i}$$

The weights are computed in the following chunk of code. **S** represents the values of the target variable, area id and individual id obtained in the sample. Similarly, **XS** (which is used later) contains the values of the covariate of the individuals in the sample. **smmry** is a matrix which contains the summary statistics shown in Table 1.

```
> probs <- 1/N
> probs <- 1 - (1 - probs)^n
> w <- 1/probs
> piest <- by(S[, 2], S[, 1], sum)
> piest <- as.vector(piest) * w/(n * w)
> actmeans <- smmry$Ymean
```

Figure 2 shows the values estimated by the π -estimator against the true means, which are known. The line representing the equation $y = x$ has also been included in the figure so that the quality of the estimates can be visually assessed. The Average Empirical Mean Square Error (AEMSE) of this estimator is 217.2.

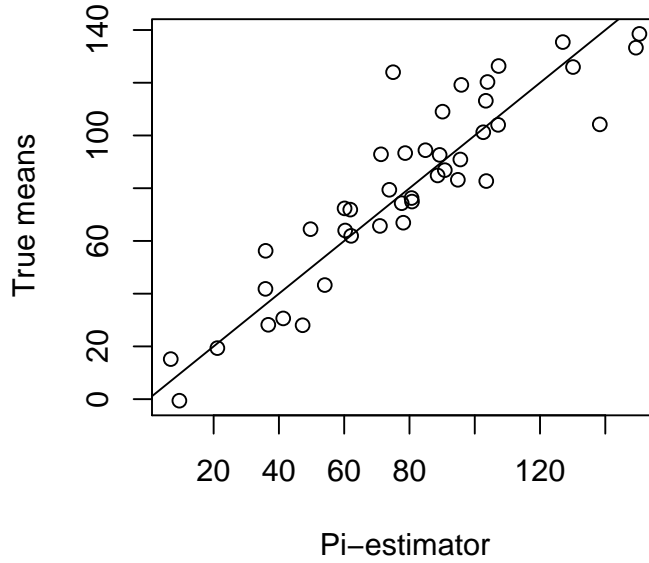


Figure 2: π -estimator against the actual mean values. The straight line represents the equation $y = x$.

The variance of the direct estimator, which is also known as *design variance*, can be estimated to assess the uncertainty about the estimates. This can be used to provide approximate confidence intervals. For the case of simple random sampling with replacement, which is the design that we are using, the design variance of the direct estimator (1) is

$$V[\widehat{Y}_{i,DIRECT}] = (1 - 1/N_i)S_i^2/n_i \quad (2)$$

Here, S_i^2 is the variance of the sample obtained from area i . The variance can be estimated by

$$\widehat{V}[\widehat{Y}_{i,DIRECT}] = (1 - 1/N_i)\widehat{S}_i^2/n_i \quad (3)$$

That is, we substitute the variance of a generic sample S_i^2 by the actual variance of the observed data \widehat{S}_i^2 .

```
> vardir <- matrix(as.numeric(tapply(y[, 2], y[, 1], var)) * (1 -
+ 1/N)/n, ncol = 1)
```

Although sampling with replacement have been used here, it is less efficient (i.e. the direct estimator has higher variance) than sampling without replacement [see Lehtonen and Pahkinen, 2004, for a detailed discussion and examples on this topic].

5.2 GREG

The Generalised Regression Estimator [GREG, Särndal et al., 1992] combines direct information from the sample with aggregated data in order to improve the quality of the estimates.

$$\widehat{Y}_{i,GREG} = \widehat{\beta X}_i^T + \sum_{j=1}^{n_i} w_{ij} (y_{ij} - \widehat{\beta} \frac{x_{ij}^T}{N_i}) = \widehat{\beta X}_i^T + \widehat{Y}_{i,DIRECT} - \sum_{j=1}^{n_i} w_{ij} \widehat{\beta} \frac{x_{ij}^T}{N_i} \quad (4)$$

This estimator is based on a linear predictor (computed using weighted regression of the sample data) plus some weighted correction terms based on the sampled units and the difference between the observed and predicted values of each individual.

```
> d <- data.frame(Y = S[, 2], X = XS[, 1], INTERCEP = 1)
> wgreg <- w[XS[, 2]]
> lmgreg <- lm(formula = Y ~ -1 + INTERCEP + X, data = d, weights = wgreg)
> lmgreg
```

Call:

```
lm(formula = Y ~ -1 + INTERCEP + X, data = d, weights = wgreg)
```

Coefficients:

```
INTERCEP      X
 23.4513    0.2050
```

```
> p1 <- as.vector(predict(lmgreg, data.frame(INTERCEP = 1, X = Xpop[,
+ 2])))
> p2 <- as.vector(predict(lmgreg, data.frame(INTERCEP = 1, X = XS[,
+ 1])))
> p2 <- (S[, 2] - p2) * wgreg
> p2 <- as.vector(by(p2, S[, 1], sum))/(n * w)
> gregest <- p1 + p2
```

In the code above, p1 estimates the vector of values

$$\widehat{\beta X}_i^T, i = 1, \dots, d$$

and p2 estimates the values

$$\sum_{j=1}^{n_i} w_{ij} (y_{ij} - \widehat{\beta} \frac{x_{ij}^T}{N_i}), i = 1, \dots, d$$

Furthermore, Xpop is a vector with the area mean values of the covariate.

The GREG estimates have been plotted against the true values in Figure 3. It can be seen how the fit seems better now. The AEMSE of this estimator is 34.91, which is lower than the one obtained with the direct estimator, which confirms that the GREG provides better estimates.

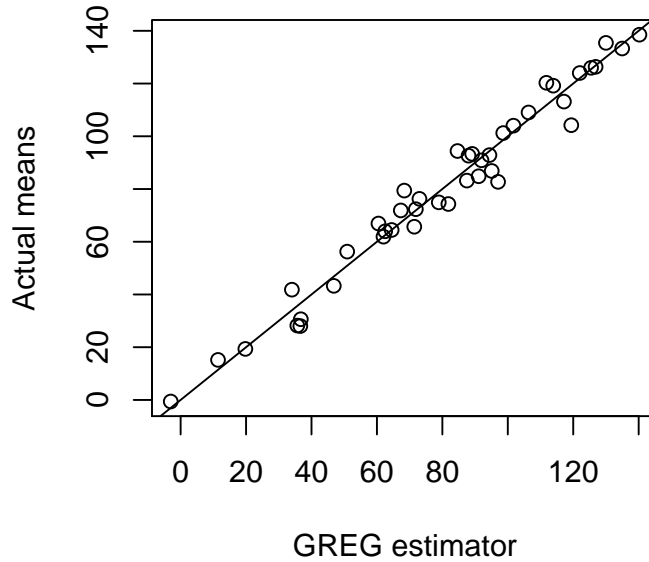


Figure 3: GREG estimator against the true mean values. The straight line represents the equation $y = x$.

6 Model-based estimators

6.1 Synthetic estimator

The synthetic estimator is based on assuming a (linear) model for the data so that the values of the areas that have not been sampled are estimated from the model using only information for available covariates. For the mean, the synthetic estimator is based on the following model:

$$\bar{Y}_i = \beta \bar{X}_i + u_i$$

where u_i is an area-based random error, which is Normally distributed with zero mean and variance σ_u^2 .

The synthetic estimator can be obtained by using the estimate of β from linear regression of the individual level sample data and computing

$$\hat{\bar{Y}}_{i,SYNTH} = \hat{\beta} \bar{X}_i \quad (5)$$

as the *synthetic estimate* in area i . Note that this estimator doesn't make any use of the random effects u_i and that for this reason it may lead to biased estimates of the area means.

```

> synthd <- data.frame(Y = ydir1, COV = Xpop[, 2])
> synthmodel <- lm(Y ~ 1 + COV, data = synthd)
> synthmodel

Call:
lm(formula = Y ~ 1 + COV, data = synthd)

Coefficients:
(Intercept)      COV
    21.2845      0.2058

> syntheest <- predict(synthmodel, data.frame(COV = Xpop[, 2]))

```

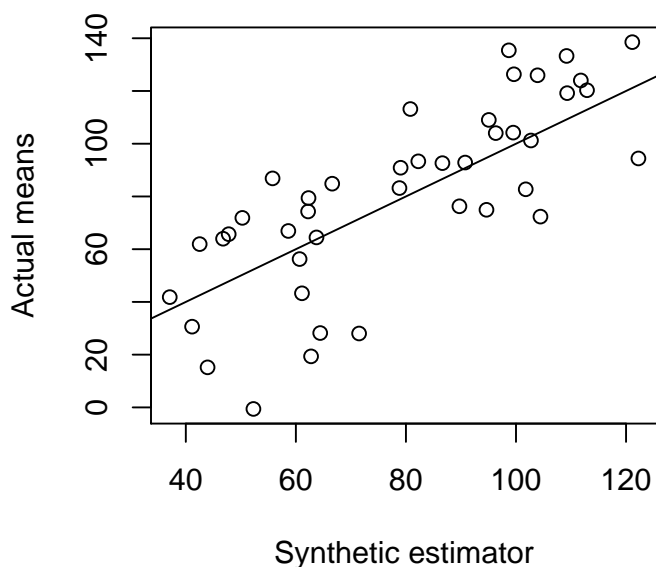


Figure 4: Estimates obtained with a Synthetic estimator against the true values.

Figure 4 shows the synthetic estimates against the true values. The AEMSE of this estimator is 489.6. This is clearly higher than the one obtained with the π -estimator, probably because we have neglected the effect of the random effects u_i and the within area variation.

6.2 Multilevel modelling

Multilevel modelling [Goldstein, 2003] can be used to construct a representation of the model for SAE that is divided into different layers, which represent different effects. This type of modelling can be used to model information at the area and individual level.

Estimation of the parameters of the model and inference comes in two flavours: frequentist and bayesian. Not surprisingly many times the solutions provided by those approaches are quite similar [see, for example, Rao, 2003, Chapter 10].

6.2.1 Unit level model

Given that in Small Area Estimation we consider the problem of obtaining a reliable estimate from a sample and additional auxiliary covariates, our first layer is at the individual level. This can be expressed as follows:

$$y_{ij}|\beta, u_i, \sigma_e^2 \sim N(\beta x_{ij} + z_i u_i, \sigma_e^2)$$

where z_i are usually 1 but can be used to model different types of area effects u_i . The next level specifies the area effects:

$$u_i|\sigma_u^2 \sim N(0, \sigma_u^2)$$

By considering different structures for the area random effects it is possible to model different spatial and temporal (or both) effects. Furthermore, in this example we have only considered two levels, but in some applications we may have more than two.

This model can be written in a matrix form as follows:

$$\begin{aligned} \mathbf{y}|\beta, U, \sigma_e^2 &\sim \text{Normal}(\beta \mathbf{x} + ZU, \sigma_e^2 I_N) \\ U|\sigma_u^2 &\sim \text{Normal}(0, \sigma_u^2 I_m) \end{aligned}$$

where $N = \sum_{i=1}^m N_i$ and I_d is the identity matrix of $d \times d$ dimension.

Unconditionally on the area effects, we have the following model:

$$\mathbf{y}|\beta, \sigma_e^2, \sigma_u^2 \sim \text{Normal}(\beta \mathbf{x}, \sigma_e^2 I_N + \sigma_u^2 Z I_m Z^T)$$

6.2.2 Area level model

The previous model can be used to provide estimates when only aggregated data are available

$$\bar{Y}_i|\beta, \sigma_e^2, \sigma_u^2 \sim N(\beta \bar{X}_i, \sigma_e^2/n_i + \sigma_u^2)$$

6.3 Linear mixed models with R

6.3.1 Area level models

Package `nlme` provides some functions to fit linear mixed-effects models and we have used it to compute small area estimates of the previous area model. Note that we need to fix the structure of the variance matrix of the sampling errors (object `vf` below) to $\sigma_e^2 \text{diag}(1/n_1, \dots, 1/n_d)$. Note that in order to be able to estimate the model, we have substituted the unknown value of \bar{Y}_i by its direct estimator (`piest` or `ydir1` in the code chunks).

The previous model cannot be estimated because we have one observation per area and we need to estimate two variances: the one for the random effects (σ_u^2) and that of the error term (σ_e^2). The use of direct estimators with area level models is better explained in Section 6.5, and it usually involves estimating (and

fixing) σ_e^2 using the sampling scheme before estimating the parameters of the model.

6.3.2 Unit level models

Using the individual level data we have:

```
> library(nlme)
> dunit <- data.frame(Y = S[, 2], X = XS[, 1], region = S[, 1])
> lmmunit <- lme(Y ~ 1 + X, random = ~1 | region, data = dunit,
+ method = "ML")
> lmmunit
```

Linear mixed-effects model fit by maximum likelihood

```
Data: dunit
Log-likelihood: -1463.871
Fixed: Y ~ 1 + X
(Intercept)          X
24.4188236    0.2008334
```

Random effects:

```
Formula: ~1 | region
(Intercept) Residual
StdDev:    21.05765 18.31329
```

Number of Observations: 327

Number of Groups: 42

```
> yunit <- data.frame(X = X[, 1], region = X[, 2])
> ypred <- as.vector(predict(lmmunit, yunit))
> lmmunitest <- as.vector(by(ypred, yunit$region, mean))
```

Figure 5 shows the estimates against the true values. The average MSE of this estimator is 36.95.

6.4 Composite estimator

The composite estimator is constructed as a weighted sum of the direct estimator and the synthetic estimator. The weights are defined so that if the sample size is “large” the direct estimate is given more weight than the synthetic one and when the sample is not reliable, the synthetic estimate will be given more weight. The equation of this estimator is as follows:

$$\widehat{Y}_{i,COMP} = \widehat{\gamma}_i \widehat{Y}_{i,DIRECT} + (1 - \widehat{\gamma}_i) \widehat{Y}_{i,SYNTH} \quad (6)$$

$\widehat{\gamma}_i$ is a value between 0 and 1 which controls the shrinkage of the two estimators. That is, depending on how large is the sample in the small area it will give more weight to the direct estimate (if the sample is large) or to the synthetic estimate (if information is needed from other areas). $\widehat{\gamma}_i$ is chosen so that it minimises the MSE of (6) or the average MSE of all synthetic estimators Ghosh and Rao [1994].

In the model with individual and area level effects, the weights are taken as

```
> plot(lmmunitest, actmeans, xlab = "True means", ylab = "Fitted means using a LMM (UNIT 1
> abline(coef = c(0, 1))
```

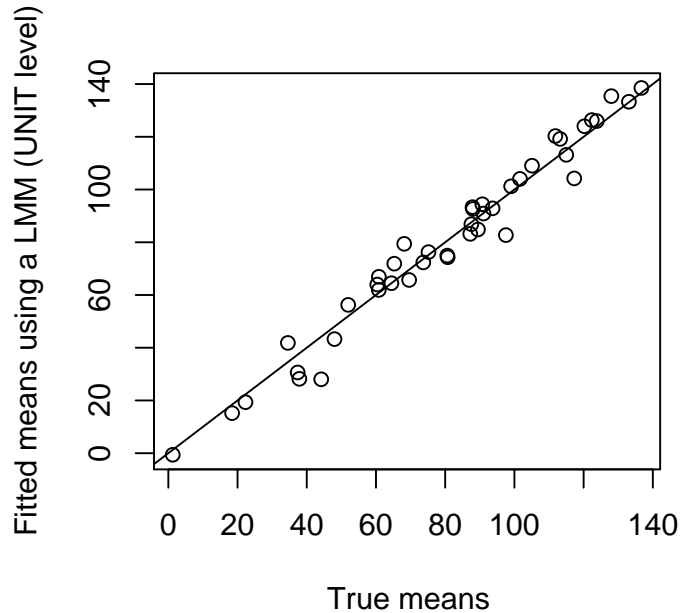


Figure 5: Estimates provided by a Multilevel model using unit level data.

$$\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}_e^2/n_i}$$

```
> sigma2u <- exp(as.numeric(attr(lmmunit$apVar, "Pars"))[1])
> sigma2e <- exp(as.numeric(attr(lmmunit$apVar, "Pars"))[2])
> gamma <- sigma2u/(sigma2u + sigma2e/n)
> compest <- gamma * synthest + (1 - gamma) * piest
```

Figure 6 displays the composite estimates against the true values. The average MSE of this estimator is 401.2. It fits better than the synthetic estimator, but considerably worse than the π -estimator.

6.5 EBLUP estimators

The EBLUP estimators considered in this vignette are the same as described in Rao [2003]. In addition, we have included the Spatial EBLUP [Petrucci and Salvati, 2005], which takes into account the spatial structure of the data by modelling the random effects according to a SAR specification.

Basically, we assume that the parameter of interest $\theta = (\theta_1, \dots, \theta_m)$, which can be the vector of totals or area means, can be expressed as

```
> plot(compest, actmeans, xlab = "True means", ylab = "Composite estimator")
> abline(coef = c(0, 1))
```

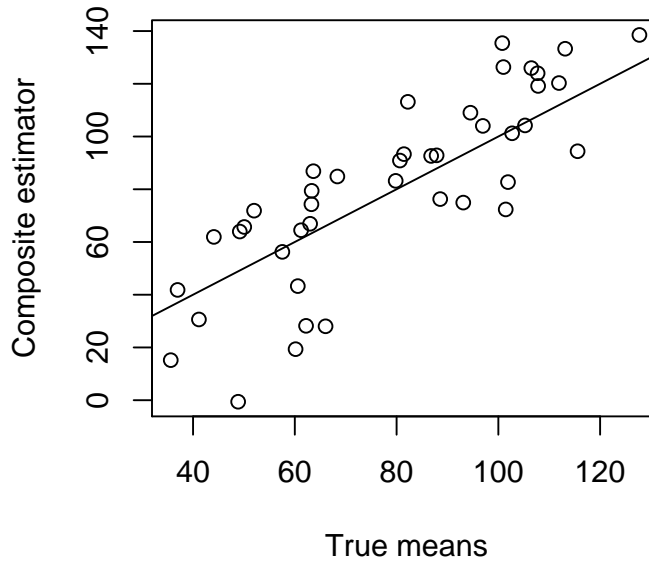


Figure 6: Composite estimator against the true means.

$$\theta = X\beta + Zu$$

where X is a set of covariates, β their associated coefficients, u area random effects and Z a matrix that models the (possible) structure of u . The random effects u are distributed with zero mean and variance σ_u^2 .

Given that a direct estimator $\hat{\theta}$ is available, we will combine both using a Fay-Herriot model [Rao, 2003]:

$$\hat{\theta} = \theta + e$$

where e represents the sampling error (i.e., variance) and is a diagonal matrix of known constants obtained from the survey design. In addition, we will assume that u and e are independent.

The estimates of β are obtained using standard Generalised Least Square techniques, whilst the estimates of u are computed using their Empirical Best Linear Unbiased Predictor (EBLUP):

$$\hat{u} = E[u|y]$$

where y represents the sampled data. \hat{u} can be computed using Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML). Computational details can be found in McCulloch and Searle [2001], Rao [2003].

The EBLUP estimator of θ is then defined as

$$\tilde{\theta} = X\hat{\beta} + Z\hat{u}$$

6.5.1 EBLUP estimator

In this case we consider the Fay-Herriot area level model to estimate the area level means. Given that we have area-level data, the variance of the sampling error e is plugged-in and it is a diagonal matrix made of the design variances of the different areas, computed as in equation (3).

```
> eblupml <- EBLUP.area(ydir1, Xpop, vardir, m)
> eblupreml <- EBLUP.area(ydir1, Xpop, vardir, m, method = "REML")
```

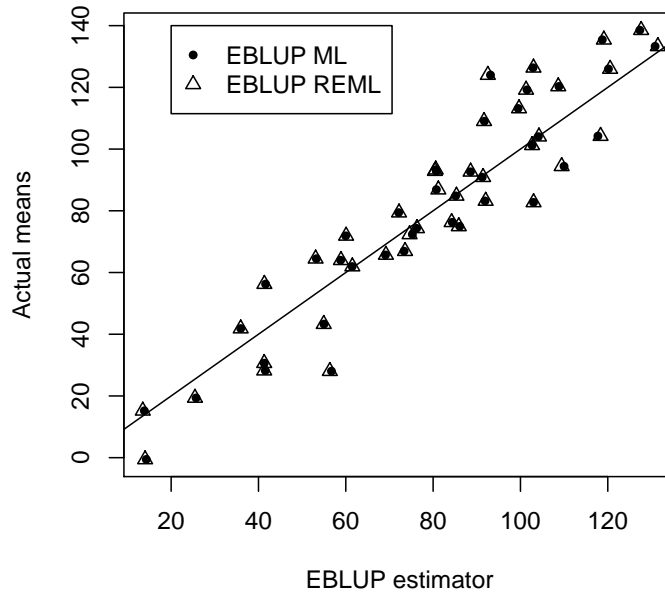


Figure 7: EBLUP estimates.

The EBLUP estimates have been plotted in Figure 7 against their true values. The average MSE of this estimator are (ML and REML) 155.3 and 154.9.

6.5.2 Spatial EBLUP estimator

This estimator is basically an EBLUP estimator before but considering a Simultaneously Autoregressive (SAR) specification as for the area random effects. In

this case, the parameter of interest θ is

$$\theta = X\beta + Zv$$

where $v = \rho Wv + u$. Here, ρ is a spatial autoregressive coefficient, W the adjacency matrix and u is defined as before.

Hence, the relationship between the direct estimator and the true value can be written as follows:

$$\hat{\theta} = \theta + e = X\beta + Z(I - \rho W)^{-1}u + e$$

The estimation of β is done using GLS and the estimates of v are computed using an EBLUP estimator $\hat{v} = E[v|y]$. See Petrucci et al. [2005], Petrucci and Salvati [2005] for computational details.

The EBLUP estimator of θ is then defined as

$$\tilde{\theta} = X\hat{\beta} + Z\hat{v}$$

```
> seblupml <- SEBLUP.area(ydir1, Xpop, vardir, m, W)
> seblupreml <- SEBLUP.area(ydir1, Xpop, vardir, m, W, method = "REML")
```

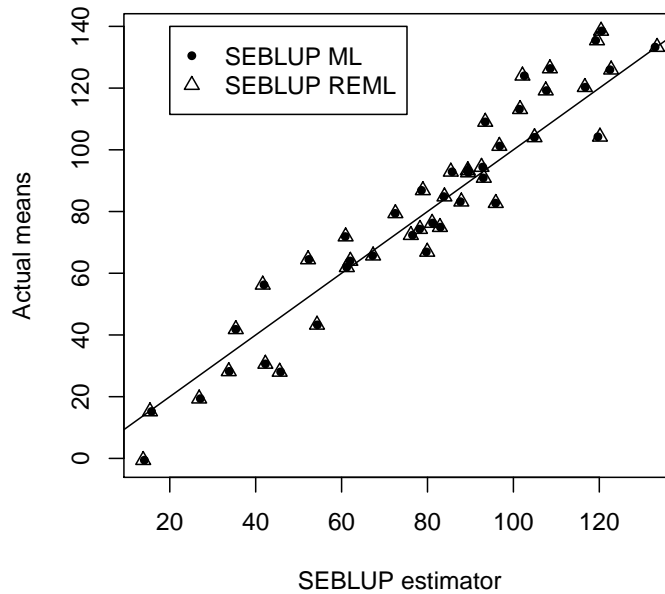


Figure 8: SEBLUP estimates.

The Spatial EBLUP estimates are shown in Figure 8 against their true values. The average MSE of this estimator are (ML and REML) 101 and 101.6. The SEBLUP clearly provides better estimates than the EBLUP because the former does not take into account the spatial structure of the data.

6.5.3 EBLUP vs SEBLUP

Figure 9 compares the EBLUP and SEBLUP estimators. In addition, Figure 10 displays these estimates and the deviation from the true mean of the SEBLUP and how it is, in most cases, lower than the distance between the EBLUP and the true mean.

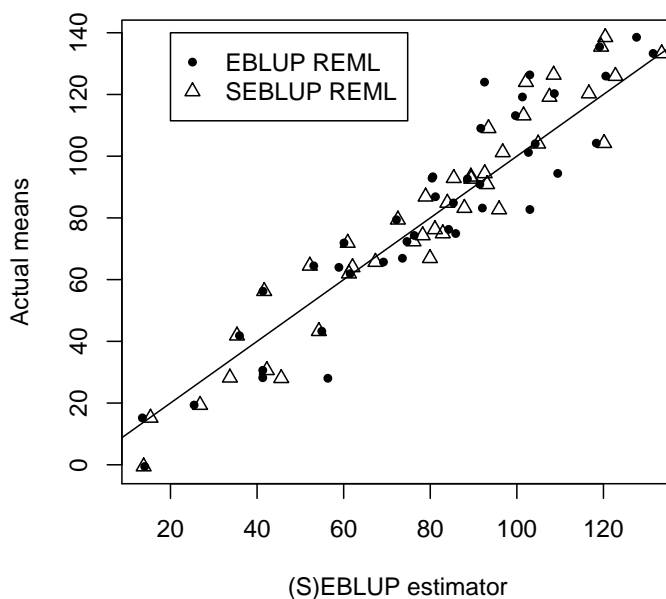


Figure 9: EBLUP estimates.

7 Acknowledgements

Part of this work has been done within the BIAS Project (<http://www.bias-project.org.uk>), funded by the Economic and Social Research Council, United Kingdom. Nicola Salvati provided the code to compute the EBLUP and SEBLUP estimators.

References

- EURAREA Consortium. Project reference volume. Technical report, EURAREA Consortium, 2004.
- M. Ghosh and J. N. K. Rao. Small area estimation: An appraisal. *Statistical Science*, 9(1):55–76, 1994.

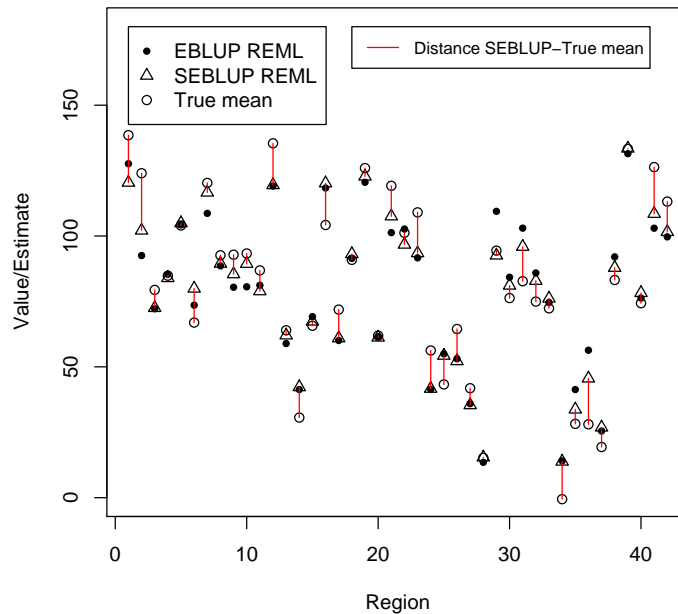


Figure 10: EBLUP and SEBLUP estimates compared to the true means.

- H. Goldstein. *Multilevel Statistical Models, 3rd ed.* Arnold Publishers, London, 2003.
- R. Lehtonen and E. Pahkinen. *Practical Methods for Design and Analysis of Complex Surveys.* John Wiley & Sons Ltd., Chichester, 2004.
- C. E. McCulloch and S. R. Searle. *Generalized, Linear, and Mixed Models.* John Wiley & Sons, Inc., New York, 2001.
- A. Petrucci, M. Pratesi, and N. Salvati. Geographic information in small area estimation: Small area models and spatially correlated random area effects. *Statistics in Transition*, 3(7):609–623, 2005.
- A. Petrucci and N. Salvati. Small area estimation considering spatially correlated errors: the unit level random effects model. Technical report, University of Florence, Department of Statistics, 2004.
- A. Petrucci and N. Salvati. Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological & Environmental Statistics*, 11(2):169–182, 2005.
- J. N. K. Rao. *Small Area Estimation.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.

C.-E. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*.
Springer-Verlag Inc., New York, 1992.