

# Small Area Estimation - Practicum

## EMOS Learning Materials

Service Contract n. 2019.0249 between Eurostat and the University of Pisa, Italy

September 2021

## Introduction

Usually sample surveys are used to infer parameters from a population of interest. Often the population of interest can be divided in geographical areas (e.g. regions) or domains obtained cross-classifying some characteristics of the units (e.g. age groups and gender). When the sample size of a specific area or domain is small, then estimators based only on specific area/domain observations - i.e. direct estimators - do not usually reach a minimum level of precision. We refer to these areas or domains as small areas. Small area estimation methods use the whole sample data assisted by auxiliary data to obtain a reliable estimator for small areas.

Modern small area estimation methods make use of linking model to bring strength from all the sample and the auxiliary variables.

The family of small area models can be divided into two groups: area-level models and unit-level models. The first uses area-level (aggregate) data for model fit and for estimating target statistics. The latter uses unit-level survey data (also called micro-data) and area-level (or aggregate) covariates for estimating means or totals, while to estimate other target statistics, such as quantiles, inequality indexes, etc., there is need of unit-level covariates for all the population units. Data availability plays a crucial role in the choice of the models. We will start by an application of the basic area-level model, or Fay and Herriot (1979) model and then by an application of the basic unit-level model (Battese, Harter, and Fuller 1988).

The two applications (exercises) are carried out using the R computing language (R Development Core Team 2013), a powerful free statistical environment. This practicum has been obtained using R-markdown. We suggest to practitioners and student to make use (together with R) of RStudio, an integrated development environment (IDE) for the R language.

## Packages for SAE

There are few packages available for small area estimation. Some of them are `sae`, `rsae`, `saeRobust`, `JoSAE`, `hbsae`, `BayesSAE` and `emdi`. In this basic practicum we will use the `emdi` package (Kreutzmann et al. 2019) and the `sae` package (Molina and Marhuenda 2015).

```
library(emdi)

## Registered S3 method overwritten by 'MuMIn':
##   method      from
##   predict.merMod lme4
##
## Attaching package: 'emdi'
##
## The following object is masked from 'package:stats':
##
##   step
```

```
library(sae)
```

```
## Loading required package: MASS
## Loading required package: lme4
## Loading required package: Matrix
##
## Attaching package: 'sae'
## The following object is masked from 'package:emdi':
##
##     direct
```

Other R packages are useful to manipulate and display data:

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:MASS':
##
##     select
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

## Area-level approach to SAE

To illustrate the application of the basic area-level model we use the data sets `direct.estimateds` and `auxiliary.vars` available from the `Practicum.Dataframes.arealevel.RData` file.

```
load("Practicum.Dataframes.arealevel.RData")
```

These are synthetic data for Italian provinces, obtained from the 2012 household budget survey and from the Italian Tax Agency register.

Our goal is to obtain reliable estimates of the mean consumption expenditure at the province level (NUTS 3) in Italy. We need direct estimates of the target, its estimated standard error and a set of auxiliary variables at provincial level.

The data frame `direct.estimateds` contains information about direct estimates of the consumption expenditure:

- `area.code`: Italian provinces id's;
- `cons.exp`: direct estimates of the mean consumption expenditure at provincial level;
- `cons.exp.se`: estimated standard errors of the mean consumption expenditure at provincial level;
- `sample.size`: effective sample size.

The data frame `auxiliary.vars` contains information about taxable income that comes from the Italian Tax Agency Register:

- `taxable.income.percapita`: mean per capita taxable income at provincial level;
- `area.code`: Italian provinces id's direct.

First, we analyse the reliability of the direct estimates by computing the estimated coefficient of variation (CV):

```
direct <- direct.estimates$cons.exp
se.direct <- direct.estimates$cons.exp.se
cv.direct <- se.direct/abs(direct)

summary(cv.direct)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.06111 0.11773 0.16515 0.20129 0.22136 1.29496
```

National statistical institutes use different definition for assessing the quality of their estimates. For the sake of simplicity, in this practicum we consider reliable those estimates with a CV smaller than 16%. As we can see from the summary of the CVs, about half of the provinces have reliable direct estimates, while half have not. Using R and the `dplyr` package is easy to obtain a table with the number of reliable and unreliable estimates:

```
direct.estimates %>% mutate(cv.direct = cons.exp.se/cons.exp) %>%
  mutate(direct.reliable = cv.direct <= 0.16) %>%
  group_by(direct.reliable) %>%
  count()
```

```
## # A tibble: 2 x 2
## # Groups:   direct.reliable [2]
##   direct.reliable     n
##   <lgl>           <int>
## 1 FALSE             57
## 2 TRUE              53
```

We have 53 provinces where the direct estimates of the mean consumption expenditure is considered reliable, and 57 provinces where direct estimates are unreliable.

We now set the data to be used in package `emdi` to get EBLUPs and their estimated MSE by the `fh` function.

The `fh` function requires data to be properly prepared by the `combined_data` functions as follows:

```
direct.estimates$cons.exp.var <- direct.estimates$cons.exp.se^2
comb.data <- combine_data(pop_data = auxiliary.vars, pop_domains = "area.code",
  smp_data = direct.estimates, smp_domains = "area.code")
```

Once the data are prepared we can obtain our estimates:

```
fh.estimates <- fh(fixed = cons.exp ~ 1 + taxable.income.percapita,
  vardir = "cons.exp.var",
  combined_data = comb.data, domains = "area.code",
  method = "reml", MSE = TRUE)
```

The object `fh.estimates` contains EBLUPs and their estimated MSE as well as other statistics, such as model fit diagnostics. A way to get an overview of the model fit is using `summary`:

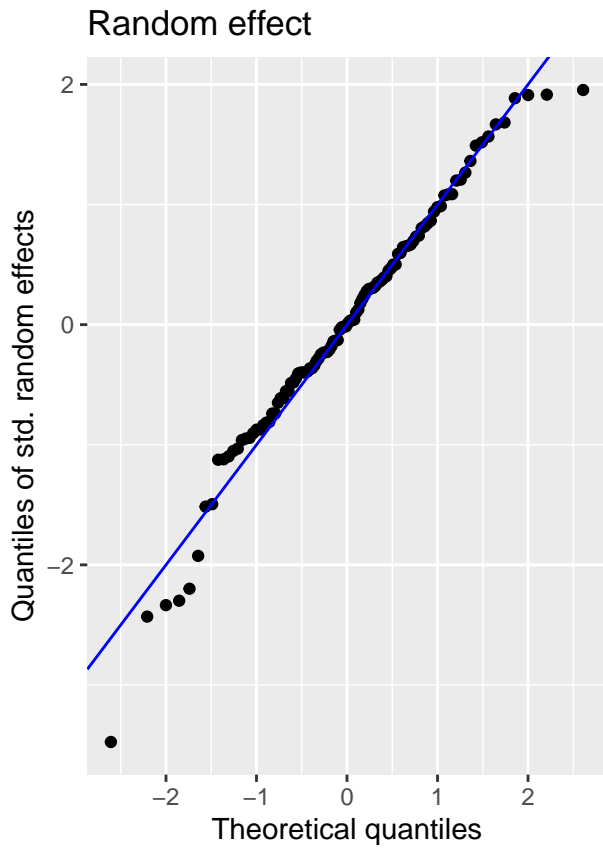
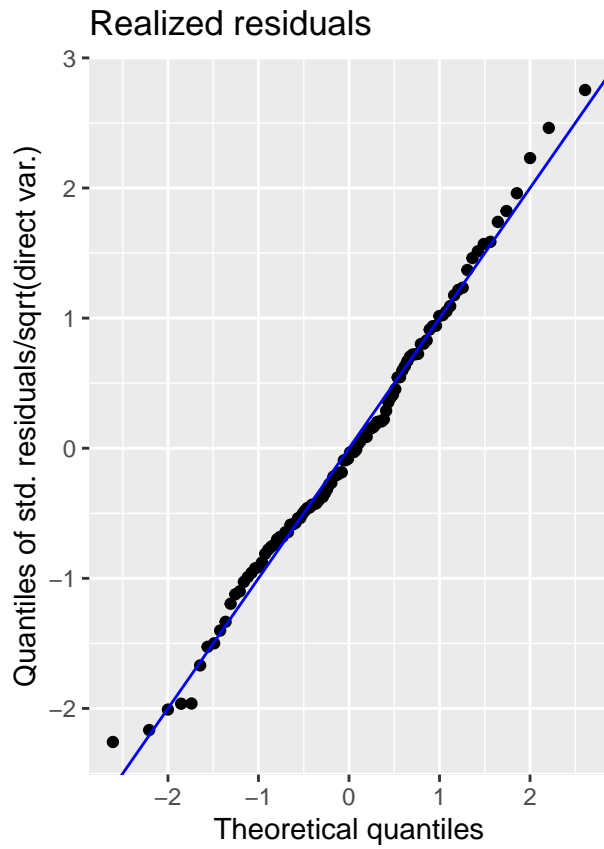
```
summary(fh.estimates)
```

```
## Call:
## fh(fixed = cons.exp ~ 1 + taxable.income.percapita, vardir = "cons.exp.var",
##     combined_data = comb.data, domains = "area.code", method = "reml",
##     MSE = TRUE)
```

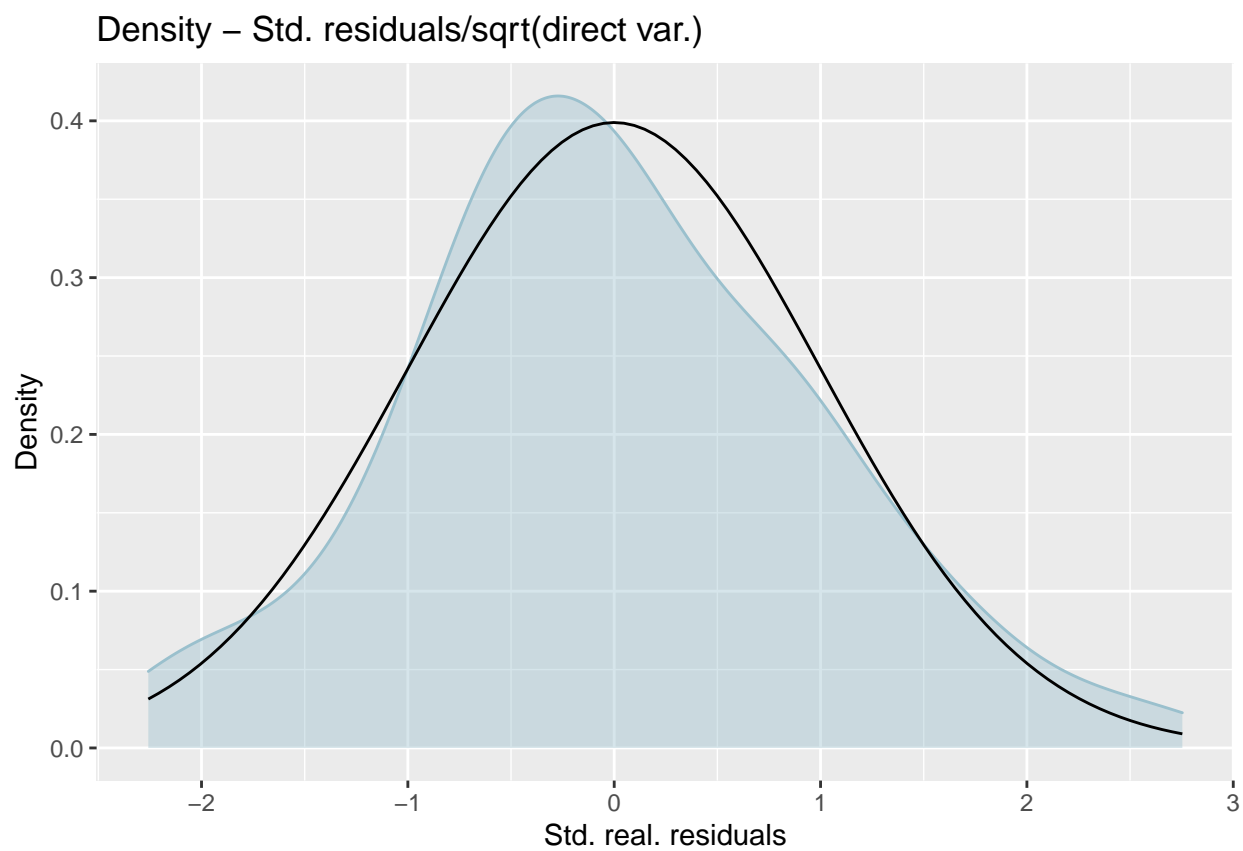
```
##
## Out-of-sample domains: 0
## In-sample domains: 110
##
## Variance and MSE estimation:
## Variance estimation method: reml
## Estimated variance component(s): 92712.18
## MSE method: prasad-rao
##
## Coefficients:
##               coefficients    std.error  t.value    p.value
## (Intercept)      588.3545565 294.24564694 1.999535 4.555047e-02
## taxable.income.percapita 0.0966379 0.01630044 5.928545 3.056299e-09
##
## Explanatory measures:
##      loglike      AIC      AICc      AICb1      AICb2      BIC      KIC      KICc
## 1 -845.0276 1696.055 1695.465 1696.441 1694.857 1704.157 1699.055 1699.085
##      KICb1      KICb2      R2      AdjR2
## 1 1700.061 1698.477 0.1197622 0.2646471
##
## Residual diagnostics:
##               Skewness Kurtosis Shapiro_W Shapiro_p
## Standardized_Residuals 0.2129190 3.045442 0.9909489 0.68138338
## Random_effects      -0.4844142 3.768139 0.9767722 0.05154707
##
## Transformation: No transformation
```

From the summary we can see we have 0 out-of-sample areas (areas for which sample size is 0), 110 in-sample areas, we used restricted maximum likelihood (reml) to obtain coefficients and random effects, the mse is estimated using the analytic approximation of PrasadRao:90. The model fit shows both intercept and taxable per capita income are statistically significant and the residual diagnostic shows normality assumption is reasonable for random effects (p-value > 0.05). Normality assumptions can also be checked graphically using plot:

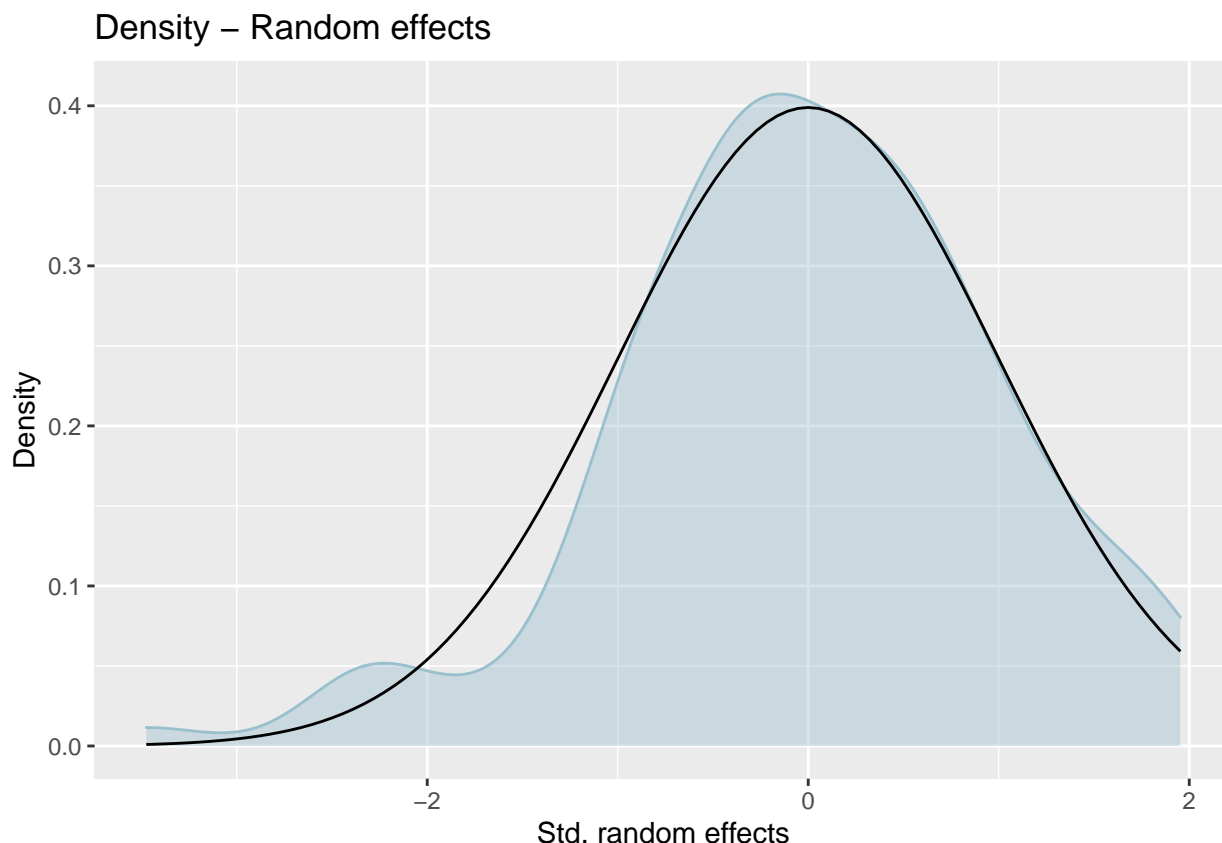
```
plot(fh.estimates)
```



## Press [enter] to continue



## Press [enter] to continue



The EBLUPs and their estimated MSE can be extracted from the object as follows:

```
EBLUP <- fh.estimated$ind$FH
mse <- fh.estimated$MSE$FH
```

It is useful to build a data frame with direct estimates and their estimated standard errors, EBLUPs and their estimated MSE, province code and sample size as well as the CVs:

```
RESULTS.AREALEVEL.SAE <- data.frame(area.code = fh.estimated$ind$Domain,
                                     direct=fh.estimated$ind$Direct,
                                     direct.se=sqrt(fh.estimated$MSE$Direct),
                                     EBLUP=fh.estimated$ind$FH,
                                     EBLUP.rmse=sqrt(fh.estimated$MSE$FH))
RESULTS.AREALEVEL.SAE <- full_join(x = RESULTS.AREALEVEL.SAE,
                                   y = direct.estimated[,c(1,4)],
                                   by = "area.code")

RESULTS.AREALEVEL.SAE$cv.direct <-
  RESULTS.AREALEVEL.SAE$direct.se/RESULTS.AREALEVEL.SAE$direct
RESULTS.AREALEVEL.SAE$cv.EBLUP <-
  RESULTS.AREALEVEL.SAE$EBLUP.rmse/RESULTS.AREALEVEL.SAE$EBLUP
```

Using package `emdi` is easy to get a map of the estimates. There is need of a `SpatialPolygonsDataFrame` object as defined by the `sp` package on which the data should be visualized. In our case we to obtain such object we use the `maptools` package (which is deprecated) and the function `readShapePoly` to read the shape file `Prov01012012_g_WGS84` (freely available from the Istat - Italian national statistical office):

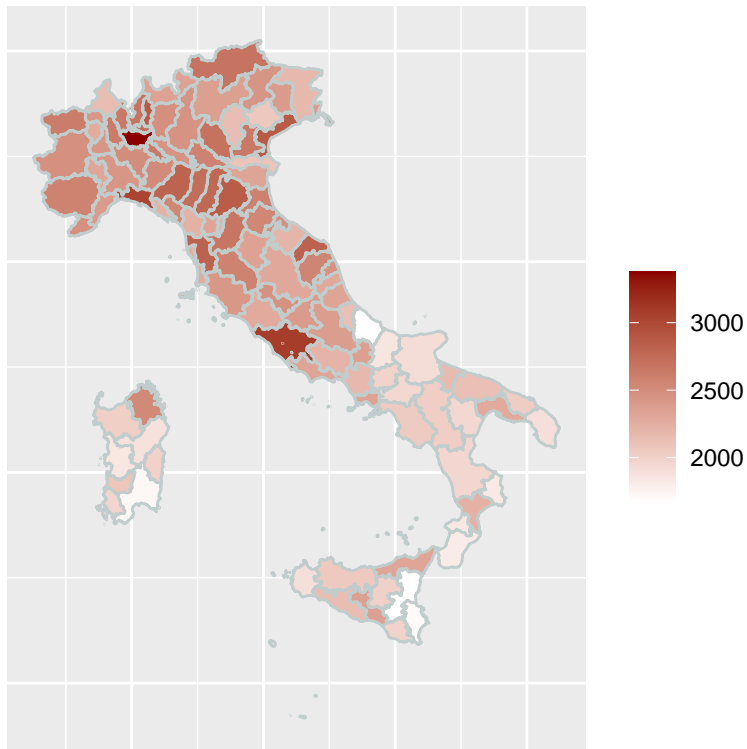
```
library(maptools)
```

```
## Loading required package: sp
```

```
## Checking rgeos availability: TRUE
prov.spdf <- readShapePoly("Prov01012012_g_WGS84")

## Warning: readShapePoly is deprecated; use rgdal::readOGR or sf::st_read
map_plot(object = fh.estimates,
          indicator = "FH",
          MSE = FALSE,
          map_obj = prov.spdf,
          map_dom_id = "COD_PROV")
```

FH



Finally, we check how many EBLUPs are considered reliable, that is  $CV \leq 16\%$ :

```
RESULTS.AREALEVEL.SAE$direct.reliable <-
  factor(x = RESULTS.AREALEVEL.SAE$cv.direct <= 0.16,
        levels = c(FALSE, TRUE),
        labels = c("Direct Unreliable", "Direct Reliable"))

RESULTS.AREALEVEL.SAE$EBLUP.reliable <-
  factor(x = RESULTS.AREALEVEL.SAE$cv.EBLUP <= 0.16,
        levels = c(FALSE, TRUE),
        labels = c("EBLUP Unreliable", "EBLUP Reliable"))

addmargins(table(RESULTS.AREALEVEL.SAE$direct.reliable,
                 RESULTS.AREALEVEL.SAE$EBLUP.reliable))
```

```
##
##               EBLUP Unreliable EBLUP Reliable Sum
## Direct Unreliable             0             57  57
## Direct Reliable               0             53  53
```



```
##      Sum              0          110 110
```

As discussed previously, using direct estimation we get reliable estimates of the mean consumption expenditure in 53 provinces and unreliable estimates in 57 provinces. Using the EBLUP we obtain reliable estimates for all the provinces.

## Unit-level approach to SAE

To illustrate the application of the basic unit-level model we use the data sets `amelia.smp.data` and `pop.means` available from the `Practicum.Dataframes.unitlevel.RData` file.

```
load("Practicum.Dataframes.unitlevel.RData")
```

AMELIA is an artificial data set that enables and fosters comparative and reproducible research (Burgard et al. 2020, 2017). AMELIA comprises the following four regional levels which are listed in descending order of area size: 1. REG (Region), 2. PROV (Province), 3. DIS (District), 4. CIT (City/Community). The AMELIA population data frame consists of approx. 10 million observations of 33 variables on personal level.

For the purpose of this practicum a stratified simple random sample has been drawn from the AMELIA population, where strata are the districts, which are also the areas of interest. The total sample size is 2015 units (persons), enough to get reliable estimates at the country level. The sample data are in the `amelia.smp.data` data frame.

The goal of this example is to estimate the mean income at the district level. The sample sizes at district level are small, therefore, direct estimation of mean income is unreliable and there is need to resort to small area estimation.

The data frame `amelia.smp.data` contains the following variables:

- AGE: age of the person;
- BAS: basic activity status (Work, Unemployment, Retired, Other Inactivity);
- COB: country of birth (Local, EU, Other);
- HHS: household size;
- MST: marital status (Never married, Married, Separated, Widowed, Divorced);
- SEX: sex of the person (male, female);
- INC: income (euros);
- DIS: district code.

Together with the sample we have the `pop.means` data frame, which has the population means of some variables that will be used in the small area model:

- `BASUnemployment`: district proportion of unemployed persons in the population;
- `BASRetired`: district proportion of retired persons in the population;
- `BASOther Inactivity`: district proportion of inactive persons (other than retired) in the population;
- `COBEU`: district proportion of persons born in EU (not in the AMELIA country) in the population;
- `COBOther`: district proportion of persons born outside EU in the population;
- `SEXFemale`: district proportion of female in the population;
- `DIS`: district code;
- `Pop.size`: district population size.

Ideally, these data can be obtained from population registers.

We start computing direct estimates of the mean income at district level and checking their reliability, using the same definition of reliable estimates used in the area level section (i.e.  $CV \leq 0.16$ ). Direct estimates and their estimated standard error should be obtained according to the sampling design. For example the package `survey` is used to analyse complex survey samples (Lumley 2019, 2004). Here, the sample design is very simple, and conditional to the district is a simple random sampling design without replications. For such a simple design we can use the function `direct` of the `sae` package:

```
direct.estimates <- sae::direct(y = INC,
                               dom = DIS,
                               domsize = data.frame(1:40, pop.means$Pop.size),
                               data = amelia.smp.data)
head(direct.estimates)
```

```
##   Domain SampSize   Direct      SD      CV
## 1     1         5 19925.746 5442.319 27.31300
## 2     2         5  2032.053 1447.459 71.23133
## 3     3         5  7095.637 3520.348 49.61285
## 4     4        32 15729.332 3453.198 21.95388
## 5     5        51  8079.921 1800.165 22.27949
## 6     6        49 15808.117 3080.455 19.48654
```

Note: the CV is in percentage.

The number of districts with reliable and unreliable estimates of the mean income are obtained as follows:

```
direct.estimates$CV <- direct.estimates$CV/100
direct.estimates %>% mutate(direct.reliable = CV <= 0.16) %>%
  group_by(direct.reliable) %>%
  count()
```

```
## # A tibble: 2 x 2
## # Groups:   direct.reliable [2]
##   direct.reliable     n
##   <lgl>           <int>
## 1 FALSE           34
## 2 TRUE             6
```

We have 6 districts with reliable estimates of mean income and 34 with unreliable estimates.

Next step is to estimate the random intercept model on the sample data and obtain EBLUPs and their estimated MSEs. The variables available at population level (district means/proportions) are BAS, COB and SEX, therefore the random intercept model for the income must use the same variables in its fixed part. The MSE is estimated using a parametric bootstrap even though analytic approximation of the MSE of the unit-level EBLUP exists (not implemented in sae package). The function to obtain EBLUPs and their estimated MSE (by bootstrap) is `pbmseBHF` (parametric bootstrap mse Battese, Harter and Fuller, see Battese, Harter, and Fuller (1988)):

```
set.seed(1)
bhf.estimates <- sae::pbmseBHF(formula = INC ~ BAS + COB + SEX,
                               dom = DIS,
                               meanxpop = data.frame(1:40, pop.means[,1:6]),
                               popnsiz = data.frame(1:40, pop.means$Pop.size),
                               B = 399, data = amelia.smp.data)
```

From the object `bhf.estimates` we can check the model:

```
bhf.estimates$est$fit$summary

## Linear mixed model fit by REML ['lmerMod']
## Formula: ys ~ -1 + Xs + (1 | dom)
##
## REML criterion at convergence: 38641.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.7232 -0.5272 -0.2022  0.3005 16.5208
##
## Random effects:
##   Groups   Name      Variance Std.Dev.
##   dom      (Intercept) 5814470 2411
##   Residual              368830434 19205
## Number of obs: 1717, groups: dom, 40
##
## Fixed effects:
##               Estimate Std. Error t value
## XsXs(Intercept)      24024.5      883.2  27.201
## XsXsBASUnemployment   -11379.8     2113.7  -5.384
## XsXsBASRetired        -15328.3     1222.2 -12.541
## XsXsBASOther Inactivity -14357.6     1193.7 -12.028
## XsXsCOBEU              9123.1     2660.3   3.429
## XsXsCOBOther           194.0     1564.6   0.124
## XsXsSEXFemale         -2666.4      941.3  -2.833
##
## Correlation of Fixed Effects:
##           XsX(I)  XXBASU  XXBASR  XXBASI  XXCOBE  XXCOBO
## XsXsBASUnmp  -0.189
## XsXsBASRtrd  -0.345  0.177
## XsXsBASOthI  -0.327  0.178  0.302
## XsXsCOBEU    -0.064 -0.061 -0.087 -0.042
## XsXsCOB0thr  -0.166 -0.029 -0.043  0.019  0.064
## XsXsSEXFeml  -0.464 -0.068 -0.067 -0.147  0.004 -0.008
```

The variables we used are statistically significant, but COBOother that is not different from the base COBLocal. Anyway, we are not interested in the interpretation of the coefficient, because the model is used for predictive purposes. Usually, we try different combination of variables in the model until we get a good working model, which fit the best to the data in order to obtain efficient model-based estimates. For simplicity, in this example we use the proposed model without investigating further on models comparisons. We are going to see that it fit well enough to improve efficiency with respect to the direct estimates.

To check the normality assumptions on area random errors and unit errors we use the Shapiro and Wilk (1965) test:

```
shapiro.test(bhf.estimates$est$fit$random$`(Intercept)`)
```

```
##
## Shapiro-Wilk normality test
##
## data:  bhf.estimates$est$fit$random$`(Intercept)`
## W = 0.93838, p-value = 0.0305
```

```
shapiro.test(bhf.estimates$est$fit$residuals)
```

```
##
## Shapiro-Wilk normality test
##
## data:  bhf.estimates$est$fit$residuals
## W = 0.74615, p-value < 2.2e-16
```

Both area random errors and unit errors are not normally distributed. However, the area random effect are symmetric (according to Miao, Gel, and Gastwirth (2006)), that is enough to get EBLUPs and approximately unbiased MSE estimates. In practical applications many alternative are possible, for example fit the model using a Box-Cox transformation, resort to robust EBLUP or to the M-quantile approach.

We collect in a data frame the direct estimates and their estimated standard errors together with the EBLUPs and their estimated MSEs:

```
RESULTS.UNITLEVEL.SAE <- data.frame(area.code = bhf.estimates$est$eblup$domain,
                                     EBLUP=bhf.estimates$est$eblup$eblup,
                                     EBLUP.rmse=sqrt(bhf.estimates$mse$mse))
RESULTS.UNITLEVEL.SAE <- merge(x = RESULTS.UNITLEVEL.SAE, y = direct.estimates[,c(1:4)],
                               by.x = "area.code", by.y = "Domain")

names(RESULTS.UNITLEVEL.SAE)[6] <- "Direct.se"

RESULTS.UNITLEVEL.SAE$cv.direct <-
  RESULTS.UNITLEVEL.SAE$Direct.se/RESULTS.UNITLEVEL.SAE$Direct

RESULTS.UNITLEVEL.SAE$cv.EBLUP <-
  RESULTS.UNITLEVEL.SAE$EBLUP.rmse/RESULTS.UNITLEVEL.SAE$EBLUP
```

Finally, we check how many EBLUPs are considered reliable, that is  $CV \leq 16\%$ :

```
RESULTS.UNITLEVEL.SAE$direct.reliable <-
  factor(x = RESULTS.UNITLEVEL.SAE$cv.direct <= 0.16,
         levels = c(FALSE, TRUE),
         labels = c("Direct Unreliable", "Direct Reliable"))

RESULTS.UNITLEVEL.SAE$EBLUP.reliable <-
  factor(x = RESULTS.UNITLEVEL.SAE$cv.EBLUP <= 0.16,
         levels = c(FALSE, TRUE),
         labels = c("EBLUP Unreliable", "EBLUP Reliable"))

addmargins(table(RESULTS.UNITLEVEL.SAE$direct.reliable,
                 RESULTS.UNITLEVEL.SAE$EBLUP.reliable))
```

```
##
##               EBLUP Unreliable EBLUP Reliable Sum
## Direct Unreliable             2             32 34
## Direct Reliable              0              6  6
## Sum                         2             38 40
```

As discussed previously, using direct estimation we get reliable estimates of the mean income in 6 districts and unreliable estimates in 34 districts. Using the EBLUP we obtain reliable estimates for 38 districts and unreliable just for 2 districts.

## Conclusions

In this practicum we have shown how to apply basic small area estimation methods.

The basic area-level model (Fay and Herriot 1979) has been applied to synthetic data to obtain reliable province level estimates of the mean consumption expenditures. The unit-level model to estimate means or proportions (Battese, Harter, and Fuller 1988) has been applied to the AMELIA synthetic data to estimate district level mean income.

Area-level estimation has been carried out using the `emdi` package, while for the unit-level estimation we used the `sae` package. This choice has been made to show the reader two of the most used package in small area estimation in R. Please, note that both the package can compute both unit-level and area-level small area estimates. Other package for small area estimation are available too, as mentioned in the introduction.

Many extensions to the presented small area estimation methods exist in literature and most of them are

implemented in the used packages.

## References

- Battese, G.E., Harter R.M., and W.M. Fuller. 1988. “An Error Component Model for Prediction of County Crop Areas Using Survey and Satellite Data.” *Journal of the American Statistical Association* 83: 28–36.
- Burgard, J.P., F. Ertz, H. Merkle, and R. Münnich. 2020. “AMELIA - Data Description V0.2.3.1.” University of Trier; Research Institute for Official; Survey Statistics.
- Burgard, J.P., J.P. Kolb, H. Merkle, and R. Münnich. 2017. “Synthetic Data for Open and Reproducible Methodological Research in Social Sciences and Official Statistics.” *AStA Wirtschafts- Und Sozialstatistisches Archiv* 11 (3): 233–44.
- Fay, R.E., and R.A. Herriot. 1979. “Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data.” *Journal of the American Statistical Association* 74: 269–77.
- Kreutzmann, Ann-Kristin, Sören Pannier, Natalia Rojas-Perilla, Timo Schmid, Matthias Templ, and Nikos Tzavidis. 2019. “The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators.” *Journal of Statistical Software* 91 (7): 1–33.
- Lumley, T. 2019. “Analysis of Complex Survey Samples.” R package version 3.35-1.
- Lumley, Thomas. 2004. “Analysis of Complex Survey Samples.” *Journal of Statistical Software* 9 (1): 1–19.
- Miao, W., Y.R. Gel, and J.L. Gastwirth. 2006. “A New Test of Symmetry About an Unknown Median.” In *Random Walk, Sequential Analysis and Related Topics – a Festschrift in Honor of Yuan-Shih Chow*, edited by A. Hsiung, C. Zhang, and Z. Ying. World Scientific Publisher.
- Molina, Isabel, and Yolanda Marhuenda. 2015. “sae: An R Package for Small Area Estimation.” *The R Journal* 7 (1): 81–98. <https://journal.r-project.org/archive/2015/RJ-2015-007/RJ-2015-007.pdf>.
- R Development Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Shapiro, S.S, and M.B. Wilk. 1965. “An Analysis of Variance Test for Normality (Complete Samples).” *Biometrika* 67: 215–16.