Small Area Estimation with R

Unit 1: Introduction to Small Area Estimation

V. Gómez-Rubio

Department of Mathematics
Universidad de Castilla-La Mancha, Spain

Tutorial Outline

9:00-9:25 Introduction to Small Area Estimation

9:25-9:50 Design-based estimators

9:50-10:15 Model-based estimators

10:15-10:30 *Coffee break*

10:30-11:00 EBLUP estimators

11:00-11:30 Bayesian Small Area Estimation

11:30-12:00 Non-linear models

Course Materials

All materials can be downloaded from github:
`https://github.com/becarioprecario/SAERTutorial`

- Slides
- `R` packages (SAE not available on CRAN yet)
- Data sets: Simulated Swedish data (income and employment)
- `R` scripts
- WinBUGS code (for the Bayesian models)

Other on-line resources

- European Working Group on Small Area Estimation
  `http://sae.wzr.pl/`
- EURAREA Project
  `http://www.statistics.gov.uk/eurarea/`
- Course on Spatial Data Analysis with R (unit 9)
  `http://www.bias-project.org.uk/ASDARcourse/`
- Office for National Statistics
  `http://www.statistics.gov.uk/`
- Task View on Official Statistics
  `https://cran.r-project.org/web/views/OfficialStatistics.html`

Small Area Estimation

Definition (Rao, 2003)

A domain (area) is regarded as "small" if the domain-specific sample is not large enough to support direct estimates of adequate precission.

Which means that...

- Small does not always refer to the size of the area (US states can be small areas!)
- Small refers to the fraction between the sample size and the total population in the area
- Depending on the problem we will prefer the term domain (and not area), as it has been done in the survey literature

Examples of Case studies

Batting rates

- Eforn and Morris (1975) considered the problem of estimating the batting average of Roberto Clemente and another 17 baseball players
- The batting average is the number of hits divided by the number of times that the player batted
- All had batted 45 times that season
- The aim is to estimate the batting averages for the remainder of the 1970

Population

- Governments carry out surveys to measure the total population in the country
- Census are collected on a regular basis (sually, every 10 years)
- However, estimates are required for the between-Census years
- The same data may be used to provide estimates for smaller domains

Examples of Case studies

Income per household

- Governments carry out surveys to measure income of the population
- The target of the survey is to collect the income at a number of household
- A few areas are selected to be representative of the population

Disease Mapping

- Public health authorities regularly collect data on mortality
- The aim is to estimate the risk of suffering (or dying) from a particular disease
- When other covariates area available (i.e., risk factors, environmental covariates) mortality can be linked to its causes
- These data are different from survey data because they provide an exhaustive accounting of the target variable

Estimation in domains and sub-domain

- Sometimes survey data are collected to provide estimates of the variable of interest in a single domain (for example, unemployment rate in California)
- However, other smaller sub-domains may be of interest as well (for example, counties in California)
- The survey design will depend on the target (sub-)domains
- Estimation of the sub-domains may led to an improved estimation in the domain (for example, stratification)

Survey design and data collection

Survey design

- The survey design determines what (sub-)domains are sampled and how the units within are taken
- To reduce costs not all (sub)domains are sampled
- The sample is taken so that it is significant at the domain level
- Stratification is useful to improve estimation
- For example, to estimate the unemployment rate in England the survey can be carried in different districts (subdomains)

Data collection

The survey will produce a set of observations

$$\{(y_{ij}, \mathbf{x}_{ij})\};\ i = 1, \ldots, K; j = 1, \ldots, n_i$$

in the K subdomains, each one with a sample size n_i.

Other data sources

Administrative data

- Statistical bureaus and other similar institutions host a myriad of other data sets that can be useful. For example, the Census.
- Previous similar surveys may be available
- Volumes with aggregated data are regularly published by statistical offices
- Aggregate data are usually easier to obtain due to confidentiality issues
- In the UK, the Office for National Statistics is a good source of data: Family Resources Survey,
- In the US, the Census Bureau provides data on a number of topics

Combining aggregate and individual data

- Surveys seldom cover all domains of interest
- Aggregate data (from official sources) often provide covariates for every area in the domain
- In order to provide estimates in all subdomains, efficient ways of *combining* this information are required
- Some models can cope with both individual and aggregate data
- The main idea is to use the observed/collected data to fit a suitable model and then produce estimates in all the domains using the available covariates

Some R packages for Small Area Estimation

- `sampling`
  Methods for sampling
- `survey`
  Methods for the analysis of complex survey data
- `nlme`
  Mixed-effects models
- `lme4`
  Next generation package to fit mixed-effects models
- `sae`
  Some functions for SAE. Described in Rao & Molina (2015)
- `SAE2`
  Provides some (spatial) EBLUP estimators (under development)

Software available for other statistical packages

General software packages

In principle, any *programmable* software package can be used for small area estimation

EURAREA Project

`Enhancing small area estimation techniques to meet European needs`

- The EURAREA project was a research programme funded by Eurostat under the Fifth Framework (FP5) Programme of the European Union to investigate methods for Small Area Estimation and their application. The project ran from January 2001 until June 2004.
- The research outputs include a final Project Reference Volume and macro language programs written in SAS.
- `http://www.statistics.gov.uk/eurarea/`

Challenges in the development of software for SAE (2008)

- Provide model-based estimators that account for complex effects: space and time
- Provide appropriate estimates of the variance of the estimators, to develop approximate confidence intervals
- Develop documentation and vignettes
- At the moment, there is no equivalent to the EURAREA SAS macros in `R`
- `nlme` and `lme4` can be a suitable starting point and development framework
  - Develop other structure for the random effects (spatial random effects, etc.)
  - Enhance the output returned to compute other estimates of the variance of the small area estimates, etc.

Official Statistics Task View

- Provides an annotated list of packages for Official Statistics (not only SAE)
  - Complex Survey Design
  - Visualisation
  - Imputation of missing values
  - Statistical Disclosure Control
  - Small Area Estimation
  - Indices and Indicators
  - Microsimulation

References

- BIAS Project. `http://www.bias-project.org.uk`
- Efron, B. and C. Morris (1975). Data Analysis Using Stein's Estimator and its Generalizations. *Journal of the American Statistical Association*, 70 (350): 311-319
- EURAREA Consortium (2004). Project reference volume. Technical report, EURAREA Consortium.
- Ghosh, M. and J. N. K. Rao (1994). Small area estimation: An appraisal. *Statistical Science 9*(1), 55–76.
- Lehtonen, R. and E. Pahkinen (2004). *Practical Methods for Design and Analysis of Complex Surveys*, 2nd ed. Wiley & Sons, Chichester.
- Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation, 2nd ed.*. John Wiley & Sons, Inc., Hoboken, New Jersey.
- Särndall, C.E., B. Swensson and J. Wretman (2003, reprinted). *Model assisted survey sampling*. Springer, New York.