University of Applied Sciences and Arts Northwestern Switzerland
School of Business

Small Area Estimation (Part III)

Short Course – Institute of Statistics, Republic of Albania

Tobias Schoch, Feburary 17 and 18, 2022

About

This slide deck has been prepared for the “Short Course on Small Area Estimation with R” at the Institute of Statistics of the Republic of Albania on February 17 and 18, 2022.

Contact

University of Applied Sciences Northwestern Switzerland
School of Business – Institute ICC
Prof. Dr. Tobias Schoch
Riggenbachstrasse 16
CH-4600 Olten
Switzerland

E-Mail: tobias.schoch@fhnw.ch
Phone: +41 62 957 21 02

Outline

1. Introduction
2. Developing a SAE Plan
3. Applications

Literature

1 Introduction

Why SAE?

More **disaggregated** and **granular statistics** (with acceptable precision) can help facilitating **more specific** and efficient policy- or intervention-targeting.

- Focus beyond broad trends and averages at national level
- Focus towards identifying subgroups of the population

⇨ Identifying subpopulations + Formulating specific policies + Monitoring

Challenges?

Disaggregation & higher granularity ⇨ **more information**

1 Introduction (ctd.)

More information: Options?

1. **Larger surveys** (more financial, personal, and administrative resources)
2. **Decentralizing data collection**
   - Local governments are mandated with data collection; National Statistical Institute oversees the process
   - The success depends heavily on the budget and technical capabilities of local governments.
   - [other issues: coordination, consistency of concepts and definitions]
3. **Minimum level of disaggregation is negotiated and fixed**
   - E.g., national government needs reliable poverty rates at provincial-level ⇨ survey (cost-effective)
   - Further disaggregation suffers from lack data (only a few observations)

⇨ Exploiting the feasibility of **SAE**

1 Introduction (ctd.)

SAE…

- is a **cost-effective** strategy to enhance granularity of statistics
- provides the **analytical framework** for improving the level of granularity without necessarily collecting large amounts of data
- **integrates multiple data sources** and capitalizes on the strength of each source
- builds on **borrowing strength** from other data sources (census, administrative data, etc.)
- is **commonly used** in many National Statistical Institutes all over the world

2 Developing a SAE Plan

2.1 Goal and Purpose of SAE

- What are **relevant policies** or programs that (potentially) require granular data?
- What are the **stakeholders' strategic goals** & desired results (context)?
- What **specific** granular data are required (regions, subpopulations, etc.)?
- What are the **inevitable consequences** for the policies/ programs if the disaggregated data are imprecise (e.g., 20%, 30% coefficient of variation)
- Are there social or economic **models** (and/or expert knowledge) that can help identifying explanatory variables?
- What are the relevant **administrative-level data** that can serve as auxiliary data?

2 Developing a SAE Plan (ctd.)

2.1 Goal and Purpose of SAE (ctd.)

- **Budget** and funding?
- By how much will SAE estimators **outperform** the survey estimators in terms of precision?

2.2 Variable and Characteristic of Interest

- **How much variation** does the estimator of interest exhibit over the areas? If the estimator is almost constant across the areas, we do not need auxiliary information and a model with random effects.
- The **choice of characteristics** (mean, total or non-linear statistics, e.g., poverty gap) must be **made early** in the process because it impacts the SAE methodology.

2 Developing a SAE Plan (ctd.)

2.2 Variable and Characteristic of Interest (ctd.)

- The variable or **characteristic** may assume **different forms** depending on the information available. This has implications for the SAE methodology.
  - **Poverty rate.** The total population size N is assumed known. The rate is defined as the number of poor individuals (total) divided by N ⇨ linear statistic, Fay-Herriot model is applicable
  - **Unemployment rate.** The size of the labor force is usually not known, nor is the number of people employed. Both figures are computed with survey data. The unemployment rate is the ratio of the two quantities; thus, a non-linear statistic, Fay-Herriot model is not applicable

2 Developing a SAE Plan (ctd.)

2.3 Level of Disaggregation and Data Requirements

- The desired **level of disaggregation** should usually be **dictated** by the goal or **purpose** of conducting SAE.
- However, it is **often the case** in Official Statistics that some domains (or areas) are finer than the pre-defined granularity of the survey domains. For those domains, the quality requirements (e.g., in terms of coefficient of variation) cannot be met.
  - Is it worth considering SAE if the direct estimator does not meet quality standard in only a handful of domains (or areas)?
  - SAE becomes more appealing when many (or "important") domains fall short of meeting the standards.
- The advantage of SAE over the classical estimator hinges on the **availability of disaggregated information** (with strong predictive power).

2 Developing a SAE Plan (ctd.)

2.3 Level of Disaggregation and Data Requirements (ctd.)

- SAE plays out its strength if there exists a **strong relationship within** and/or **between** the survey estimator and other sources of data.
  - between direct estimator and auxiliary data
  - between direct estimators over time
  - in information collected for adjacent/ neighboring areas
  - between areas sharing the same features (similarity)
  - [combinations of the sources]
- **Strength of a relationship** should checked/ tested (⇨ **selection** of variables)
  - Scatter plot, correlation, etc.
  - Model: information criteria, significance, etc.

2 Developing a SAE Plan (ctd.)

2.3 Level of Disaggregation and Data Requirements (ctd.)

- Major **difficulty** in choosing disaggregated auxiliary information
  - Coverage (target population of auxiliary data vs. survey population)
  - Consistency of definitions
  - Reference period
  - Quality of administrative data
- What about **"new" data sources**?
  - Social media (self-selection bias?)
  - "Big Data" (e.g., satellite or mobile phone data; confidentiality?)

2 Developing a SAE Plan (ctd.)

2.4 Approach to SAE: Choosing a Specific Model or Technique

- There is a wide range of models…
- In my humble opinion, always start with the area-level model. Then, set a target where you want to improve on with the unit-level model. Stick with the unit-level model if and only if you can really meet the targets.

2.5 Quality Assessment of SAE Estimates

- **Internal evaluation** (within team); model: goodness of fit, diagnostics (check assumptions), compare SAE-estimates and direct estimates; compute measures of precision, accuracy and reliability
- **External evaluation** with stake holders or end users (focus group discussion, consultation, etc.): Are their requirements met?

2 Developing a SAE Plan (ctd.)

2.6 Dissemination Strategy for Presentation of the SAE Estimates

- Of course, we communicated the results in the way we are used to in Official Statistics (tables, figures, explanations, infographics, etc.)
- There are **2 additional complications** that occur with SAE
  - **Direct estimator and EBLUP (usually) differ.** To see this, consider the rate of "coronary artery bypass surgeries"; rate estimated by EBLUP is 3.5 per 1000 persons; survey: 48 cases in 15000 persons ⇨ rate = 3.2 per 1000 ⇨ Users may claim that National Statistical Institute made a mistake…
  - **Benchmarking / disaggregation.** Consider estimating a total. The sum of the area-specific totals estimated by EBLUP is (usually) not equal to the national total.
- SAE involves devising a **specific** dissemination strategy

3 Applications

In this section, I summarized applications found in Asian Development Bank (2020)

Poverty statistics

- **Philippines** and **Thailand**
- Level: Provinces and cities
- Goal: Economic and social policies program for poverty reduction
- Survey: family income and expenditure survey
- Auxiliary information: Census of population + housing register data
- Method: EBLUP + World Bank methodology (Elbers et al., 2002)

Source: Asian Development Bank (2020, p. 8–10)

3 Applications (ctd.)

Employment statistics

- **Philippines**
- Level: Municipalities and cities
- Goal: Estimates of unemployment rate among women and young & proportion of households with working children aged 5 – 17
- Survey: Labor force survey
- Auxiliary information: Census and registers
- Method: EBLUP

Source: Asian Development Bank (2020, p. 10–11)

3 Applications (ctd.)

Health and nutrition statistics

- **Philippines**
- Level: Provinces
- Goal: Proportion of underweight children aged 0 –5 years, vitamin A-deficient children, maternal mortality, etc.
- Survey: National nutrition survey
- Auxiliary information: Census and registers
- Method: EBLUP + Poisson model (count data)
- Other countries mentioned: Australia (disability), Italy (doctor visits), and USA (obesity)

Source: Asian Development Bank (2020, p. 11)

3 Applications (ctd.)

Other SAE applications

- India: Non-farming activities
- USA: Forest coverage
- Senegal: Literacy rates for men and women (mobile phone data)
- Philippines: Drug abuse

Source: Asian Development Bank (2020, p. 11–12)

Literature

Asian Development Bank (2020) *Introduction to Small Area Estimation Techniques: A Practical Guide for National Statistics Offices*, Manila (Philippines).

Elbers, Lanjouw & Lanjouw (2002) Micro–Level Estimation of Poverty and Inequality, *Econometrica* 71 (1), p. 355–364.