From Start to Finish: Specify-Analyse/Adapt-Evaluate (SAE)

Nikos Tzavidis[1], Angela Luna Hernandez, Li-Chun Zhang
(University of Southampton)
Timo Schmid & Natalia Rojas (Freie Universität Berlin)

NCRM Research Methods Festival
Bath, July 2016

http://www.ncrm.ac.uk/research/ISAEM/

[1] Presenting author

Background

The problem of estimating finite population parameters (means, proportions, totals...) has been addressed using Sample Surveys.

Design based inference (Cochran 1953, Kish 1965, Särndal et al. 1992)

- U is a finite population with fixed values of the variable of interest Y_i.
- A sample s is selected from U using a probabilistic sampling design.
- As the Y_i are assumed non-stochastic, statistical inference is based only on the probability distribution induced by the sample selection process.
- Distribution free methodology, dominated the production of Official Statistics.

The SAE problem

Users's requirements for more disaggregated estimates have been increasing in the past 10 years or so. Now we need estimates for many small areas:

- Geographic areas: municipalities, districts, neighbourhoods,...
- Domains: combinations of factors such as Age, Sex, Ethnicity, Labour Force status,...

For design based inference to work well, s needs to be big enough

- Areas with 2, 3 observations?
- Areas with no observations at all?

The SAE problem

How small is a small area?

Estimates based only on the domain-specific sample information are called direct estimates.

A small area is as a domain for which the domain-specific sample is not large enough to produce direct estimates with acceptable precision.

- In order to allocate funds (7 billion U$) to meet the educational needs of disadvantaged children, USA needs to estimate the number of school children 5-17 in families under poverty. Small Areas: county and school district.
- The World Bank supports the development of poverty maps in many countries. The definition of the geographic unit depends on data availability. Province, municipality,...

The SAE problem

Small Area Estimation (SAE) methods face the lack of domain-specific information by using models to *borrow strength* from other areas/domains.

$$\hat{Y}_{1i} = \hat{\beta}_i X_i$$
$$\hat{Y}_{2i} = \hat{\beta} X_i$$
$$\hat{Y}_{3i} = \hat{\beta} X_i + \hat{u}_i$$

- $V(\hat{Y}_{2i}) \leq V(\hat{Y}_{1i})$. However, $B(\hat{Y}_{2i}) \geq B(\hat{Y}_{1i})$. What about $MSE(\hat{Y}_{2i})$ and $MSE(\hat{Y}_{1i})$?

- $V(\hat{Y}_{3i}) \geq V(\hat{Y}_{2i})$, but hopefully not that much and $B(\hat{Y}_{3i})$ can be considerably smaller than $B(\hat{Y}_{2i})$.

Tradeoff between bias and variance.

Three stages

Aim: Outline the main stages towards the implementation of Small Area Estimation (SAE) project in practice.

Stage I. Specification

1. Specify user needs
2. Specify a set of target indicators to be estimated and a target geography/set of domains

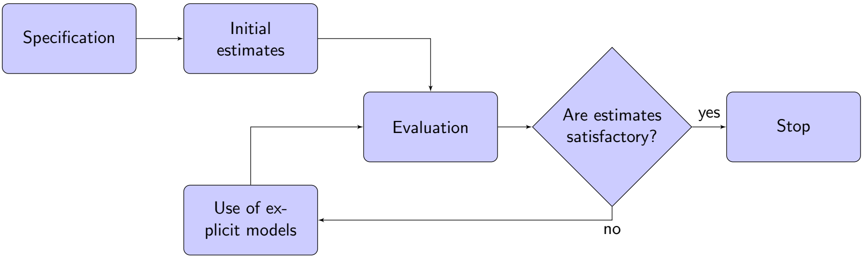
Stage II. Analysis/Adaptation

3. Initial estimates
4. Use of explicit models

Stage III. Evaluation

5. MSE estimation
6. Model and Design based evaluation
7. Further evaluation tasks

Three stages



Stage I. Specification

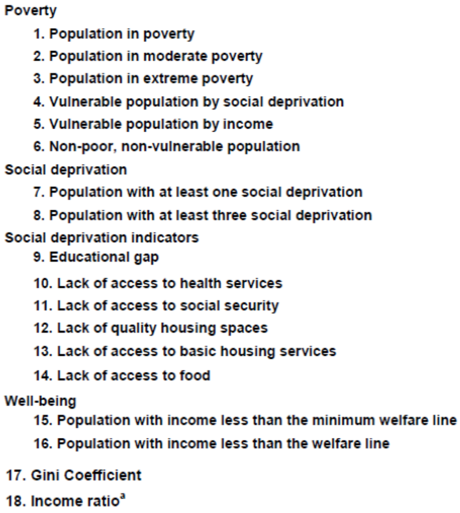
A chosen level of geography should provide meaningful (background of the problem) and useful (data availability) estimates

Follow in decreasing level of aggregation and avoid the temptation of getting unrealistically low.

- SAE is a prediction problem. Access to good auxiliary data is, in most cases, crucial.
- Survey, Census, Administrative data can be used for modelling and evaluation purposes.
- For indicators such as totals, means and proportions, area level information can be enough. More complex indicators such as percentiles may require unit level information, i.e., access to microdata.
- Consider the coverage of the sources in relation to the target geography.

Stage I. Specification

18 Indicators specified by law at the municipal level.



Poverty

1. Population in poverty
2. Population in moderate poverty
3. Population in extreme poverty
4. Vulnerable population by social deprivation
5. Vulnerable population by income
6. Non-poor, non-vulnerable population

Social deprivation

7. Population with at least one social deprivation
8. Population with at least three social deprivation

Social deprivation indicators

9. Educational gap
10. Lack of access to health services
11. Lack of access to social security
12. Lack of quality housing spaces
13. Lack of access to basic housing services
14. Lack of access to food

Well-being

15. Population with income less than the minimum welfare line
16. Population with income less than the welfare line

17. Gini Coefficient

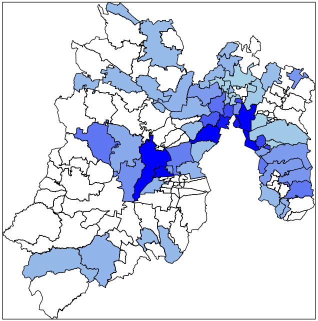
18. Income ratio[a]

- Totals
- Proportions
- More complex indicators

Feasibility will depend on the data availability.

Stage I. Specification

Geographic coverage of the data sources



- 125 municipalities in State of Mexico (EDOMEX). Only 58 are included in the survey. For the municipalities in the sample, the average sample size is 47 households.

- All municipalities are covered by the Census.

Stage II. Analysis/Adaptation

3. Initial estimates

Using only the information of the main survey, produce a triplet of estimates (direct, synthetic, composite) for each area at the given level of geography:

- Direct: uses only-domain specific data, e.g., $\hat{\bar{Y}}_k^D = \bar{X}_k \hat{\beta}_k$
- Synthetic: borrows information from other areas/domains, e.g., $\hat{\bar{Y}}_k^S = \bar{X}_k \hat{\beta}$
- Composite: it is a convex combination of a Direct and a Synthetic estimators, e.g., $\hat{\bar{Y}}_k^C = \phi \hat{\bar{Y}}_k^D + (1 - \phi) \hat{\bar{Y}}_k^S$

Unlikely these estimators to produce estimates with acceptable coefficients of variation (CVs).

Stage II. Analysis/Adaptation

4. Use of explicit models

General considerations

- Access to microdata? Unit-level or Area-level models. Complexity of the target parameters
- Continuous responses: start with Linear Models
- Discrete responses: start with Generalized Linear Models
- Unexplained heterogeneity: Mixed Models
- Out of sample areas? Synthetic estimators

Stage II. Analysis/Adaptation

4. Use of explicit models

Model Building

- No single approach to model building.
- Fixed effects play a key role. Build the fixed part of the model as well as possible given the covariates available, before to focus on the inclusion of random effects.
- To choose the covariates for the fixed part of the model:
  - The triplet of estimators obtained in the previous stage can be useful
  - Simple measures: Use AIC and R^2 based on a linear model without random effects

Stage II. Analysis/Adaptation

4. Use of explicit models

Residual diagnostics

For the selected model use residual diagnostics

- QQ plots of residuals at different levels
- Influence diagnostics: Plots of Cook's distances
- Plot standardised residuals vs fitted values - Heteroscedasticity
- Plot standardised residuals vs design weights - Informative sampling

Stage II. Analysis/Adaptation

4. Use of explicit models

Adaptations

If the residual diagnostics indicate violation of model assumptions. Adapt the model

- Explore the use of transformations. Deciding on appropriate transformations is not straightforward, but offers a possible avenue for improving the model
- Use robust methods as an alternative to transformations (Chambers & Tzavidis, 2006; Ghosh et al., 2008; Sinha & Rao, 2009; Chambers et al., 2014; Dongmo Jiongo et al., 2013)
- Use non-parametric models (Opsomer et al., 2006; Ugarte et al., 2009)
- Elaborate the random effects structure e.g. include spatial structures (Pratesi & Salvati, 2008; Schmid & Münnich, 2014)
- Consider extensions to two-fold models (Morales et al., 2015)

Stage II. Analysis/Adaptation

3. Initial estimates. EDOMEX

In EDOMEX, direct/composite estimation is only possible for 58 municipalities. Even in those cases, most municipalities have small/moderate sizes.

4. Use of explicit models. EDOMEX

- Continuous outcomes: Unit-level nested error regression model (Battese et al., 1988) - BHF model

Mixed effects predictors were used for the areas in the sample and synthetic ones for out of sample areas.
If only area level data were available, a Fay-Herriot model (Fay & Herriot, 1988) could be used. However, the feasibility of the estimation of percentiles or complex indicators in this case is less clear.

Stage II. Analysis/Adaptation

4. Use of explicit models. EDOMEX

Some complex Income-based indicators

- FGT measures (Foster et al.,1984))

$$FGT(\alpha,t)=\sum_{i=1}^{N}\left(\frac{t-y_i}{t}\right)^{\alpha}\mathbb{1}(y_i\leq t)$$

$\alpha=0$ - Head Count Ratio; $\alpha=1$ - Poverty Gap

- The Gini coefficient

$$Gini=\frac{N+1}{N}-\frac{2\sum_{i=1}^{N}(N+1-i)y_{(i)}}{N\sum_{i=1}^{N}y_{(i)}}$$

- Quintile Share Ratio

$$QSR_{80/20}=\frac{\sum_{i=1}^{N}[y_i\mathbb{1}(y_i>q_{0.8})]}{\sum_{i=1}^{N}[y_i\mathbb{1}(y_i\leq q_{0.2})]}$$

Stage II. Analysis/Adaptation

4. Use of explicit models. EDOMEX

SAE methodologies for complex Income-based indicators

- The World Bank Approach (Elbers et al., 2003)
- The EBP Approach (Molina & Rao, 2010)
- The M-Quantile Approach (Marchetti et al., 2012 ; Chambers & Tzavidis, 2006)
- EBP based on normal mixtures (Elbers & Van der Weidel, 2014; Lahiri and Gershunskaya, 2011)
- MvQ methods based on Asymmetric Laplace distribution (Tzavidis et al., 2015)

Stage II. Analysis/Adaptation

4. Use of explicit models. EDOMEX

The EBP Method (under normality)

Point of departure: Unit-level Mixed effects model

$$y_{ik} = \boldsymbol{x}_{ik}^T\boldsymbol{\beta} + u_k + \epsilon_{ik}, u_k \sim N(0, \sigma_u^2); \epsilon_{ik} \sim N(0, \sigma_e^2)$$

Summary of the Method

- Use sample data to estimate β, σ_u^2, σ_ϵ^2, γ_k
- Generate $u_k^* \sim \ N(0, \hat{\sigma}_u^2(1-\gamma_k))$ and $\epsilon_{ik}^* \sim \ N(0, \hat{\sigma}_\epsilon^2)$

$$y_{ik}^* = \mathbf{x}_{ik}^T\hat{\boldsymbol{\beta}} + \hat{u}_k + u_k^* + \epsilon_{ik}^*$$

- Calculate the indicator of interest using the y_{ik}^*.

Micro-simulation of a synthetic population. Repeat the process L times.

Stage II. Analysis/Adaptation

4. Use of explicit models. Adaptation. EDOMEX.

Use of Transformations for the EBP method

- Molina & Rao (2010) use a logarithmic transformation
- Alternative 1 (Molina, 2015) use a logarithmic transformation with shift: $log(y_{ik} + s)$
- Alternative 2 (Rojas et al., 2015; Gurka et al., 2006): Box-Cox-Transformations under the linear mixed model

$$y^*_{ik}(\lambda) = \begin{cases} \frac{(y_{ik}+s)^\lambda - 1}{\alpha^{\lambda-1}\lambda}, & \lambda \neq 0 \\ \alpha \log(y_{ik} + s), & \lambda = 0 \end{cases},$$

for $y_{ik} > -s$ and α is the geometric mean of y_{ik}. Optimal power transformation parameter λ is estimated by ML

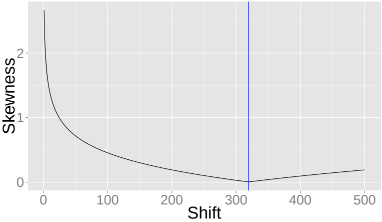
Stage II. Analysis/Adaptation

4. Use of explicit models. Adaptation. EDOMEX

Log-Shift transformation (Molina, 2015)

- $y_{ij}^{*} = log(y_{ij} + s)$, with s, the shift parameter

- Find s that makes skewness of the residuals close to 0



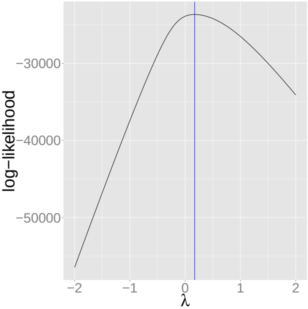
Stage II. Analysis/Adaptation

4. Use of explicit models. Adaptation. EDOMEX

Box-Cox transformation (Rojas et al., 2015; Gurka et al., 2006)

$$y^*_{ik}(\lambda) = \begin{cases} \frac{(y_{ik}+s)^\lambda - 1}{\alpha^{\lambda-1}\lambda}, & \lambda \neq 0 \\ \alpha \log(y_{ik}+s), & \lambda = 0 \end{cases}$$

for $y_{ik} > -s$ and α is the geometric mean of y_{ik}.

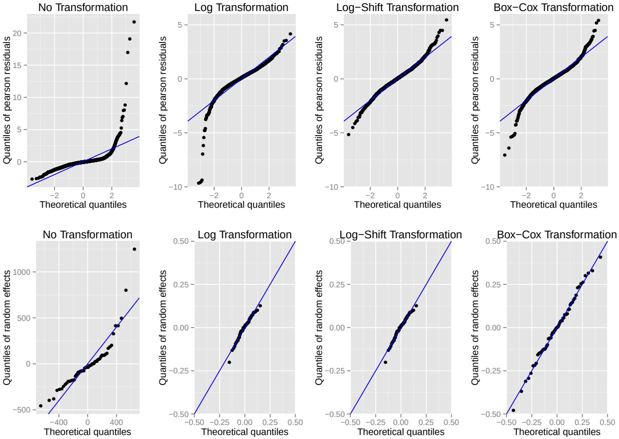


- Define a grid of λ values
- Optimal power transformation parameter λ obtained by the best fitting model within this grid

Stage II. Analysis/Adaptation

4. Use of explicit models. Adaptation. EDOMEX

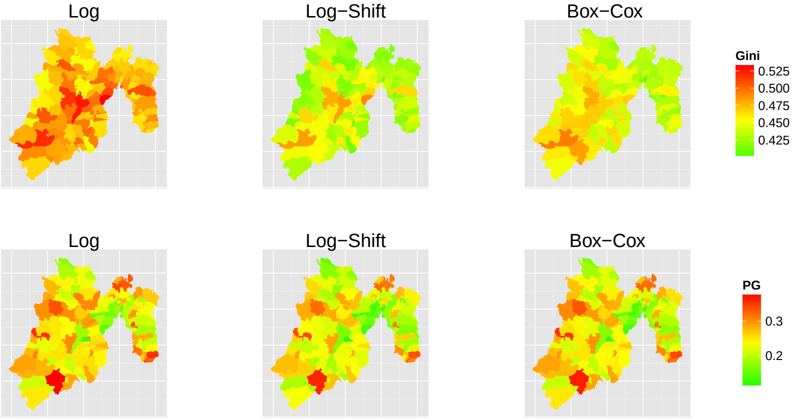
Residual diagnostics



Transformation	Original	Log	Optimal Shift	Box-Cox
R^2	0.32	0.45	0.52	0.50

Stage II. Analysis/Adaptation

4. Use of explicit models. Adaptation. EDOMEX



Choice of transformation possibly important for parameters involving the whole distribution. Gini more sensitive than PG

Stage III. Evaluation

5. MSE estimation
6. Model and Design based evaluation
7. Further evaluation tasks

Stage III. Evaluation

5. MSE estimation

- For Direct estimators, quality evaluation is commonly performed via variance estimation. In the case of small sample sizes, though, such estimates can be very unstable.
- Indirect SA estimates, in general, have smaller variances but can show bias. MSE estimation is necessary.
- For indicators such as totals, means or proportions, analytic MSE expressions are available (Prasad & Rao, 1990; Rao, 2003; Chambers et al., 2011)
- For more complex indicators, we increasingly rely on computer intensive methods. Bootstrap has become common in SAE application.
  - Parametric bootstrap (Hall & Maiti, 2006; Sinha & Rao, 2009)
  - Non-parametric/semi-parametric bootstrap (Correa & Pfeffermann, 2012; Chambers & Chandra, 2013; Mokhtarian & Chambers, 2013; Dongmo Jiongo & Nguimkeu, 2014)

Stage III. Evaluation

6. Model and Design based evaluation

Two complementary evaluation tools:

- Model-based evaluation:
  - Uses synthetic data generated under a model
  - Sampling is performed repeatedly from the population generated in each Monte-Carlo round
  - Useful for evaluating performance and sensitivity of new methods under different assumptions
- Design-based evaluation:
  - Uses Frame data (census data, for instance) or Synthetic data preserving the survey characteristics
  - Sampling is performed repeatedly from a fixed population
  - Useful for comparing different methods in a particular case

Stage III. Evaluation

7. Further evaluation tasks

- Compare SA estimates to direct estimates. Direct estimates are unstable but unbiased. Check for systematic departures from them: Bias, Over shrinkage.
- Compare aggregates of the SA estimates to the corresponding direct estimates
- Compare SA estimates to external data
- Evaluate estimates by consulting with local experts

Stage III. Evaluation

6. Model and Design based evaluation. EDOMEX

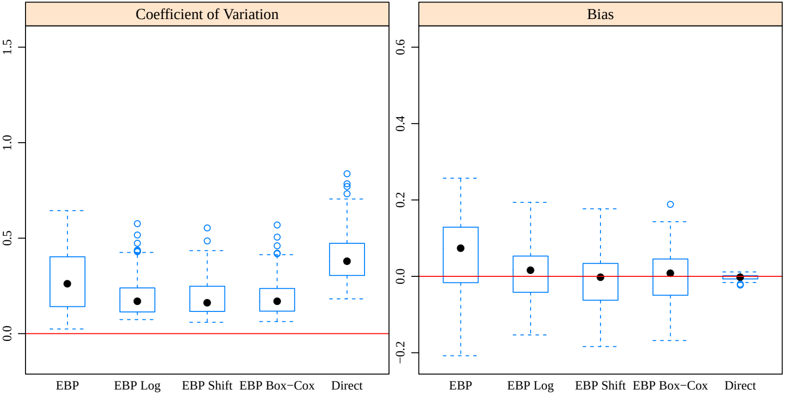
Design-based evaluation

- Two income variables are available in the survey.
- The target variable is available only on the survey. *Earned per capita income from work* is also available on the Census micro data.
- Target indicators Gini, Head Count Ratio, Poverty Gap, Quintile Share Ratio
- Setup
  - Design-based simulation with 500 MC-replications repeatedly drawn from EDOMEX Census
  - 6 covariates used leading to a R^2 around $40 - 50\%$
  - Unbalanced design leading to a sample size of $n = 2195$ ($min = 8$, $mean = 17.6$, $max = 50$)
  - Sampling from each municipality
  - Modification: More realistic to have some areas with 0 sample

Stage III. Evaluation

6. Model and Design based evaluation. EDOMEX

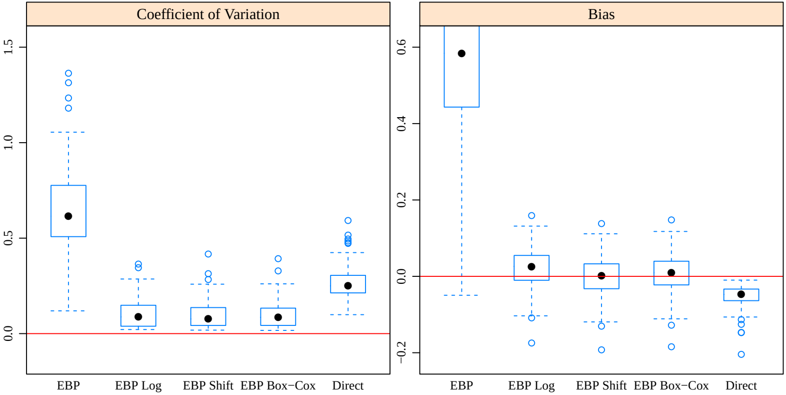
Design-based evaluation. Results: Head Count Ratio



Stage III. Evaluation

6. Model and Design based evaluation. EDOMEX

Design-based evaluation. Results: Gini



Related topics

Software availability

- Code for the majority of SAE methods is written in *R*. Open source. Easy to access, modify and extend
- Some attempts to collect code in a single place
    - SAMPLE PROJECT - Deliverable 13 *http://www.sample-project.eu/en/the-project/deliverables-docs.html*
    - SAE package in *R* (Molina & Marhuenda, 2015)
    - *saeSim* for setting up simulations in SAE (Warnholz & Schmid, 2015)
    - More methods will appear in packages in the near future
- SAE research community has a culture of sharing code. Ask authors of papers to provide code if not already available
- Widespread use of open access code promotes better understanding and validation of methods

Thank you

http://www.ncrm.ac.uk/research/ISAEM/