The role of models in model-assisted and model-dependent estimation for domains and small areas

Risto Lehtonen
University of Helsinki
risto.lehtonen@helsinki.fi

Workshop on Survey Sampling Theory and Methodology
Ventspils, Latvia
August 24-28, 2006

Outline

- PART I
  Estimation for domains and small areas:
  GREG and EBLUP methodology

- PART II
  Results on properties (bias and accuracy) of GREG and EBLUP estimators of totals for domains

PART I
Estimation for domains and small areas: GREG and EBLUP methodology

Some key terminology

- Population subgroup = Domain
  - NUTS regions, demographic subgroups…
  - Usually assumed non-overlapping
- Small area
  - Domain whose sample size is small (even zero)
- Domain estimation
  - Estimation of statistics (totals, means, proportions…) for domains
- Small area estimation SAE
  - Estimation of statistics for small domains
  - SIE – Small Island Estimation (Canary Islands)!
- Generalized regression GREG estimator
  - Model-assisted estimator
  - Assisting models
    - Fixed-effects model
    - Mixed model with fixed effects and random effects
- Synthetic estimator SYN
  - Model-dependent estimator
  - Underlying model
    - Fixed-effects model
- Empirical best linear unbiased predictor EBLUP
  - Model-dependent estimator
  - Underlying model
    - Mixed model with fixed effects and random effects

Background

- **World-wide trend**
  - Increasing need in society for official statistics for regional and other domains
    - Labour market, Economy, Demography, Welfare and health, Environment
- **SAIPE – Small Area Income & Poverty Estimates**
  - U.S. Census Bureau
- **EURAREA Project** (2001-2004)
  - Adaptation of model-dependent SAE methods into the European context
- **SAE in conferences**
  - Warsaw (1992)
  - Riga (1999) ISI Satellite
  - Berlin (2003) ISI Session
  - Sydney (2005) ISI Session
  - Jyväskylä (2005) SAE I
  - Pisa (2007) SAE II
  - Spain (2009) SAE III?
- **Statistics in Transition Journal, SAE papers**
  - December 2005
  - March 2006
- **Recent SAE books**
  - Rao (2003)
  - Longford (2005)

Approaches to be discussed

- **Design-based model-assisted methods**
  - Family of **generalized regression GREG** type estimators
    - Särndal, Swensson and Wretman (1992)
    - Särndal (1996)
    - Estevao and Särndal (1999, 2004)
    - Lehtonen and Veijanen (1998, 1999)
    - Lehtonen, Särndal and Veijanen (2003, 2005, 2006)
- **Model-dependent methods**
  - Family of **Synthetic SYN** type estimators
  - Family of **Empirical Best Linear Unbiased Predictor EBLUP** type estimators
    - Ghosh (2001)
    - Rao (2003)
    - Longford (2005)
- **Parameters to be estimated**
  - **Totals for domains**

Two different domain structures

- **Planned domain structure**
  - Domains are defined as strata in the sampling design
    - Domains are treated as independent subpopulations (strata)
    - Domain sample sizes are fixed in the sampling design
    - Stratification for the domain structure is an efficient option!
- **Unplanned domain structure**
  - Domain structure is not a part of the sampling design
    - Domain sample sizes are random variables
      - Extra variation due to randomness must be taken into account in variance estimation
    - Common situation in practice

Components of estimation procedure

- **(1) Sample survey data**
  - Unit-level sample survey data
  - Data are collected with a given (simple or complex) sampling design
  - Here we discuss
    - SRSWOR
    - PPS
  - Measurement
    - Study variables *y*
- **(2) Auxiliary x-data from the population**
  - Unit-level population data
    - Covariates **x**
    - Domain membership data
    - Sampling design identifiers
- **(3) Micro-merging**
  - Sample survey data and auxiliary data are merged at the unit level
    - Use of PIN:s and similar unique identifiers

Components of estimation procedure

- **(4) Model choice**
  - Choice of assisting model for GREG
  - Choice of underlying model for SYN and EBLUP
  - Here we discuss
    - Linear and logistic fixed-effects models
    - Linear and logistic mixed models
    - Members of generalized linear mixed models (GLMM) family
- **(5) Choice of estimator of domain totals**
  - Design-based model-assisted estimators
    - GREG family estimators
  - Model-dependent estimators
    - Synthetic SYN and EBLUP family estimators
- **(6) Estimation phase**
  - Point estimates
  - Variance and MSE estimation
  - Diagnostics

More on model choice

Model choice depends on the type of response variable *y*

- Continuous response variables
  - Linear (mixed) models
- Binary response variables
  - Binomial logistic (mixed) models
- Polytomous response variables
  - Multinomial logistic (mixed) models
- Count responses
  - Poisson (mixed) models

- Generalizd linear mixed models (GLMM:s)
  - McCullogh and Searle (2001)

$$E_m(y_k \mid \mathbf{x}_k, \mathbf{u}_d) = f(\mathbf{x}_k'(\boldsymbol{\beta} + \mathbf{u}_d))$$

- NOTE
  - The same GLMM:s can be incorporated both in design-based GREG estimation procedure and in model-dependent SYN and EBLUP estimation procedures
    - The role of model is different!

Estimation of model parameters

- Linear fixed-effects models
  - OLS Ordinary least squares
  - WLS Weighted least squares
- Logistic fixed-effects models
  - ML Maximum likelihood
  - PML Pseudo maximum likelihood (weighted)
- Linear and logistic mixed models
  - GLS Generalized least squares
  - GWLS Generalized weighted least squares
  - REML Restricted (residual) maximum likelihood
  - Pseudo REML (weighted)

Estimation of model parameters: Tools

- SAS for fitting GLMM:s
  - Proc REG for linear fixed-effects models
  - Proc MIXED for linear mixed models
  - Proc LOGISTIC for logistic fixed-effects models
  - Proc GENMOD for generalized linear models
  - Proc GLIMMIX for generalized linear mixed models
  - Proc NLMIXED for nonlinear mixed models
- R software functions for fitting GLMM:s
  - `lme` linear mixed models
  - `nlme` nonlinear mixed models

Population frame and parameters

$U = \{1, 2, ..., k, ..., N\}$ Population (fixed, finite)

$U_1, ..., U_d, ..., U_D$ Domains of interest (non-overlapping)

$Y_d = \sum_{U_d} y_k, d = 1, ..., D$ Target parameters (domain totals)

$\mathbf{x}_k = (x_{1k}, ..., x_{pk})'$ Auxiliary variable vector

$I_{dk} = 1$ if $k \in U_d$ Domain membership indicators,

$I_{dk} = 0$ otherwise $d = 1, ..., D$

NOTE: We assume that the vector value $\mathbf{x}_k$ and domain membership are known for every population unit $k \in U$

Sampling design

Sampling designs:

Simple random sampling without replacement SRSWOR

Systematic PPS with sample size n

s Sample from U

$s_d = s \cap U_d$ Random part of s falling in domain d

$\pi_k = n / N$ Inclusion probability for $k \in U$ in SRSWOR

$\pi_k = n \frac{x_k}{\sum_{k \in U} x_k}$ Inclusion probability for $k \in U$ in PPS

$a_k = 1/\pi_k$ Sampling weight for $k \in s$

We observe response variable values y_k for $k \in s$

Linear models

Linear fixed-effects models

$$y_k = \beta_0 + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} + \varepsilon_k$$

$$y_k = \beta_{0d} + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} + \varepsilon_k,\ d = 1,\ldots,D$$

Linear mixed models

$$y_k = \beta_0 + u_d + \beta_1 x_{1k} + \ldots + \beta_p x_{pk} + \varepsilon_k$$

$$y_k = \beta_0 + u_{0d} + (\beta_1 + u_{0d}) x_{1k} + \ldots + \beta_p x_{pk} + \varepsilon_k,\ d = 1,\ldots,D$$

where β_j are fixed effects, $j = 1,\ldots,p$

u_{0d} are domain-specific random intercepts

u_{1d} are domain-specific random slopes

Generalized linear models (GLMM)

- Special cases
  - Linear mixed model

$$\mathrm{E}_m(y_k \mid \mathbf{u}_d) = \mathbf{x}'_k(\boldsymbol{\beta} + \mathbf{u}_d)$$

  - Multinomial logistic mixed model

$$\mathrm{E}_m(y_{ik} \mid \mathbf{u}_d) = \frac{\exp(\mathbf{x}'_k(\boldsymbol{\beta}_i + \mathbf{u}_{id}))}{1 + \sum_{r=2}^{m} \exp(\mathbf{x}'_k(\boldsymbol{\beta}_r + \mathbf{u}_{rd}))}$$

Estimators of domain totals

Model-assisted GREG estimators

$$\hat{Y}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k)$$

Model-dependent SYN estimators

$$\hat{Y}_{dSYN} = \sum_{k \in U_d} \hat{y}_k$$

Model-dependent EBLUP estimators

$$\hat{Y}_{dEBLUP} = \sum_{k \in s_d} y_k + \sum_{k \in U_d - s_d} \hat{y}_k$$

where $d = 1,\ldots,D$

Fitted values under linear models

Fitted values under linear fixed-effects models

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{1k} + \ldots + \hat{\beta}_p x_{pk}$$

$$\hat{y}_k = \hat{\beta}_{0d} + \hat{\beta}_1 x_{1k} + \ldots + \hat{\beta}_p x_{pk}, \quad d = 1,\ldots,D$$

Fitted values under linear mixed models

$$\hat{y}_k = \hat{\beta}_0 + \hat{u}_{0d} + \hat{\beta}_1 x_{1k} + \ldots + \hat{\beta}_p x_{pk}$$

$$\hat{y}_k = \hat{\beta}_0 + \hat{u}_{0d} + (\hat{\beta}_1 + \hat{u}_{1d}) x_{1k} + \ldots + \hat{\beta}_p x_{pk}, \quad d = 1,\ldots,D$$

NOTE: Fitted values are calculated for every $k \in U$

Fitted values under logistic models

For a binary y_k

Fitted values under logistic fixed-effects models

$$\hat{y}_k = \exp(\mathbf{x}'\hat{\boldsymbol{\beta}})/(1+\exp(\mathbf{x}'\hat{\boldsymbol{\beta}}))$$

Fitted values under logistic mixed models

$$\hat{y}_k = \exp(\mathbf{x}'(\hat{\boldsymbol{\beta}}+\hat{\mathbf{u}}_d))/(1+\exp(\mathbf{x}'(\hat{\boldsymbol{\beta}}+\hat{\mathbf{u}}_d)))$$

Variance estimation for GREG

Planned domain structure under SRSWOR

$$\hat{v}_{srs}(\hat{Y}_d) = N_d^2(1-\frac{n_d}{N_d})(\frac{1}{n_d})\sum_{k\in s_d}\frac{(\hat{e}_k-\bar{\hat{e}}_d)^2}{n_d-1}$$

where $\hat{e}_k = y_k - \hat{y}_k$ and $\bar{\hat{e}}_d = \sum_{k\in s_d}\hat{e}_k / n_d$

Unplanned domain structure under SRSWOR

$$\hat{v}_{srs}(\hat{Y}_d) = N^2(1-\frac{n}{N})(\frac{1}{n})p_d\hat{s}_{d\hat{e}}^2(1+\frac{q_d}{c.v_{d\hat{e}}^2})$$

where $p_d = n_d / n$ and $q_d = 1 - p_d$

$c.v_{d\hat{e}} = \hat{s}_{d\hat{e}} / \bar{\hat{e}}_d$ is sample coefficient of variation of residuals in domain d with $\hat{s}_{d\hat{e}}$ as the sample standard deviation of residuals in domain d

Estimation for domains and small areas: EURAREA tools

- **SAS/IML Macro EBLUPGREG**
  - GREG, SYN and EBLUP estimation of totals and means for domains and small areas using linear mixed models
  - Freeware
  - www.statistics.gov.uk/eurarea
  - Developed by Statistics Finland and University of Jyväskylä
- Models
  - Linear mixed models with area-specific random intercepts
  - Modelling of spatial correlations
    - Exponential decay model
  - Modelling of temporal correlations
    - Fixed time effect

PART II
Results on properties (bias and accuracy) of GREG and EBLUP estimators of totals for domains

Quality measures: Bias, precision, accuracy

Bias

$$\text{Bias}(\hat{Y}_d) = \text{E}(\hat{Y}_d) - Y_d$$

Precision

Measured by variance

$$\text{V}(\hat{Y}_d) = \text{E}(\hat{Y}_d - \text{E}(Y_d))^2$$

Accuracy

Measured by mean squared error MSE

$$\text{MSE}(\hat{Y}_d) = \text{E}(\hat{Y}_d - Y_d)^2$$

where $d = 1,\ldots,D$

Quality measures of estimators in Monte Carlo simulation experiments

- Bias
  - Absolute relative bias ARB (%)
- Accuracy
  - Relative root mean squared error RRMSE (%)

$$\text{ARB}(\hat{Y}_d) = \left| \frac{1}{K}\sum_{v=1}^{K} \hat{Y}_d(s_v) - Y_d \right| / Y_d$$

$$\text{RRMSE}(\hat{Y}_d) = \sqrt{\frac{1}{K}\sum_{v=1}^{K} (\hat{Y}_d(s_v) - Y_d)^2} \,/ Y_d$$

Known design-based properties of estimators for domain totals

	Design-based model-assisted methods GREG family	Model-dependent methods EBLUP and SYN families
Bias	Unbiased (approximately) by the construction principle	Biased Bias does not necessarily approach zero with increasing sample size
Precision (Variance)	Variance may be large for small domains Variance tends to decrease with increasing domain sample size	Variance can be small even for small domains Variance tends to decrease with increasing domain sample size
Accuracy (Mean Squared Error, MSE)	MSE = Variance (or nearly so)	MSE = Variance + Squared Bias Accuracy can be poor if the bias is substantial
Confidence intervals	Valid intervals can be constructed	Valid intervals not necessarily obtained

Properties of estimators Monte Carlo simulation designs

- **Experiment 1**
  - Binary response variable
  - Logistic mixed models
    (a) Artificial population
    (b) Generated LFS population
- **Accounting for domain differences**
  - By domain-specific fixed effects OR by random effects?
- **Experiment 2**
  - **Unequal probability sampling design (PPS)**
  - Continuous response variable
  - Linear mixed models
  - Artificial population
- **Accounting for sampling complexities**
  - By GREG family OR by EBLUP family estimators?

Experiment 1a) Artificial population Monte Carlo design

- Artificial population
  - One million elements
  - *D* = 100 domains
- Population generating model
  - **Logistic mixed model**

$$P\{Y_k = 1\} = \frac{\exp(\eta_k)}{1+\exp(\eta_k)}$$

$$\eta_k = 1 + u_d + (1.5 + v_d)x_k$$

- Random effects
  - Bivariate normal distribution with mean zero and variances 0.25
  - Correlation between random effects

$$Corr(u_{0d}, u_{1d}) = -0.5$$

- Simulation
  - *K* = 1000 simulated samples of *n* = 10,000 with SRSWOR

Logistic SYN and logistic GREG estimators for a binary or polytomous response variable by model choice and estimator type (Lehtonen, Särndal and Veijanen 2005)

Model choice			Estimator type	
Model abbreviation	**Model specification**	**Effect type**	**Model-dependent synthetic**	**Model-assisted generalized regression**
CC	Common intercepts Common slopes	Fixed effects	**LSYN-CC**	**LGREG-CC**
SC	Separate intercepts Common slopes	Fixed effects	**LSYN-SC**	**LGREG-SC**
		Fixed and random	**MLSYN-SC**	**MLGREG-SC**
SS	Separate intercepts Separate slopes	Fixed effects	**LSYN-SS**	**LGREG-SS**
		Fixed and random	**MLSYN-SS**	**MLGREG-SS**

Average ARB (%) and average RRMSE (%) of logistic SYN and logistic GREG estimators of totals of a binary response variable for the artificial population (Lehtonen, Särndal and Veijanen 2005)

Estimator	Average ARB (%)			Average RRMSE (%)		
	Domain size class			Domain size class		
	Minor (20−69)	Medium (70−119)	Major (120+)	Minor (20−69)	Medium (70−119)	Major (120+)
SYN estimators						
LSYN-CC	**45.5**	**36.0**	**29.5**	45.7	36.1	29.7
LSYN-SC	1.1	0.6	0.4	42.6	24.0	16.1
MLSYN-SC	**20.9**	9.8	4.7	31.1	20.1	14.3
LSYN-SS	1.1	0.5	0.4	43.8	24.2	16.1
MLSYN-SS	**20.7**	9.6	4.6	31.3	20.2	14.4
GREG estimators						
LGREG-CC	0.1	0.6	0.4	43.6	24.3	16.3
LGREG-SC	1.1	0.6	0.4	42.6	24.0	16.1
MLGREG-SC	1.0	0.6	0.4	41.4	23.8	16.0
LGREG-SS	1.1	0.5	0.4	43.8	24.2	16.1
MLGREG-SS	1.1	0.5	0.4	41.4	23.8	16.0

Experiment 1b) Generated LFS population: Monte Carlo design

- Binary response *y*
  - ILO unemployed (0/1)
- Auxiliary x-variables from registers:
  - Sex, Age, Area (NUTS2)
  - Unemployed jobseeker indicator Reg-UE (0/1)
- Generated LFS population
  - N = 3 million units duplicated from the LFS data
- Domains of interest
  - D = 85 NUTS4 regions
- **Logistic models**
  - Fixed effects: Gender, Age, Reg-UE (plus interactions)
  - Domain-specific random intercepts (mixed model)
- 1000 independent samples with SRSWOR
  - n = 12,000 units

Average ARB (%) and average RRMSE (%) of logistic SYN and logistic GREG estimators of the number of unemployed of the generated LFS population (Lehtonen, Särndal and Veijanen 2005)

Estimator	Average ARB (%)			Average RRMSE (%)		
	Domain size class			Domain size class		
	Minor (20-69)	Medium (70-119)	Major (120+)	Minor (20-69)	Medium (70-119)	Major (120+)
SYN estimators						
LSYN-CC	**32.3**	**20.0**	**13.5**	32.5	20.3	13.9
LSYN-SC	2.7	1.0	0.3	43.0	28.7	16.3
MLSYN-SC	**17.2**	9.7	3.6	28.8	17.8	11.9
GREG estimators						
LGREG-CC	0.1	0.7	0.3	41.4	28.5	16.4
LGREG-SC	2.7	1.0	0.3	43.0	28.7	16.3
MLGREG-SC	0.8	0.7	0.3	40.6	28.1	16.2

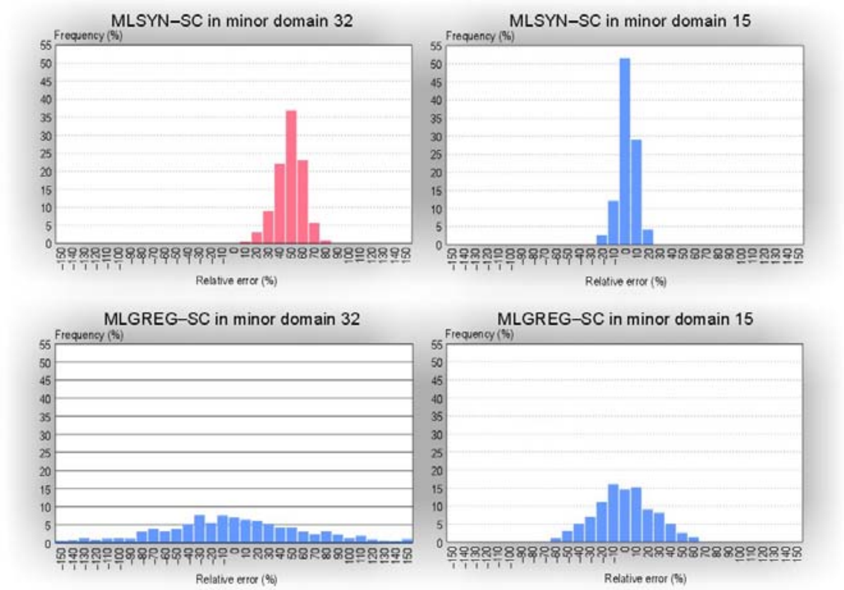
Experiment 1b) LFS population Closer examination for four domains

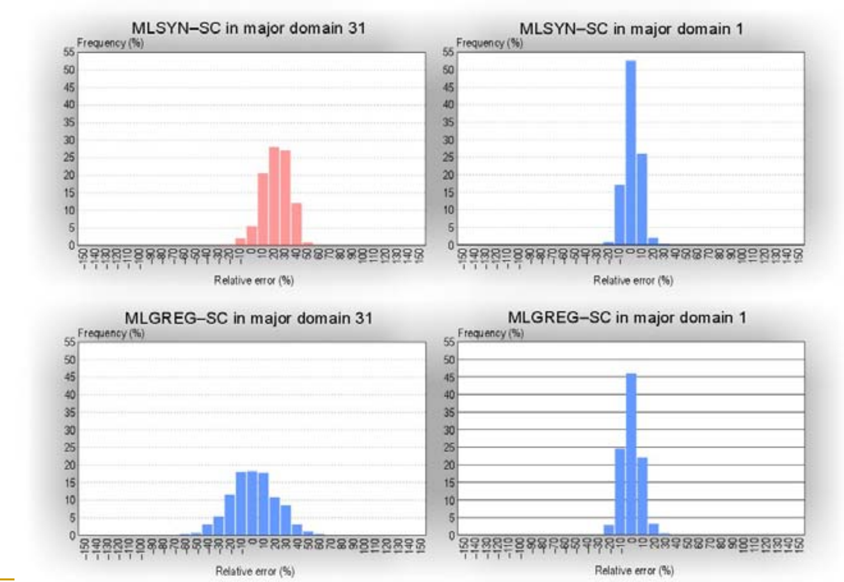
- Selected domains
  - Minor domains 32 and 15
  - Major domains 31 and 1
- Relative error of
  - MLSYN-SC
  - MLGREG-SC

(Lehtonen, Särndal and Veijanen 2005)

$$(\hat{Y}_d(s_v) - Y_d)/Y_d$$

	Minor domains		Major domains	
	Domain 32	Domain 15	Domain 31	Domain 1
Domain size	11689	16950	40699	299978
True domain total	466	1866	3263	23672
Means of estimated domain totals over simulations				
MLSYN-SC estimates	**693**	1825	**3946**	23968
MLGREG-SC estimates	468	1870	3308	23687


MLSYN–SC in minor domain 32
MLSYN–SC in minor domain 15
MLGREG–SC in minor domain 32
MLGREG–SC in minor domain 15
Frequency (%)
Relative error (%)


MLSYN–SC in major domain 31
MLSYN–SC in major domain 1
MLGREG–SC in major domain 31
MLGREG–SC in major domain 1
Frequency (%)
Relative error (%)

Experiment 2. Artificial population Monte Carlo simulation design

- Artificial finite population
  - One million elements
  - *D* = 100 domains
- Population generating model
  - Linear mixed model

$$y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$$

$$\beta_0 = 1 \quad \beta_2 = 1.5$$

$$\beta_1 = 2 \quad \delta_u^2 = 0.25$$

- *K* = 1000 systematic PPS samples
  - *n* = 10,000 elements
- Sampling weights vary between 50 and 600
- Correlations

$$corr(y, x_1) = 0.779$$

$$corr(y, x_2) = 0.607$$

$$corr(x_1, x_2) = -0.001$$

Two questions

- Accounting for sampling complexities under PPS sampling
  - By inclusion of sampling weights in the estimation procedure of model parameters

    OR
  - By inclusion of the PPS size variable in the model
- Accounting for domain differences
  - By domain-specific fixed intercepts

    OR
  - By domain-specific random intercepts
- How do GREG and EBLUP compare for these choices?

Estimators of domain totals

Schematic presentation of the model-dependent and model-assisted estimators of domain totals for a continuous response variable by model choice and estimator type.

Model choice and estimation procedure				Estimator type	
Model abbreviation	**Model specification**	**Effect type**	**Use of weights in estimation of model parameters**	***Model-dependent estimators***	***Model-assisted estimators***
CC	Common intercepts Common slopes	Fixed effects	No	SYN-CC	Not applicable
			Yes	Not applicable	GREG-CC
SC	Separate intercepts Common slopes	Fixed effects	No	SYN-SC	Not applicable
			Yes	Not applicable	GREG-SC
		Fixed and random	No	EBLUP-SC	Not applicable
			Yes	EBLUPW-SC	MGREG-SC

NOTE: In SYN, weights are ignored in the estimation procedure by default.
In GREG, weights are incorporated in the estimation procedure by default.

Estimators of domain totals

- **GREG-CC and GREG-SC**
  - Fixed-effects models
  - Estimation with WLS
- **MGREG-SC**
  - Mixed model
  - Estimation with GWLS and a pseudo (weighted) REML
- NOTE: Weights are incorporated in GREG estimation procedures by default
- **In the simulation experiment, a total of 15 models and 27 estimators were considered**

- **SYN-CC and SYN-SC**
  - Fixed-effects models
  - Estimation with LS
  - Weights are ignored
- **EBLUP-SC**
  - Mixed model
  - Estimation with GLS and REML
  - Weights are ignored
- **EBLUPW-SC**
  - Mixed model
  - Estimation with GWLS and a weighted modification of REML
  - **Weights are included**

Robustness of estimators to model choice: Means of average ARB (%) and average RRMSE (%) figures of model-dependent and model-assisted estimators of domain totals (number of estimators in parenthesis).

Estimator	Means of average ARB (%) figures			Means of average RRMSE (%) figures		
	Domain size class			Domain size class		
	Minor (20–69)	Medium (70–119)	Major (120+)	Minor (20–69)	Medium (70–119)	Major (120+)
Model-dependent estimators						
SYN (8)	12.1	12.3	11.7	13.4	13.0	12.1
EBLUP (7)	8.3	8.3	7.7	8.7	8.7	8.1
All (15)	10.3	10.4	9.8	11.2	11.0	10.2
Model-assisted estimators						
GREG (8)	0.4	0.2	0.1	8.8	5.1	3.7
MGREG (4)	0.2	0.1	0.1	8.7	5.1	3.6
All (12)	0.3	0.2	0.1	8.8	5.1	3.6

Average ARB (%) and average RRMSE (%) of model-assisted GREG type estimators of domain totals.

Model and estimator	**Average ARB (%)**			**Average RRMSE (%)**		
	Domain size class			Domain size class		
	Minor (20-69)	Medium (70-119)	Major (120+)	Minor (20-69)	Medium (70-119)	Major (120+)
Model A1 $y_k = \beta_{0d} + \varepsilon_k$						
GREG-SC	**1.4**	**0.5**	**0.3**	**13.7**	**8.1**	**5.7**
Model A2 $y_k = \beta_0 + u_d + \varepsilon_k$						
MGREG-SC	**0.2**	**0.2**	**0.1**	**13.7**	**8.1**	**5.6**
Model B1 $y_k = \beta_{0d} + \beta_1 x_{1k} + \varepsilon_k$						
GREG-SC	**0.2**	**0.1**	**0.0**	**7.8**	**4.6**	**3.2**
Model B2 $y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$						
MGREG-SC	**0.2**	**0.1**	**0.0**	**7.8**	**4.6**	**3.3**
Model C1 $y_k = \beta_{0d} + \beta_2 x_{2k} + \varepsilon_k$						
GREG-SC	**1.4**	**0.5**	**0.3**	**11.6**	**6.8**	**4.8**
Model C2 $y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$						
MGREG-SC	**0.2**	**0.1**	**0.1**	**11.6**	**6.8**	**4.7**
Model D1 $y_k = \beta_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$						
GREG-SC	**0.0**	**0.0**	**0.0**	**1.7**	**1.0**	**0.7**
Model D2 $y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$ (Population generating model)						
MGREG-SC	**0.0**	**0.0**	**0.0**	**1.7**	**1.0**	**0.7**

Average ARB (%) and average RRMSE (%) of model-dependent EBLUP type estimators of domain totals.

Model and estimator	Average ARB (%)			Average RRMSE (%)		
	Domain size class			Domain size class		
	Minor (20-69)	Medium (70-119)	Major (120+)	Minor (20-69)	Medium (70-119)	Major (120+)
Model A $y_k = \beta_0 + u_d + \varepsilon_k$						
EBLUP-SC	**22.9**	**23.1**	**21.7**	**22.9**	**23.3**	**21.8**
EBLUPW-SC	**3.7**	**3.5**	**3.3**	**3.9**	**3.6**	**3.5**
Model B $y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$						
EBLUP-SC	**1.8**	**1.4**	**0.7**	**2.8**	**2.5**	**2.2**
EBLUPW-SC	**3.5**	**3.5**	**3.3**	**3.5**	**3.6**	**3.3**
Model C $y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$						
EBLUP-SC	**22.3**	**23.1**	**21.8**	**22.4**	**23.2**	**21.9**
EBLUPW-SC	**3.7**	**3.6**	**3.2**	**3.9**	**3.7**	**3.3**
Model D $y_k = \beta_0 + u_d + \beta_1 x_{1k} + \beta_2 x_{2k} + \varepsilon_k$ (Population generating model)						
EBLUP-SC	**0.3**	**0.1**	**0.0**	**1.3**	**0.8**	**0.6**

Concluding remarks

- **Model-assisted GREG family**
  - Bias remained negligible for all model choices
  - Double-use of the same auxiliary information appeared to be beneficial
    - Sampling design phase
    - Modelling phase
  - Mixed model formulation did not outperform fixed-effects model formulation
- **Model-dependent EBLUP family**
  - Bias can be large for a misspecified model
  - PPS design could be accounted for with two options
    - Inclusion of size variable into the model
    - Use of a weighted version of EBLUP
  - The squared bias component still dominated strongly the MSE

References: Books and reports

- Lehtonen R. and Djerf K. (eds). *Lecture Notes on Estimation for Population Domains and Small Areas*. Helsinki: Statistics Finland, Reviews 2001/5.
  - Ghosh M. Model-dependent small area estimation: theory and practice.
  - Särndal C.-E. Design-based methodologies for domain estimation.
- Lehtonen R. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys. Second Edition*. Chichester: Wiley. (Chapter 6)
- Longford N.T. (2005). *Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician.* New York: Springer.
- Rao J.N.K. (2003). *Small Area Estimation*. Hoboken, NJ: Wiley.
- Särndal C.-E., Swensson B. and Wretman J. (1992). *Model Assisted Survey Sampling*. New York: Springer

References: Journal articles

- Estevao V.M. and Särndal C.-E. (1999). The use of auxiliary information in design-based estimation for domains. *Survey Methodology*, 25, 213–221.
- Estevao V.M. and Särndal C.-E. (2004). Borrowing strength is not the best technique within a wide class of design-consistent domain estimators. *Journal of Official Statistics,* 20, 645–669.
- Lehtonen R., Särndal C.-E. and Veijanen A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology,* 29, 33–44.
- Lehtonen R., Särndal C.-E. and Veijanen A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition,* 7, 649–673
- Lehtonen R., Myrskylä M., Särndal C.-E. and Veijanen A. (2006). Model-assisted and model-dependent estimation for domains and small areas under unequal probability sampling. 9th International Vilnius Conference on Probability Theory and Mathematical Statistics, June 25–30, 2006.
- Lehtonen R. and Veijanen A. (1998). Logistic generalized regression estimators. *Survey Methodology*, 24, 51–55.
- Lehtonen R. and Veijanen A. (1999) Domain estimation with logistic generalized regression and related estimators. *Proceedings, IASS Satellite Conference on Small Area Estimation*, Riga, August 1999. Riga: Latvian Council of Science, 121–128.