HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Small area estimation by calibration and model-assisted methods with applications

Risto Lehtonen, University of Helsinki

Mini Course

University of Pisa

17 May 2017

Lecture topics

- Topic 1: Introduction
- Topic 2: Basic concepts and approaches
- Topic 3: Traditional (direct) estimators for domains
- Topic 4: Direct GREG and calibration
- Topic 5: Indirect GREG estimators
- Topic 6: Extended GREG and model-assisted calibration
- CASE STUDY 1: SAE in the SILC data
- CASE STUDY 2: EBLUP example

Topic 1 INTRODUCTION

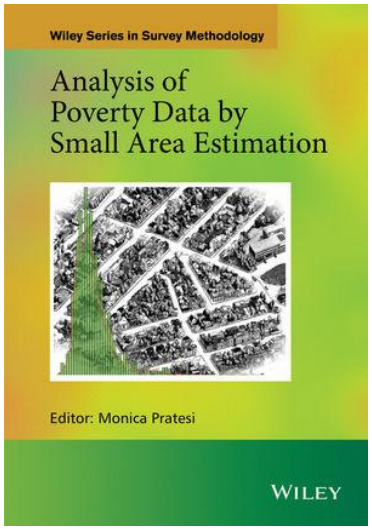
- **Introduction to small area estimation (SAE)**
  - Motivation:
    what is small area estimation?
    why SAE?
  - Estimation tasks in SAE
  - Main literature

Small area estimation: World-wide trend

- An increasing need in society for reliable statistics for regional and other population subgroups or domains, including sub-populations with small sample sizes
- SAE: Challenge for official statistics describing the society
  - SAIPE (U.S.) Allocation of federal state funds to small local areas based on model-based estimates obtained by SAE methods
- SAE: Challenge for scientific research
  - Good review on SAE research: Pfeffermann (2013)

Lively SAE Research under EU's Framework Programmes

- Small area estimation research under Framework Programmes (FPs) in Europe
- Actors: Universities & NSIs
  - EURAREA Project (2001-2004), EU/FP5
    AMELI Project (2008-2011), EU/FP7
    SAMPLE Project (2008-2011), EU/FP7
    InGRID Project (2013-2017), EU/FP7

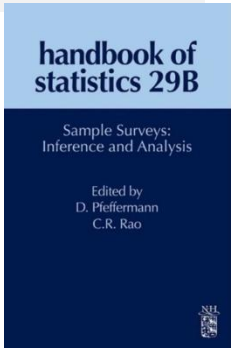


- Main results of AMELI and SAMPLE:
  - Pratesi M. (Ed.) (2016) Analysis of Poverty Data by Small Area Estimation. Chichester: Wiley.

Series of SAE Conferences

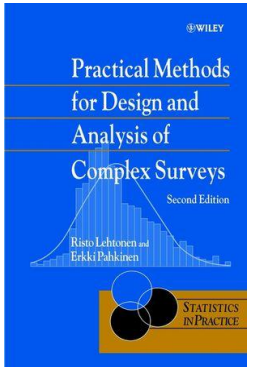
- EWORSAE European Working Group on Small Area Estimation http://sae.wzr.pl/
  - SAE1993 (Warsaw, Poland)
  - SAE2000 (Riga, Latvia)
  - SAE2005 (University of Jyväskylä, Finland)
  - SAE2007 (University of Pisa, Italy)
  - SAE2009 (University of M. Hernandez, Elche, Spain)
  - SAE2011 (University of Trier, Germany)
  - SAE2013 (Bangkok, Thailand)
  - SAE2014 (Poznan University of Economics, Poland)
  - SAE2015 (Santiago, Chile)
  - SAE2016 (Maastricht, The Netherlands)
  - SAE2017 (Paris, France)

Main materials for this course



Model-assisted SAE:

Lehtonen R. & Veijanen A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Sample Surveys: Inference and Analysis. Vol. 29B.* New York: Elsevier.

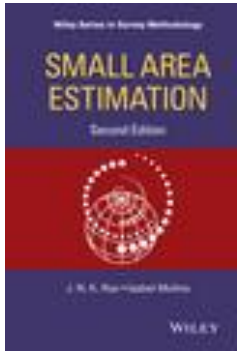


Lehtonen R. & Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys.* Second Edition. Chichester: John Wiley & Sons. Chapter 6.

Lehtonen R. & Veijanen A. Model-assisted methods to small area estimation of poverty indicators. In: Pratesi M. (Ed.) (2016). Analysis of Poverty Data by Small Area Estimation. Wiley.

Model-based SAE:

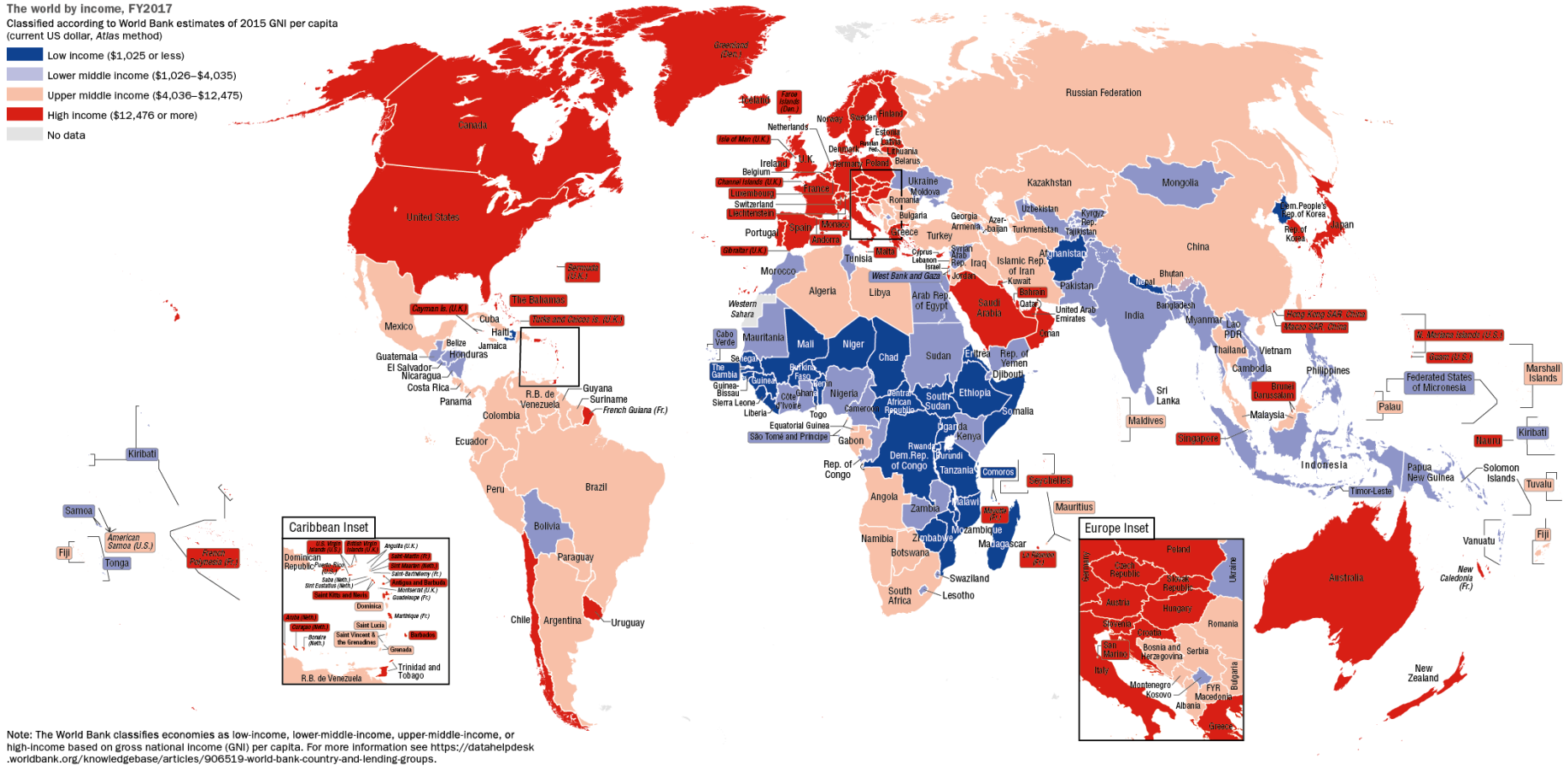
Rao J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons. (Second edition Rao & Molina 2015)



Estimation for domains and small areas

- **Domains of interest**
  - Well-defined (non-overlapping) population subgroups
  - Regional areas
  - Sex-age groups within regions
  - Grouping of enterprises into domains according to the type of industry
- **Estimation for domains**
  - Estimation of population quantities for the desired domains of interest
- **Small area estimation SAE**
  - Estimation for domains whose **sample size is small** or very small (even zero)
- **Alternative SAE definition** (Partha Lahiri):
- Small area = Domain of interest for which the sample size is not adequate to produce reliable **direct estimates**

EXAMPLE 1: The world by income



- http://data.worldbank.org/products/wdi-maps

EXAMPLE 2: Disposable income per capita by NUTS3 regions in the EU

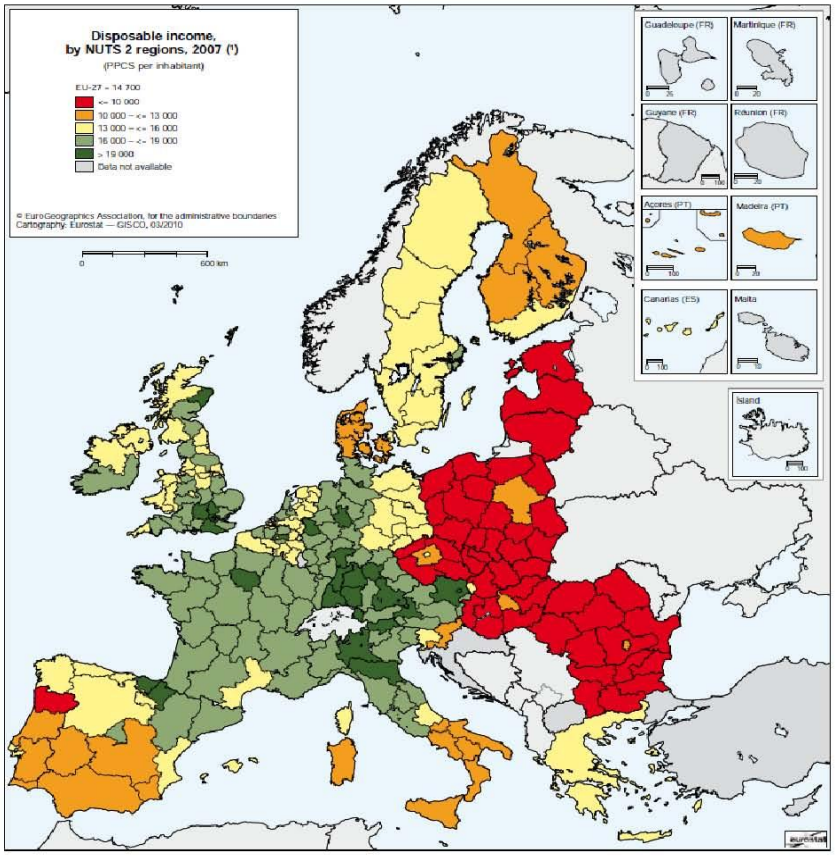


Figure 1. Disposable income by NUTS 2 regions in 2007 in the European Union

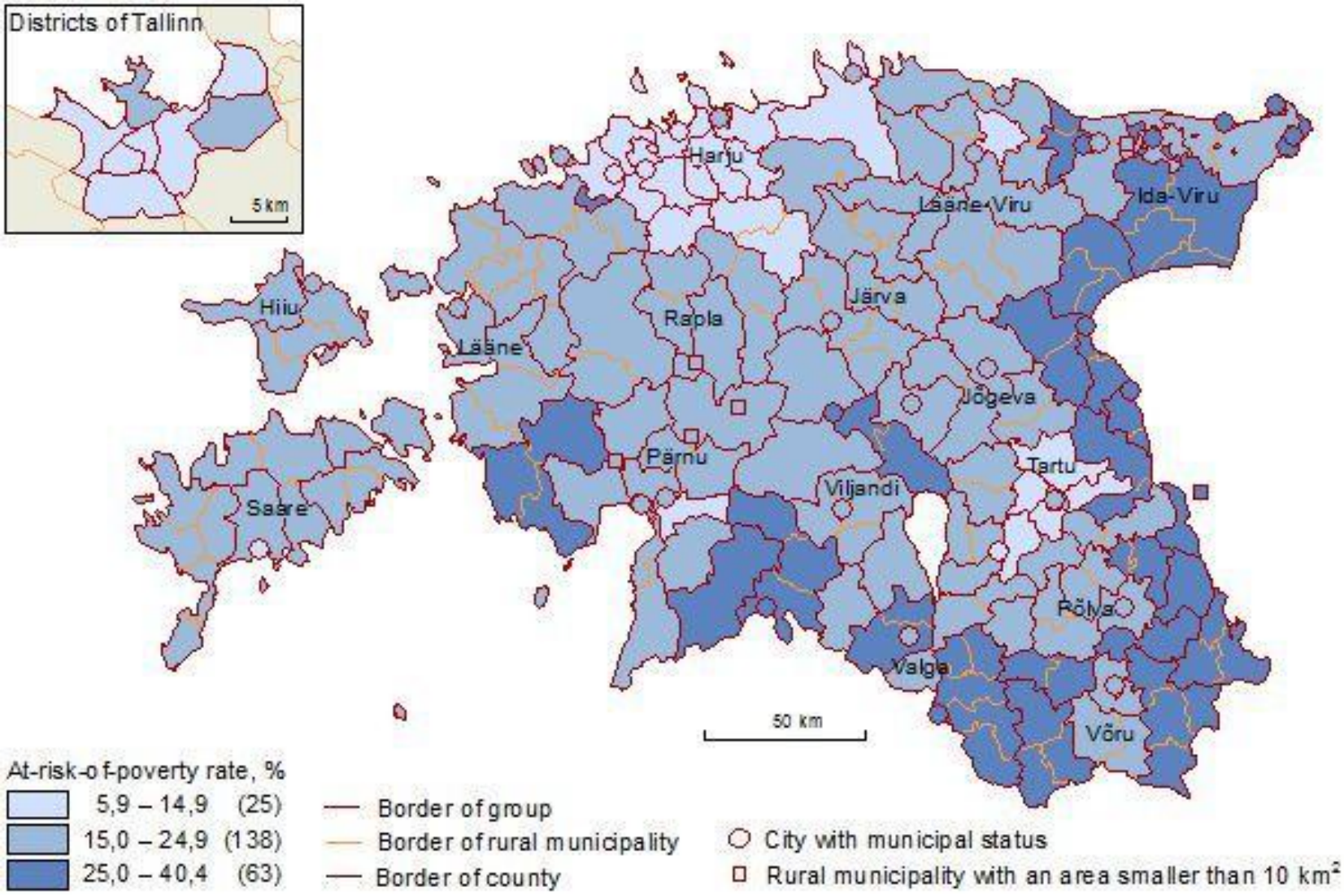
Source: Eurostat Regional Yearbook 2010, p.93, Section on Household Accounts. Information about the metadata is available at http://epp.eurostat.ec.europa.eu/cache/ITY_SDDS/EN/reg_ecohh_esms.htm

More recent: 2003-2013

http://ec.europa.eu/eurostat/tgm/mapToolClosed.do?tab=map&init=1&plugin=1&language=en&pcode=tgs00026&toolbox=types

EXAMPLE 3: Poverty map for Estonia

World Bank 2014 – Regional poverty rates based on SILC data



Components of typical estimation task - 1

- Domains of interest
  - Breakdown of population into sub-populations (areas, domains)
  - The number of domains of interest is usually large
- Study variable(s)
- Target parameters for the domains
  - Totals
  - Means
  - Ratios
  - Percentiles, medians
  - Poverty indicators
  - AROPE indicators, ...

Components of typical estimation task - 2

- Data sources
  - **Sample survey data source**
  - Unit-level values of study variable
  - **Auxiliary data sources**
  - Alternatives: Domain-level (area-level) aggregates of auxiliary variables or unit-level values of auxiliary variables x that can be merged with sample data at the unit level
  - NOTE: Availability depends on the statistical data infrastructure
- Statistical models
  - Example: Generalized linear mixed models GLMM family
  - Alternatives: Domain-level (area-level) models or unit-level models

Components of typical estimation task - 3

- Estimators of domain parameters
  - **Model-assisted design-based estimators**
  - Examples: Generalized regression (GREG) estimators and calibration estimators
  - **Model-based estimators**
  - Examples: Empirical best linear unbiased predictor (EBLUP) and empirical best predictor (EBP) type estimators
- Variance estimators, MSE estimators
- Computation, graphical illustration
- Quality assurance
- Publication

Topic 2 BASIC CONCEPTS & APPROACHES

- Two main SAE approaches:
  Design-based and model-based SAE
- Two different domain structures:
  Planned and unplanned domains
- Two different types of estimators for domains:
  Direct and indirect estimators
- “Borrowing strength”
- Simple examples

Approaches for domain estimation and SAE

- **Design-based approach**
- **Model-based approach**

- Additional variants
  - Bayesian methods: Empirical Bayes, Hierarchical Bayes
  - Poverty mapping: World Bank, Peter Lanjouw, Chris Elbers,…
  - PovMap Software
    http://iresearch.worldbank.org/PovMap/PovMap2/PovMap2Main.asp
  - Spatial microsimulation:
    Rahman A & Harding A. (2016) Small Area Estimation and Microsimulation Modeling. Chapman and Hall/CRC.

Main methods for domain estimation and SAE

- **Design-based methods**
  - Horvitz-Thompson (HT) estimator
  - Hájek estimator
  - Generalized regression (GREG) estimators
  - Model-free calibration estimators
  - Model-assisted calibration estimators
- **Model-based methods**
  - Synthetic (SYN) estimators
  - Empirical best linear unbiased predictor (EBLUP) estimators
  - Empirical best predictor (EBP) type estimators

Basic small area estimation approaches

A. Design-based direct estimation	B. Indirect estimation "Borrowing strength" from other domains by using models and auxiliary data (unplanned domains)		
Domains are considered as independent sub-populations (strata, planned domains)	B1. Design-based model-assisted estimation	B2. Model-based estimation	
		Unit-level models	Area-level models
Horvitz-Thompson Hajék Model-free calibration Direct GREG	Extended GREG Model-assisted calibration	SYN EBLUP EB	Fay-Herriot

Design-based approach

- The randomness is introduced by the **sampling design**
- Statistical properties (design bias, design accuracy) are evaluated under the sampling design
- **Examples of estimators**
- Horvitz-Thompson (HT)
- Model-free calibration methods
- Model-assisted methods e.g. generalized regression (GREG) assisted by linear model (Särndal et al. 1992)
- Overview: see Lehtonen & Veijanen (2009)

Model-based approach

- The randomness is introduced by an assumed **superpopulation model**
- Statistical properties (model bias, model accuracy) are evaluated under the model
- **Examples of estimators**
- Empirical best linear unbiased predictor (EBLUP) estimator with area-level model e.g. Fay-Herriot model
- Nested error linear regression model with unit-level data (Battese et al. 1988)
- Synthetic estimators
- Overview: see Datta (2009)

EXAMPLES of estimators under SRS

Model-based synthetic estimator of population total $t = \sum_{k \in U} y_k$

$$\hat{t}_{SYN} = \sum_{k \in U} \hat{y}_k$$

Design-based GREG estimator of population total $t = \sum_{k \in U} y_k$

$$\hat{t}_{GREG} = \sum_{k \in U} \hat{y}_k + \sum_{k \in S} a_k (y_k - \hat{y}_k)$$

$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_{1k}$, $k \in U$, are y-values predicted by linear model

$$y_k = \beta_0 + \beta_1 x_{1k} + \varepsilon_k, \; k \in U$$

x_{1k} are auxiliary variable values known for all $k \in U$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are LS estimates of beta parameters

$a_k = 1/\pi_k$ is design weight for element k in sample $s \subset U$

$\pi_k = n/N$ is SRS inclusion probability for element k in population U

NOTE: Role of sampling complexities

- **Design-based approach:**
- Estimators are constructed such that the properties of the sampling design are accounted for
  - Stratification, clustering, weighting
  - EXAMPLE: HT estimator

$$\hat{t}_{HT} = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in s} a_k y_k$$

π_k inclusion probability

$a_k = 1/\pi_k$ design weight

for element k in sample s

- **Model-based approach:**
- Accounting for sampling design properties is **not necessarily** an issue BUT **is possible if desired**

EXAMPLES:

  - Pseudo EBLUP (Rao 2003)
  - Mixed models in accounting for clustering & stratification
  - Incorporation of stratification variables in the model
  - NOTE: No consensus within statistical communities
  - NOTE: CASE STUDY 2

Key properties of estimators - 1

- **Key properties of design-based estimators**
  - (Nearly) design unbiased (by construction principle)
  - Models are used as assisting tools in inference
  - Estimators remain unbiased even under a wrong model
  - Accuracy can be good with a strong model
  - Accuracy can be poor if domain sample size is small
- **Key properties of model-based estimators**
  - Design biased (by construction principle)
  - Inference relies on the correctness of the model
  - Accuracy can be good with a strong model, even for small domains
  - Accuracy can be poor with an incorrect model, irrespective of domain sample size

Key properties of estimators - 2

Source: Lehtonen and Veijanen (2009)

Table 1
Design-based properties of model-assisted and model-dependent estimators for domains and small areas

	Design-based model-assisted methods	Model-dependent methods
	GREG and calibration estimators	Synthetic and EBLUP estimators
Bias	Design unbiased (approximately) by the construction principle	Design biased Bias does not necessarily approach zero with increasing domain 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 (MSE)	MSE = Variance (or nearly so)	MSE = Variance + squared bias Accuracy can be poor if the bias is substantial
Confidence intervals	Valid design-based intervals can be constructed	Valid design-based intervals not necessarily obtained

NOTE on the role of models

- The role of model differs in *model-assisted design-based* estimators and *model-based* estimators
  - **Model assisted design-based methods use models as assisting tools**
    - Benefit: design bias near to zero
    - Cost to be paid: poor accuracy for small domains
  - **Model-based methods rely solely on models**
    - Benefit: improved accuracy for small domains
    - Cost to be paid: the risk of nonzero design bias
  - NOTE: **Trade-off between bias and accuracy!**
  - NOTE: “All models are wrong but some are useful” (George Box 1978)

model

EXAMPLE 4. Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005): Does the matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition* 7, 649-673.

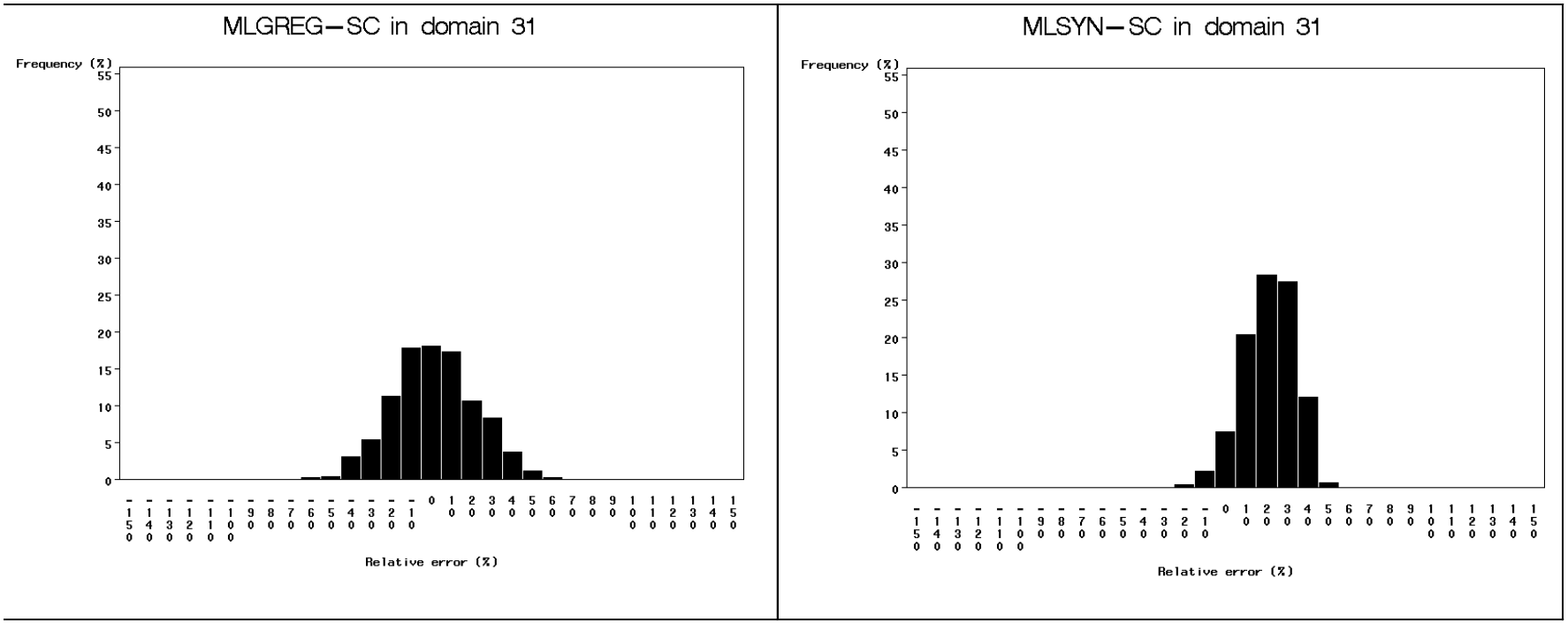


FIGURE 1 Distribution of relative error (%) of design-based MLGREG (left-hand side) and model-based MLSYN (right-hand side) estimators of domain totals of binary study variable in domain 31 of the generated LFS population. (Design-based simulation experiment, 1,000 independent simple random samples of 12,000 elements from population of three million elements and 84 domains)

Relative error of an estimator $\hat{t}_d$ for sample s_i, $i = 1,...,1000$, in domain *d* is defined as

$$\mathrm{RE}(\hat{t}_d) = (\hat{t}_d(s_i) - t_d) / t_d,\ d = 1,...,84$$

Lessons learned – EXAMPLE 4

- MLGREG: *design-based generalized regression* (GREG) estimator assisted by logistic mixed model
- MLSYN: *model-based synthetic* estimator using the same logistic mixed model formulation as GREG

- QUESTIONS:
- Which one of the two estimators indicates smaller design bias?
- Which one of the estimators indicates smaller design variance?

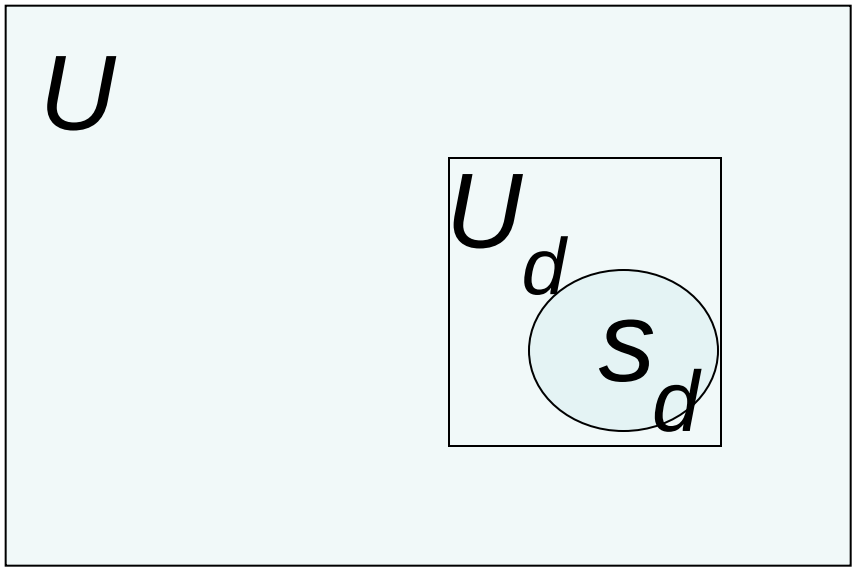
  - NOTE: Design bias refers to the difference between the expected value (or mean) of the distribution (in repeated sampling from the population) of the estimator and the true parameter value
  - Design variance refers to the spread of the distribution of the estimator around its expectation

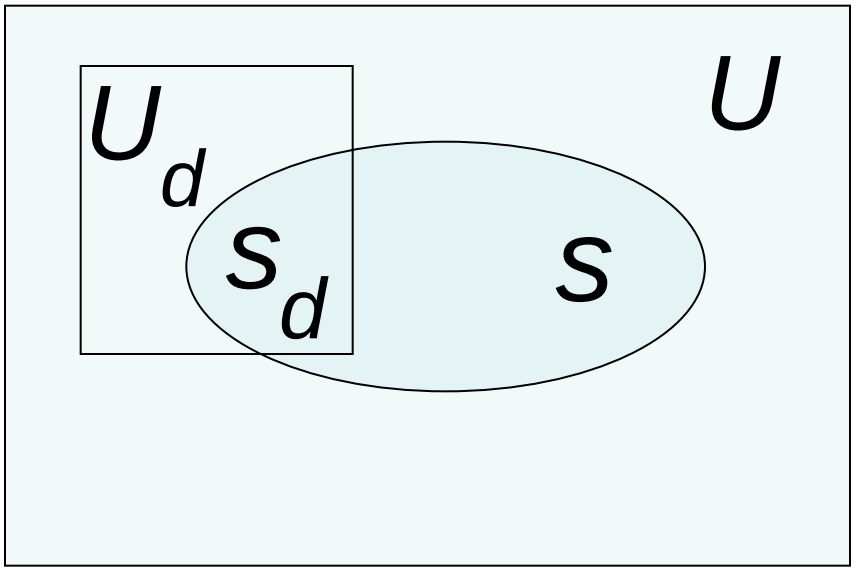
Important concepts

- **Type of domains of interest**
  - Planned domains / Unplanned domains
- **Type of domain estimator**
  - Direct / Indirect
- **Availability of auxiliary (population) data**
  - Unit-level / Aggregate-level (area-level)
- **Type of model**
  - Linear model / Non-linear model
  - Fixed-effects model / Mixed model
- **Accuracy measures**
  - Variance estimators / MSE estimators

Two main domain structures

- **Planned domains**
  - The most important domains are defined as **strata** in the sampling design (stratified sampling)
  - The strata are independent sub-populations
  - Domain sample sizes are fixed in advance
  - Domain sample sizes are controlled by allocation scheme
  - Small sample sizes can be avoided if desired
- **Unplanned domains**
  - Domain structure is not connected to the sampling design
  - Domain sample sizes are not fixed but are random
  - Small domain sample sizes can occur
  - **Typical in SAE practice**
  - NOTE: Similarity of unplanned domains with **post-stratifiction**, see Lehtonen & Pahkinen (2004) pp. 89-92





Planned domains

U Population

U_d Population domain d, $d = 1,...,D$

Domains = Strata

Several ($= D$) independent samples

Sample $s_d \subset U_d$ drawn in domain d

Sample size n_d is **fixed** by sampling design

Unplanned domains

U Population

A single sample s is drawn

$s \subset U$ Sample

U_d Population domain d, $d = 1,...,D$

$s_d = s \cap U_d$ Sample falling in domain d

Sample size n_d in domain d is **random**

Direct and indirect estimation

- **Direct estimation for domains**
  - *Direct* domain estimator uses values of the variable of interest *y* only from the time period of interest and only from units in the domain of interest
    (Federal Committee on Statistical Methodology, 1993)
  - Often in connection to *planned* domain structures
- **Indirect estimation for domains**
  - *Indirect* domain estimator uses values of the variable of interest *y* from a domain and/or time period other than the domain and time period of interest
  - Often in connection to *unplanned* domain structures

Domain type and estimator type 1

Domain type	Estimator type	
	Direct	**Indirect**
Planned	Typical set-up	More rarely
Unplanned	More rarely	Typical set-up

"Borrowing strength" in SAE

- *Indirect estimators* are attempting to "borrow strength" from other (similar) domains (spatial dimension) and/or from previous time points (temporal dimension)
- For domains with small sample size, this is a well justified goal – Why?
- The concept of "borrowing strength" is often used in model-based small area estimation
  - E.g. Rao & Molina (2015)
- Borrowing strength also is used for *design-based model assisted* estimators
  - E.g. Lehtonen & Veijanen (2009)
  - See EXAMPLE 11

EXAMPLE 5
Direct and indirect GREG

Assume continuous y-variable and one continuous auxiliary x-variable

Domains of interest U_d, $d = 1,...,D$

Assisting **linear fixed-effects models** in two cases:

a) Planned domains case: $y_k = \beta_d x_k + \varepsilon_k, \ k \in U_d, \ d = 1,...,D$

b) Unplanned domains case: $y_k = \beta x_k + \varepsilon_k, \ k \in U$

NOTE: Intercept parameters $\beta_{0d} = \beta_0 = 0$

NOTE: Models a) and b) are different. In what essential way?

For both domain types, let us construct a GREG estimator

of domain total of y-variable $t_d = \sum_{k \in U_d} y_k$, $d = 1,...,D$

a) Direct GREG estimator for domains

Assisting model: $y_k = \beta_d x_k + \varepsilon_k, \quad k \in U_d, \; d = 1,...,D$

By noting that $\hat{\beta}_d = \dfrac{\hat{t}_{dHT}}{\hat{t}_{dxHT}}$ and $\hat{y}_k = \hat{\beta}_d x_k$ we have:

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

$$= \hat{t}_{dHT} + \frac{\hat{t}_{dHT}}{\hat{t}_{dxHT}} (t_{dx} - \hat{t}_{dxHT})$$

$$= t_{dx} \times \frac{\hat{t}_{dHT}}{\hat{t}_{dxHT}}, \; d = 1,...,D$$

which is standard textbook form of *ratio estimator*

Why this GREG estimator is direct?

NOTE: Auxiliary information needed: x-totals t_{dx} for domains

b) Indirect GREG estimator for domains

Assisting model: $y_k = \beta x_k + \varepsilon_k, \quad k \in U$

By noting that $\hat{\beta} = \dfrac{\hat{t}_{HT}}{\hat{t}_{xHT}}$ and $\hat{y}_k = \hat{\beta} x_k$ we have

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

$$= \hat{t}_{dHT} + \frac{\hat{t}_{HT}}{\hat{t}_{xHT}} (t_{dx} - \hat{t}_{dxHT})$$

which is standard textbook form of *regression estimator*

Why this GREG estimator is indirect?

NOTE: Auxiliary information needed: x-totals t_{dx} for domains

Lessons learned – EXAMPLE 5

- Which one of the two GREG estimators would YOU prefer?
  a) Ratio estimator
  b) Regression estimator

- Why?

- Further , which one of the two GREG estimators, the ratio estimator or the regression estimator, aims at “borrowing strength” for domain *d* from other domains?

Topic 3 TRADITIONAL DIRECT ESTIMATORS FOR DOMAINS

- Definitions and notation
- Why totals are important?
- Estimation of domain totals for planned and unplanned domains
- Unconditional and conditional approach
- Horvitz-Thompson estimator
- Variance estimation – different options

Some definitions and notation

Fixed and *finite population* $U = \{1,2,...,k,...,N\}$

Inclusion probability: An observation k is included in a sample s with probability $\pi_k = P\{k \in s\}$

Design weight: $a_k = 1/\pi_k$

Sample membership indicator: $I_k = I\{k \in s\} = 1$ if $k \in s$, 0 otherwise

Expectation of sample membership indicator $E(I_k) = \pi_k$

Probability of including both elements k and l $(k \neq l)$:
$\pi_{kl} = E(I_k I_l)$ with inverse $a_{kl} = 1/\pi_{kl}$ $(a_{kl} = a_k$ when $k = l)$

The covariance of I_k and I_l is $Cov(I_k, I_l) = \pi_{kl} - \pi_k \pi_l$

Estimation of domain totals

Estimation of totals

$$t_d = \sum_{k \in U_d} y_k,\ d = 1,...,D$$

of variable of interest y for D non-overlapping domains

$U_d \subset U,\ d = 1,2,...,d,...,D,$

with (known or unknown) domain sizes N_d

NOTE: For unknown N_d an estimator

$\hat{N}_d = \sum_{k \in U_d} a_k = \sum_{k \in U_d} 1/\pi_k$ is often used

Why domain totals are important?

Totals are basic and the simplest descriptive statistics for continuous (or binary) study variables

Many other, more complex statistic are functions of totals

Domain ratio: $$R_d = \frac{t_{dy}}{t_{dz}} = \frac{\sum_{k \in U_d} y_k}{\sum_{k \in U_d} z_k}$$

Estimator: $$\hat{R}_d = \frac{\hat{t}_{dy}}{\hat{t}_{dz}} = \frac{\sum_{k \in s_d} a_k y_k}{\sum_{k \in s_d} a_k z_k}$$

Domain mean: $$\bar{y}_d = t_d / N_d$$

Estimator: $$\hat{\bar{y}}_d = \hat{t}_d / N_d \quad \text{or} \quad \hat{\bar{y}}_d = \hat{t}_d / \hat{N}_d$$

Estimation for planned domains

Sample s is divided into subsamples s_d, $d = 1,...,D$

Planned domains:
Stratified sampling with domains = strata

- The population domains U_d are taken as separate subpopulations i.e. strata
- Domain sizes N_d in domains U_d are assumed known
- Sample sizes n_d in domain samples $s_d \subset U_d$ are fixed in in the allocation scheme of the stratified sampling design
- **Standard estimators for the entire population are applicable for the domains as such, because the domains are taken as independent sub-populations**

NOTE: Sample allocation for planned domains

- Stratified sampling with a suitable *allocation scheme* is advisable in practical applications, in order to obtain control over the domain sample sizes

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. The American Statistician 42, 174-177.

Choudhry, G.H., Rao, J.N.K. & Hidiroglou, M.A. (2012). On sample allocation for effective domain estimation. Survey Methodology 38, 23-29.

Falorsi, P.D. & Righi, P. (2008). A balanced sampling approach for multi-way stratification for small area estimation. Survey Methodology 34, 223–234.

Molefe W.B. & Clark R.G. (2015). Model-assisted optimal allocation for planned domain using composite estimation. Survey Methodology 41, 377–387.

Estimation for unplanned domains

Unplanned domains: A single sample *s* of size *n* is drawn from population *U*

Domain samples are $s_d = s \cap U_d$

RECALL: Domain sample sizes n_d are considered *random*

Extended variable of interest y_d defined as:

$$y_{dk} = y_k \text{ for } k \in U_d \text{ and } y_{dk} = 0 \text{ for } k \notin U_d$$

In other words, $y_{dk} = I\{k \in U_d\} y_k$

Because $t_d = \sum_{k \in U_d} y_k = \sum_{k \in U} y_{dk}$, we can estimate

domain total of y by estimating the population total of y_{dk}

NOTE: Unconditional and conditional inference

- In the **unconditional approach,** the contribution of extra variance caused by random domain sample sizes can be incorporated in variance expressions and computation
  - Variance estimates for unplanned domains are often used
- In the **conditional approach**, inference is conditional on the realized sample and domain sample sizes are considered as fixed quantities
  - Variance estimators for planned domains are often used
  - Note again the similarity with post-stratification
- Lehtonen & Pahkinen (2004) p. 90
- Lehtonen & Veijanen (2009) p. 224
- Coquet & Lesage (2012)
- Rao J.N.K. (1999)

Horvitz-Thompson estimator of domain totals

Horvitz-Thompson (HT) estimator (*expansion estimator*) is the basic *design-based direct* estimator of the domain total $t_d = \sum_{k \in U_d} y_k$, $d = 1,\ldots,D$:

$$\hat{t}_{dHT} = \sum_{k \in U_d} I_k y_k / \pi_k = \sum_{k \in s_d} y_k / \pi_k = \sum_{k \in s_d} a_k y_k \qquad (1)$$

HT estimates of domain totals are additive: they sum up to the HT estimator $\hat{t}_{HT} = \sum_{k \in s} a_k y_k$ of the population total

$$t = \sum_{k \in U} y_k \qquad (2)$$

As $E(I_k) = \pi_k$, the HT estimator is design unbiased for t_d

NOTE: More detailed treatment of HT (and GREG) under planned domains: See Lehtonen & Veijanen (2009)

Variance estimation for HT - 1

Standard *variance estimator* for $\hat{t}_{dHT}$ under **planned** domains:

$$\hat{V}\left(\hat{t}_{dHT}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) y_k y_l \qquad (3)$$

where $a_k = 1/\pi_k$ and $a_{kl} = 1/\pi_{kl}$

NOTE: Variance estimator (3) is somewhat impractical for many unequal probability sampling designs because the second-order inclusion probabilities π_{kl} are needed

Approximations have been developed for standard sampling designs to be used in practical situations

Variance estimation for HT - 2

Variance estimation for planned domains in practice

Approximations to π_{kl} for fixed- size without-replacement (WOR) probability proportional-to-size (πPS) designs :

- Hájek (1964) and Berger (2004, 2005) approximation
- Särndal (1996) approximation

Alternative methods: Resampling

- Berger and Skinner (2005) jackknife variance estimator
- Kott (2006) delete-a-group jackknife variance estimator

see Lehtonen & Veijanen (2009) page 226-227

NOTE: For some design types $\pi_{kl} = \pi_k \pi_l,\ k \neq l$

Variance estimation for HT - 3

Planned domains: Conditional variance estimator assuming fixed domain sample sizes

Approximate estimator:

$$\hat{V}_A\left(\hat{t}_{dHT}\right) = \frac{1}{n_d(n_d-1)} \sum_{k \in s_d} \left(n_d a_k y_k - \hat{t}_{dHT}\right)^2 \qquad (4)$$

where n_d refers to domain sample size

For example, SAS Procedure SURVEYMEANS uses (4) for planned type domain structures

Variance estimation for HT - 4

Unplanned domains: Unconditional variance estimator by accounting for random domain sample sizes

Approximate variance estimator by using *extended domain variables* y_{dk} :

$$\hat{V}_U\left(\hat{t}_{dHT}\right) = \frac{1}{n(n-1)} \sum_{k \in s} \left(n a_k y_{dk} - \hat{t}_{dHT} \right)^2, \qquad (5)$$

where *n* is the total sample size

NOTE: e.g. SAS procedure SURVEYMEANS uses (5) for unplanned cases

NOTE: Extended domain variables are $y_{dk} = I\{k \in U_d\} y_k$

Recall: $y_{dk} = y_k$ if $k \in U_d$, 0 otherwise

Topic 4 DIRECT GREG AND MODEL-FREE CALIBRATION

- Origins of the traditional GREG and calibration
- Components of estimation procedure
- Basic idea: Difference estimator
- Population fit regression estimator
- Direct GREG estimator for domain totals of continuous study variable
- Variance estimators and approximations
- Example

Traditional linear GREG estimator

- *GREG = Generalized regression estimator*

- Robinson P.M. & Särndal C.-E. (1983) Asymptotic properties of the generalized regression estimator in probability sampling, *Sankhyā Ser. B,* 45, 240–248.

- Särndal, C.E. (1980) On π-inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika* 67, 639–650.

- Särndal C.-E., Swensson B. & Wretman J. (1992) *Model-Assisted Survey Sampling*. New York: Springer.

Traditional model-free calibration estimator

- *Calibration estimators*

- Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *JASA* 87, 376–382.

- Estevao V.M. & Särndal C.-E. (1999) The use of auxiliary information in design-based estimation for domains. *Survey Methodology* 2, 213-221.

- Särndal C.-E. (2007) The calibration approach in survey theory and practice. *Survey Methodology* 33, 99–119.

- Kim J.K. & Park M. (2009) Calibration estimation in survey sampling

Components of GREG estimation procedure

- **Sample survey data**
  - Access to unit-level sample survey data
- **Model specification and model fitting**
  - Specification of the linear fixed-effects model
  - Estimation of model parameters from the sample data
- **Auxiliary data**
  - Access to domain-level or unit-level population data
- **Estimation of domain totals: Two alternatives**
  - Estimation with domain-level auxiliary data
  - Estimation with unit-level auxiliary data

Population fit regression estimator -1

Difference estimator of population total t of y (Särndal 1980)

Let us assume known values y_k^0 that are close to population values y_k, $k \in U$. We write the population total as

$$t = \sum_{k \in U} y_k = \sum_{k \in U} y_k^0 + \sum_{k \in U} (y_k - y_k^0)$$

In practice, sample values y_k, $k \in s$ only are available!

Difference estimator: We estimate the second sum using HT:

$$\hat{t}_{DIFF} = \sum_{k \in U} y_k^0 + \sum_{k \in s} a_k (y_k - y_k^0), \text{ where } a_k = 1/\pi_k$$

Population fit regression estimator -2

In practice, no such y_k^0, $k \in U$, exist! Let us use modelling...

Consider regression superpopulation model

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k, \quad Var(\varepsilon_k) = \sigma_k^2 = \sigma^2 \text{ (constant)}$$

where $\mathbf{x}_k = (1, x_{1k}, ..., x_{Jk})'$ is the vector of auxiliary x-variables

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)'$ is the vector of regression coefficients

If we had access to population values $y_k \in U$ then

a LS (least squares) estimator $\tilde{\boldsymbol{\beta}}_{LS}$ of $\boldsymbol{\beta}$ is:

$$\tilde{\boldsymbol{\beta}}_{LS} = \left(\sum_{k \in U} \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_{k \in U} \mathbf{x}_k y_k \right)$$

Population fit regression estimator -3

Using $\tilde{\boldsymbol{\beta}}_{LS}$ and $\mathbf{x}_k$, $k \in U$, we calculate fitted values $\tilde{y}_k = \mathbf{x}_k' \tilde{\boldsymbol{\beta}}_{LS}$ for all $k \in U$. We define **population fit regression estimator :**

$$\hat{t}_{REG} = \sum_{k \in U} \tilde{y}_k + \sum_{k \in s} a_k (y_k - \tilde{y}_k), \quad \text{where} \quad a_k = 1/\pi_k$$

Because we only have access to sample values $y_k \in s$, we estimate $\boldsymbol{\beta}$ by plugging in HT estimators for both sum components of $\tilde{\boldsymbol{\beta}}_{LS}$ for weighted LS estimator:

$$\hat{\boldsymbol{\beta}}_{WLS} = \left(\sum_{k \in s} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \left(\sum_{k \in s} a_k \mathbf{x}_k y_k \right)$$

and compute fitted y-values $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}_{WLS}$ for all $y_k \in U$

Direct GREG estimator for domains -1

Direct GREG estimator of domain total $t_d = \sum_{k \in U_d} y_k$:

Assisting linear fixed-effects model:

$$y_k = \mathbf{x}_k' \boldsymbol{\beta}_d + \varepsilon_k, \; k \in U_d \qquad (6)$$

Domain-specific parameter $\boldsymbol{\beta}_d$ is estimated using weighted LS in each domain:

$$\hat{\boldsymbol{\beta}}_{WLSd} = \hat{\boldsymbol{\beta}}_d = \left(\sum_{k \in s_d} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s_d} a_k \mathbf{x}_k y_k$$

where weights are $a_k = 1/\pi_k$

Direct GREG estimator for domains -2

Using beta estimates of (6), fitted values $\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}}_d$, $k \in U_d$, and residuals $e_k = y_k - \hat{y}_k$, $k \in s_d$, are incorporated into

direct GREG estimator

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k e_k \qquad (7)$$

First part: Synthetic (SYN) estimator

Second part: HT estimator of residual total $\sum_{k \in U_d} E_k$

(adjustment for design bias of SYN estimator)

NOTE: (7) operates with unit-level x-data from population

Traditional regression estimator

Rearranging the terms of GREG: traditional regression estimator

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\boldsymbol{\beta}}_d, \qquad (8)$$

where $\mathbf{t}_{dx} = \sum_{k \in U_d} \mathbf{x}_k = \left(N_d, \sum_{k \in U_d} x_{1k}, \ldots, \sum_{k \in U_d} x_{Jk}\right)'$

$$\hat{\mathbf{t}}_{dx} = \sum_{k \in s_d} a_k \mathbf{x}_k$$

Variance of $\hat{t}_{dGREG}$ can be approximated using sample residuals $e_k = y_k - \mathbf{x}_k' \hat{\boldsymbol{\beta}}_d$:

$$\hat{V}_1\left(\hat{t}_{dGREG}\right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) e_k e_l \qquad (9)$$

NOTE: (8) requires totals of auxiliary variables in each domain

Practical variance estimator for direct GREG for planned domains

Approximate variance estimator of GREG:

$$\hat{V}_A\left(\hat{t}_{dGREG}\right) = \frac{1}{n_d(n_d-1)} \sum_{k \in s_d} \left(n_d a_k e_k - \hat{t}_{dHTe}\right)^2 \qquad (10)$$

where

n is the total sample size and $a_k = 1/\pi_k$ (design weights)

$e_k = y_k - \hat{y}_k$ are residuals in fitting the model

$y_k = \beta_{0d} + \beta_{1d} x_{1k} + \beta_{2d} x_{2k} + \ldots + \beta_{Jd} x_{Jk} + \varepsilon_k, \quad k \in U_d$

$\hat{t}_{dHTe} = \sum_{k \in s_d} a_k e_k$ is HT estimator of residual total in domain d

NOTE: Similarity of (10) with HT variance estimator (4)

for planned domains, but there is an important difference!

Simple variance estimator for SRS

Simple **variance estimator** of $\hat{t}_{dGREG}$ under SRS sampling:

Assisting domain-specific model: $y_k = \beta_{0d} + \beta_{1d} x_k + \varepsilon_k$

GREG estimator: $\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \frac{N_d}{n_d} \sum_{k \in s_d} (y_k - \hat{y}_k)$

where N_d is population size and n_d is sample size in domain d

Variance estimator: $\hat{V}_{SRS}(\hat{t}_{dGREG}) = \hat{V}_{SRS}(\hat{t}_{dHT})(1 - \hat{\rho}^2_{dyx})$

where $\hat{V}_{SRS}(\hat{t}_{dHT})$ is variance estimator of SRS-based (HT) estimator

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k = \frac{N_d}{n_d} \sum_{k \in s_d} y_k$$

and $\hat{\rho}_{dyx}$ is sample correlation of y and x in domain d

Direct GREG as calibration estimator

GREG can be written as a weighted sum of observations incorporating so-called g-weights (Särndal et al. 1992):

$$\hat{t}_{dGREG} = \sum_{k \in s_d} a_k g_{dk} y_k, \qquad (11)$$

where $g_{dk} = I_{dk} + I_{dk} \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx} \right)' \hat{\mathbf{M}}_d^{-1} \mathbf{x}_k$ and $\hat{\mathbf{M}}_d = \sum_{i \in s_d} a_i \mathbf{x}_i \mathbf{x}_i'$

$I_{dk} = I\{k \in U_d\}$ is the domain membership indicator

g-weights are used in variance estimator

$$\hat{V}_2 \left(\hat{t}_{dGREG} \right) = \sum_{k \in s_d} \sum_{l \in s_d} (a_k a_l - a_{kl}) g_{dk} e_k g_{dl} e_l \qquad (12)$$

Calibration property

GREG as calibration estimator

$$\hat{t}_{dGREG} = \sum_{k \in s_d} a_k g_{dk} y_k$$

where $a_k g_{dk}$ are **calibration weights**

Calibration for auxiliary x-variables involves:

$$\hat{t}_{dGREG}(x_j) = \sum_{k \in s_d} a_k g_{dk} x_{jk} = t_{dx_j} = \sum_{k \in U_d} x_{jk} \text{ for } j = 1,...,J$$

NOTE: Calibration property:

Applying calibration weights $a_k g_{dk}$ for any x-variable reproduces the known population total t_{dx_j} of x-variable x_j in domain d

EXAMPLE 6
Direct HT and direct GREG for planned domains

- Comparison of HT and direct GREG
- Examination whether auxiliary data improves efficiency or not

- Population: N = 431,000 households
- Household sampling: Stratified πPS (PPS-WOR)
- Size variable in PPS-WOR: Number of household members
- Strata: D = 12 NUTS4 regions (domains)
- Planned type domains
- Proportional allocation for the strata
  - Domain (stratum) sample sizes are assumed fixed
- Total sample size: n = 1000 households

- Source: Lehtonen & Veijanen (2009) pp. 228-230

Variables

- **Study variable** *y*
  - Disposable household income

- **Auxiliary x-variables** (known for all HHs)
  - EDUC: the number of household members who had higher education
  - EMP: the number of months in total the household members were employed during last year
  - Variables are derived from administrative registers
- NOTE: for this pedagogical exercise we assume access to the total parameter values of study variable y in the domains
- This gives option to compare results with true values

Estimators of domain totals

HT estimator with variance estimator (5)

Direct linear GREG estimator with variance estimator (10)

Parameters: Domain totals $t_d = \sum_{k \in U_d} y_k$, $d = 1,...,12$

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k$$

$$\hat{V}_A\left(\hat{t}_{dHT}\right) = \frac{1}{n_d(n_d - 1)} \sum_{k \in s_d} \left(n_d a_k y_k - \hat{t}_{dHT}\right)^2$$

$$\hat{t}_{dGREG} = \hat{t}_{dHT} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\boldsymbol{\beta}}_d$$

$$\hat{V}_A\left(\hat{t}_{dGREG}\right) = \frac{1}{n_d(n_d - 1)} \sum_{k \in s_d} \left(n_d a_k e_k - \hat{t}_{dHTe}\right)^2$$

Assisting models in GREG

Direct GREG estimator with linear fixed-effects assisting model and domain-specific terms:

$$y_k = \beta_{0d} + \beta_{1d}\mathrm{EMP}_k + \varepsilon_k \quad \text{(column 2)}$$

$$y_k = \beta_{0d} + \beta_{1d}\mathrm{EMP}_k + \beta_{2d}\mathrm{EDUC}_k + \varepsilon_k \quad \text{(column 3)}$$

The models are fitted separately in each domain

Beta parameters are estimated using WLS with design weights

NOTE: Domain-specific intercepts and slopes
Therefore, this GREG estimator is direct

Measures of quality in Table 2

Absolute relative error in domain *d*:

$$ARE(\hat{t}_d) = \left|\hat{t}_d - t_d\right| / t_d$$

MARE in a domain group is the mean of absolute relative errors over domains in the group

Coefficient of variation in domain *d*:

$$\text{c.v}(\hat{t}_d) = \text{s.e}(\hat{t}_d) / \hat{t}_d$$

MCV in a domain group is the mean of coefficients of variation of the estimate, over domains in the group

Table 2. Mean absolute relative error MARE (%) and mean coefficient of variation MCV (%) of direct HT and direct calibration (GREG) estimators of totals for minor, medium-sized and major domains by using various amounts of auxiliary information for **planned domains**.

	HT		Direct GREG			
	Auxiliary information					
	1 None		2 Domain sizes and domain totals of EMP		3 Domain sizes and domain totals of EMP and EDUC	
Domain sample size class	MARE %	MCV %	MARE %	MCV %	MARE %	MCV %
Minor $8 \le n_d \le 33$	11.5	**11.9**	5.8	**7.7**	6.4	**6.8**
Medium $34 \le n_d \le 45$	7.6	**9.0**	3.7	**8.0**	3.6	**8.1**
Major $46 \le n_d \le 277$	12.5	**5.2**	4.3	**4.7**	5.2	**3.7**

Lessons learned – Example 6 Planned domains

- Domains are taken as independent sub-populations
- Direct estimators are used
- **Estimation error**
  - Mean absolute error MARE figures are smaller for GREGs when compared with HT, in all three domain sample size groups
- **Estimation accuracy (variance)**
  - Mean coefficient of variation MCV figures tend to be smaller for both GREGs, when compared with HT
  - GREG with more use of auxiliary data tends to be more accurate than the GREG with less use of auxiliary data
- Incorporation of auxiliary data in the direct GREG estimation procedure makes sense!

Topic 5 INDIRECT GREG AND CALIBRATION

- “Borrowing strength” in model-assisted methodology with indirect estimation procedures
- Indirect linear GREG estimator for domain totals of continuous study variable
- Variance estimators
- Example

Indirect estimators

- **Recall definition**
- Indirect estimator uses y-values not only from the domain of interest itself but also outside the domain or from earlier time points
- “Borrowing strength” from other domains (spatially) or in a temporal dimension
- Borrowing strength can be exercised both in design-based SAE and model-based SAE

Indirect GREG estimator for domains - 1

Indirect GREG estimator of domain total parameters

$$t_d = \sum_{k \in U_d} y_k,\ d = 1,...,D$$

Assume known vector values of auxiliary x-data with J variables

$$\mathbf{x}_k = (1, x_{1k}, ... x_{Jk})',\ k \in U$$

Assisting linear fixed-effects model:

$$y_k = \mathbf{x}_k' \boldsymbol{\beta} + \varepsilon_k,\ Var(\varepsilon_k) = \sigma^2,\ k \in U \qquad (13)$$

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_J)'$ beta coefficients **common for all domains**

Parameter $\boldsymbol{\beta}$ is estimated from the sample s by weighted least squares with weights $a_k = 1/\pi_k$:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{k \in s} a_k \mathbf{x}_k \mathbf{x}_k' \right)^{-1} \sum_{k \in s} a_k \mathbf{x}_k y_k$$

Some notes on efficiency

- The model (13) given by
  $$y_k = \mathbf{x}_k'\boldsymbol{\beta} + \varepsilon_k$$
  is not domain specific but is specified as common model for all domains
- This means **borrowing strength** for a given (possibly small) domain from other “similar” (possibly larger) domains
- Efficiency improves over HT if the explanatory power of x-variables in the model is good over the domains involving small residuals for every domain
- NOTE: GREG estimator remains (nearly) design unbiased in a domain irrespective of the correctness of the model

Indirect GREG estimator for domains - 2

Fitted values

$$\hat{y}_k = \mathbf{x}_k' \hat{\boldsymbol{\beta}},\ k \in U$$

and sample residuals

$$e_k = y_k - \hat{y}_k,\ k \in s$$

are incorporated into **indirect GREG estimator**

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k (y_k - \hat{y}_k) = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} a_k e_k,\ d = 1,\ldots,D$$

NOTE: This GREG is indirect since all y-values in the sample contribute to the beta estimates

NOTE: Difference to direct GREG (7) is in the predictions!

Examples of assisting models

Linear fixed - effects models:

Common model with J x-variables for all domains

$$y_k = \beta_0 + \beta_1 x_k + ... + \beta_J x_{Jk} + \varepsilon_k,\ k \in U$$

Domain-specific fixed intercepts and common slopes

$$y_k = \beta_{01} I_{1k} + \beta_{02} I_{2k} + ... + \beta_{0D} I_{Dk} + \beta_1 x_k + ... + \beta_J x_{Jk} + \varepsilon_k,\ k \in U$$

where $I_{dk} = I\{k \in U_d\} = 1$ if $k \in U_d$, 0 otherwise

Linear mixed model with domain-specific random intercepts

$$y_k = (\beta_0 + u_d) + \beta_1 x_k + ... + \beta_J x_{Jk} + \varepsilon_k,\ k \in U_d,\ d = 1,...,D$$

GREG as calibration estimator

Indirect GREG can be written as a weighted sum of observations incorporating *calibrated weights* (g-weights) $w_k = a_k g_{dk}$:

$$\hat{t}_{dGREG} = \sum_{k \in s_d} w_k y_k = \sum_{k \in s_d} a_k g_{dk} y_k$$

where $g_{dk} = I_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx} \right)' \hat{\mathbf{M}}^{-1} \mathbf{x}_k$ are *extended* g-weights

$I_{dk} = I\{k \in U_d\}$ is domain membership indicator

such that $I_{dk} = 1$ if $k \in U_d$, 0 otherwise

$\hat{\mathbf{M}} = \sum_{i \in s} a_i \mathbf{x}_i \mathbf{x}_i'$ NOTE: Extends over the whole sample s

NOTE: **Calibration property** holds for all x-variables x_j, $j = 1,...,J$:

$$\hat{t}_{dx_j GREG} = \sum_{k \in s_d} a_k g_{dk} x_{jk} = \sum_{k \in U_d} x_{jk} = t_{dx_j}$$

Variance estimator of indirect GREG with extended g-weights

$$\hat{V}\left(\hat{t}_{dGREG}\right) = \sum_{k\in s}\sum_{l\in s}(a_k a_l - a_{kl})g_{dk}e_k g_{dl}e_l \qquad (14)$$

where $e_k = y_k - \hat{y}_k$ are sample residuals

$$g_{dk} = I_{dk} + \left(\mathbf{t}_{dx} - \hat{\mathbf{t}}_{dx}\right)' \hat{\mathbf{M}}^{-1}\mathbf{x}_k \text{ with } \hat{\mathbf{M}} = \sum_{i\in s} a_i \mathbf{x}_i \mathbf{x}_i'$$

Extended g-weights g_{dk} are used

The whole sample data set s is used to estimate variance for given domain d

NOTE: $\hat{V}\left(\hat{t}_{dGREG}\right)$ requires weights $a_{kl} = 1/\pi_{kl}$

where π_{kl} are second-order inclusion probabilities

They are intractable for practical variance estimation

More practical variance estimator

Approximate variance estimator of indirect GREG for unplanned domains by using *extended residuals*:

$$\hat{V}_U\left(\hat{t}_{dGREG}\right) = \frac{n}{n-1}\sum_{k\in s}\left(a_k e_{dk} - \hat{t}_{dHTe}\,/\,n\right)^2 \qquad (15)$$

where n is the total sample size and $a_k = 1/\pi_k$ (design weights)

$e_{dk} = I\{k \in U_d\}y_k - \hat{y}_k$ are extended residuals, where $e_k = y_k - \hat{y}_k$

NOTE: $e_{dk} = -\hat{y}_k$ if $k \notin U_d$ and $k \in s$ (Lehtonen & Pahkinen 2004 p. 202)

$\hat{t}_{dHTe} = \sum_{k\in s_d} a_k e_k$ is HT estimator of residual total in domain d

Alternatively, it is possible to use in $\hat{V}_U\left(\hat{t}_{dGREG}\right)$ the *extended domain variables* $y_{dk} = I\{k \in U_d\}y_k$ (Lehtonen & Veijanen 2009 p. 234)

EXAMPLE 7: HT and GREG for planned and unplanned domains

- Comparison of direct HT with direct and indirect GREG for planned and unplanned domains
- Population: N = 431,000 households
- **Household sampling:**
- Planned domains: Stratified πPS (PPS-WOR) with household size as the size variable and domains as the strata
- Unplanned domains: πPS (PPS-WOR, no stratification)
- Size variable in PPS-WOR: Number of household members
- Domains: D = 12 NUTS4 regions (domains)
- Sample size: n = 1000 households
- Lehtonen & Veijanen (2009) Section 4.2.

Estimators

- **Estimators of domain totals**
  - HT estimator (1) with variance estimators (4) and (5)
  - Direct GREG estimator with assisting model (6) and variance estimator (10)
  - Indirect GREG estimator with assisting model (13) and variance estimator (15)

Assisting models in GREG

GREG estimator

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

is assisted by linear fixed-effects models

Assisting model for direct GREG

$$y_k = \beta_{0d} + \beta_{1d} \mathrm{EMP}_k + \varepsilon_k,\ k \in U_d$$

The model is fitted separately in each domain

Assisting model in indirect GREG

$$y_k = \beta_0 + \beta_1 \mathrm{EMP}_k + \varepsilon_k,\ k \in U$$

The model is fitted to the whole sample

Beta parameter vectors are estimated with WLS using design weights

Measure of quality in Table 4

Coefficient of variation in domain *d*:

$$\text{c.v}(\hat{t}_d) = \text{s.e}(\hat{t}_d) / \hat{t}_d$$

MCV in a domain group is the mean of coefficients of variation of the estimate, over domains in the group

Table 3. Mean coefficient of variation MCV (%) of direct HT and direct and indirect GREG estimators of totals for minor, medium-sized and major domains for planned and unplanned domains.

	Planned domains		Unplanned domains	
	(a) HT	(b) Direct GREG	(c) HT	(d) Indirect GREG
Domain sample size class	MCV %	MCV %	MCV %	MCV %
Minor $8 \le n_d \le 33$	11.9	7.7	28.3	9.0
Medium $34 \le n_d \le 45$	9.0	8.0	20.3	8.1
Major $46 \le n_d \le 277$	5.2	4.7	9.6	5.0

Variance estimators:

(a) $\hat{V}_A\left(\hat{t}_{dHT}\right) = \frac{1}{n_d(n_d-1)} \sum_{k \in s_d} \left(n_d a_k y_k - \hat{t}_{dHT}\right)^2$ (c) $\hat{V}_U\left(\hat{t}_{dHT}\right) = \frac{1}{n(n-1)} \sum_{k \in s} \left(n a_k y_{dk} - \hat{t}_{dHT}\right)^2$

(b) $\hat{V}_A\left(\hat{t}_{dGREG}\right) = \frac{1}{n_d(n_d-1)} \sum_{k \in s_d} \left(n_d a_k e_k - \hat{t}_{dHTe}\right)^2$ (d) $\hat{V}_U\left(\hat{t}_{dGREG}\right) = \frac{1}{n(n-1)} \sum_{k \in s} \left(n a_k e_{dk} - \hat{t}_{dHTe}\right)^2$

Lessons learned from EXAMPLES 6 & 7

- **Planned domains, direct estimators**
  - GREG better than HT in terms of accuracy, in small domains in particular
- **Unplanned domains, indirect estimators**
  - GREG much better than HT in terms of accuracy
- Use of auxiliary data makes sense!
- **Planned vs. unplanned case**
  - For HT, accuracy clearly better in planned domains case
  - For GREG, better accuracy in small planned domains
- Stratification for important domains of interest makes sense!
  - An issue of the survey planning stage

Topic 6 EXTENDED GREG AND MODEL-ASSISTED CALIBRATION

- **Extended GREG and model calibration estimators**
- This far, our study variable was of continuous type and linear assisting models were used
- Assisting generalized linear mixed models (GLMM) are needed for binary, polytomous and count variables, and for mixed model formulations
- **EXAMPLE**
- GREG and model-assisted calibration estimators for the number of ILO unemployed in regions
- Study variable is now **polytomous** with 3 classes: Employed, Unemployed, Not in labour force
- Data: LFS sample data, unit-level auxiliary data from registers
- Multinomial logistic mixed model as assisting model

Data requirements

- **Traditional linear GREG estimator and model-free calibration estimator for continuous study variable**
  - Linear fixed-effects models are used
  - Unit-level x-vectors not necessarily needed
  - Known domain totals of x-variables only are needed
  - Often used in all data infrastructures but applicable in ”survey” countries in particular (current paradigm in Official statistics)
- **Extended GREG family estimators and model-assisted calibration estimators for other study variable types**
  - Unit-level x-data are assumed for all units in population
  - Linear and generalized linear mixed models are used
  - Applicable in ”register” countries in particular
  - Active research & development in academic communities

EXAMPLE 8: Assisting model in GREG and model-assisted calibration

Linear mixed model for continuous study variable y

$$y_k = \mathbf{x}_k'\boldsymbol{\beta} + u_d + \varepsilon_k,\ k \in U_d,\ d = 1,...,D$$

where $\mathbf{x}_k = (1, x_{1k},...,x_{pk})'$, $\boldsymbol{\beta} = (\beta_0, \beta_1,..., \beta_p)'$

u_d are domain-level random intercepts

$u_d \sim N(0, \sigma_u^2)$, $\varepsilon_k \sim N(0, \sigma^2)$, u_d and ε_k independent

Estimate $\boldsymbol{\beta}$ and σ_u^2 from the sample data set s (`lme4`, MIXED)

Calculate estimates $\hat{u}_d$, $d = 1,...,D$ and calculate fitted values

$$\hat{y}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d,\ k \in U_d,\ d = 1,...,D$$

Used in linear mixed model assisted GREG estimator (MGREG)

Lehtonen & Veijanen (1999), Lehtonen, Särndal and Veijanen (2003)

EXAMPLE 9: Assisting model in GREG and model-assisted calibration

Logistic fixed - effects model

for binary response variable y

$$E_m(y_k) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_k'\boldsymbol{\beta})}$$

Estimate $\boldsymbol{\beta}$ from the sample data set s by ML

Calculate fitted values $\hat{y}_k = \frac{\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})}{1+\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}})}, \quad k \in U$

Used in logistic model assisted GREG estimator (LGREG)

Lehtonen & Veijanen (1998)

EXAMPLE 10: Assisting model in GREG and model-assisted calibration

Logistic mixed model for binary response variable y

$$E_m(y_k \mid u_d) = \frac{\exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}{1+\exp(\mathbf{x}_k'\boldsymbol{\beta} + u_d)}, \quad k \in U_d,\ d = 1,...,D$$

where u_d are domain-level random intercepts, $u_d \sim N(0, \sigma_u^2)$

Estimate $\boldsymbol{\beta}$ and σ_u^2 from the sample data set s (`lme4`, MIXED)

Calculate estimates $\hat{u}_d$, $d = 1,...,D$ and calculate fitted values:

$$\hat{y}_k = \frac{\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)}{1+\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)}, \quad k \in U_d,\ d = 1,...,D$$

Used in logistic mixed model assisted GREG estimator (MLGREG)

Lehtonen, Särndal & Veijanen (2005), Lehtonen & Veijanen (2009)

Estimation of the model

- GLMMs can be fitted in R by:
  - R packages nlme or lme4 (glmer function) using maximum likelihood
- GLMMs **with survey weights** for unit and domain level can be fitted in SAS by:
  - Procedures GLIMMIX (using ML) or MIXED (using REML or ML)
  - R options for this purpose?

- Some classical references
- Generalized linear (fixed-effects) models (GLM)
  Nelder & Wedderburn (1972) JRSS-A
  McCullagh & Nelder (1982) Generalized Linear Models. Springer.
- Generalized linear mixed models (GLMM) family models
  Demidenko (2005) Mixed Models: Theory and Applications. Wiley.

GREG estimator assisted by GLMM

- For an assisting GLMM for GREG the formulation of GREG estimator for domain total and mean or proportion remains the same. The difference is in obtaining the predicted y-values

MGREG estimator for domain total t_d of continuous y-variable

Assisting model: Linear mixed model

Predicted values: $\hat{y}_k = \mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d,\ k \in U_d,\ d = 1,...,D$

MLGREG for domain proportion p_d of binary y-variable

Assisting model: Logistic mixed model:

Predicted values: $\hat{y}_k = \dfrac{\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)}{1+\exp(\mathbf{x}_k'\hat{\boldsymbol{\beta}} + \hat{u}_d)},\ k \in U_d,\ d = 1,...,D$

For MGREG and MLGREG the estimator is of the same form:

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

Calibration estimator assisted by GLMM

- For an assisting GLMM for model calibration estimator, the formulation of the model-assisted calibration estimator for domain total and mean or proportion remains the same.
  The difference is in obtaining the predicted y-values

Calibration estimators $\hat{t}_d = \sum_{k \in s_d} w_k y_k$

w_k method-specific **calibration weight** for element k

Weights are constructed to satisfy calibration equations:

$$\sum_{k \in s_d} w_k \mathbf{z}_k = \sum_{k \in U_d} \mathbf{z}_k = \left(N_d, \sum_{k \in U_d} \hat{y}_k \right)'$$

where $\mathbf{z}_k = (1, \hat{y}_k)'$, $s_d = s \cap U_d$, $d = 1, \ldots, D$

Fitted values $\hat{y}_k = f(\mathbf{x}_k'(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d))$ with $\mathbf{x}_k = (1, x_{1k}, \ldots, x_{Jk})'$, $k \in U$

EXAMPLE 11: Poverty rate

- Source: Lehtonen & Veijanen (2016a)
- The aim: Estimation of poverty rate for regions by using GREG estimators assisted by logistic fixed-effects model (LGREG estimator, Lehtonen & Veijanen 1998) and logistic mixed model (MLGREG, Lehtonen, Särndal & Veijanen 2003)
- Methods:
- See separate paper

- NOTE: A related paper by Molina & Rao (2010)

Poverty rate: Results

Table 4 Absolute relative bias (ARB %) and relative root men squared error (RRMSE %) of estimators of poverty rate in a design-based simulation experiment of 1,000 SRSWOR samples.

	Estimator	ARB (%)			RRMSE (%)		
		Expected domain sample size			Expected domain sample size		
		5-12	12-25	25-151	5-12	12-25	25-151
Direct estimator	HT	1.7	2.2	0.9	83.7	60.1	38.9
Indirect estimators Assisting models							
(a) Fixed-effects logistic model with domain-specific intercepts	LGREG	1.8	1.9	0.9	83.7	59.9	38.5
(b) Mixed logistic model with domain-specific random intercepts	MLGREG	2.0	1.8	0.9	72.4	55.0	36.8

Lessons learned - EXAMPLE 11

- All estimators were nearly design unbiased as expected
- Model choice had larger effect on RRMSE:
- Fixed-effects logistic model with domain-specific intercepts did not yield good results with the model-assisted LGREG estimator
- The reason might be instable estimation, in the group of smallest domains in particular
- Note: There are 36 fixed intercept parameters to be estimated!
- This result suggests that a fixed-effects model with domain-specific parameters might not be a good idea if the number of domains is large
- The best results were obtained with the logistic mixed model assisted MLGREG estimator
- This estimator outperformed clearly the HT and LGREG estimators.

Generalized regression GREG: EXAMPLES from literature

- Simulation results and empirical examples statistical properties of the extended family GREG estimators
- Lehtonen & Veijanen (1998)
  - GREG assisted by logistic fixed-effects model (LGREG)
- Lehtonen & Veijanen (1999)
  - GREG assisted by linear mixed model
- Lehtonen, Särndal and Veijanen (2003, 2005)
  - GREG assisted by linear mixed model (MGREG)
  - GREG assisted by logistic mixed model (MLGREG)
- Lehtonen and Veijanen (2009)
  - GREG assisted by linear and logistic mixed models
- Lehtonen, R., Veijanen, A., Myrskylä, M. and Valaste, M. (2011)
  - AMELI project: GREG applications to poverty indicators

Model-assisted calibration

- Idea: Extension of model-free calibration beyond linear models for continuous study variables to cover nonlinear models for continuous variables and GLMs and GLMMs for binary, polytomous and count type study variables
  - E.g. Linear mixed models, Logistic mixed models
- Calibration principle in domain estimation:
  Calibration of totals of *model predictions* estimated from sample to agree with the population totals of model predictions
- NOTE: difference w.r.t. model-free calibration
- Model calibration:
  Wu and Sitter (2001)
  Montanari and Ranalli (2005, 2009)

Model-assisted calibration procedure for domains

- **Modelling phase**:
  - Model specification
    - Models with no domain-specific terms
    - Models with (fixed or random) domain-specific terms
  - The model is fitted using the entire sample data
  - “Borrowing strength” from other (similar) domains)
  - Predicted y-values are computed for every population element by using the estimated model parameters and auxiliary x-data
- **Calibration phase**:
  - Calibration of the sample total of predicted y-variable values to the **population** level, **domain** level or an **intermediate** (regional, spatial, neighborhood) level

Model-assisted calibration: EXAMPLES from literature

- Simulation results and empirical examples on statistical properties of model-assisted calibration estimators

- Lehtonen & Veijanen (2012)
  - Calibration and GREG assisted by logistic fixed-effects model and logistic mixed model
- Lehtonen & Veijanen (2016a,b)
  - Calibration and GREG assisted by logistic mixed model

- Lehtonen, R., Veijanen, A., Myrskylä, M. and Valaste, M. (2011)
  - AMELI project: Model-assisted calibration applied to poverty indicators

NOTE on computation

- **Software for calibration and GREG**
- SAS macro language programs by SCB, Statistics Canada, INSEE

- Zardetto (2015)
  ISTAT: R package `ReGenesees`
  http://www.istat.it/it/files/2014/05/Zardetto-jos-2015-0013.pdf
- R package `sampling` (Matei & Tille 2916)
  https://cran.r-project.org/web/packages/sampling/sampling.pdf
- R package `icarus`
  https://cran.r-project.org/web/packages/icarus/icarus.pdf

CASE STUDY 1: Estimation of mean of "Perceived income" for regional domains

- Comparison of regional mean estimates using direct HT and indirect GREG assisted with linear fixed-effects and mixed models
- Data sources: EU-SILC data and statistical registers of Statistics Finland

- Master Thesis in Statistics
  Nico Maunula (2012). Small Area Estimation Methods with Application to Perceived Income for Domains in Finland in 2009. Master's Thesis, University of Helsinki. (In Finnish)

Study setting

- Target population: *N* about 4,3 million
- Regions (domains): *D* = 70 NUTS4 areas
- Sizes of regions vary (2000 – 1 million):

- **Stratified unequal probability sampling**
- Sample size *n* = 11,000 households
- Domains are of **unplanned** type
  - Smallest domain sample size: 10 persons
  - Largest domain sample size: 2425 persons

- CAPI interviews with household head as respondent

- Reweighting to adjust for unit nonresponse
- Model-free calibration for final weights w_k

Auxiliary data

- Auxiliary data are taken from statistical registers covering the target population
- Registers maintained by Statistics Finland

- Auxiliary data were merged with sample survey data at the unit level by using unique identification keys
  - Personal ID number

Study variable

- HS120: **Ability to make ends meet**
- Represents ”experienced” (perceived) income (contrasted with “actual” income)
  - A subjective wellbeing indicator
- Ordinal level measurement with 6 levels
  - 1 = lowest, 6 = highest
  - Treated as continuous variable in modelling
  - Mean = 4.3 in SILC data
  - NOTE: Why “perceived income” This is because it is not available in administrative registers!

HS120: Ability to make ends meet

SOCIAL EXCLUSION (Non-monetary household deprivation indicators)
Cross-sectional and longitudinal
Reference period: current
Unit: household
Mode of collection: household respondent

Values	
1	with great difficulty
2	with difficulty
3	with some difficulty
4	fairly easily
5	easily
6	very easily

Flags	
1	filled
-1	missing

The household respondent's assessment of the level of difficulty experienced by the household in making ends meet.
A household may have different source of income and more than one household member may contribute to it. Thinking of the household's total monthly income, the idea is with which level of difficulty the household is able to pay its usual expenses.

Auxiliary variables

- Variables (for HH head) from statistical registers
  - Gender
  - Age group (4 age groups)
  - Education (3 classes)
  - Actual (register) income
  - Socio-economic status (6 classes)
  - Stage in life of household-dwelling unit (5 classes)

- Categorical variables are transformed to indicator (dummy) variables
- 16 x-variables in the regression model
- All variables statistically significant
- R squared = 15%

Models

Linear fixed-effects model

$$y_k = \beta_0 + \beta_1 x_k + ... + \beta_{16} x_{16k} + \varepsilon_k, \ , \ k \in U, \ \varepsilon_k \sim N(0, \sigma^2)$$

where beta coefficients are common for all domains

Linear mixed model

$$y_k = \beta_0 + u_d + \beta_1 x_k + ... + \beta_{16} x_{16k} + \varepsilon_k, \ k \in U_d, \ d = 1, ..., 70$$

with domain-level random intercepts u_d and common fixed slope parameters

$u_d \sim N(0, \sigma_u^2)$, $\varepsilon_k \sim N(0, \sigma^2)$, u_d and ε_k independent

Estimation of assisting models

GREG assisted by linear fixed-effects model

Model fitted by LS

Predicted values

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + \ldots + \hat{\beta}_{16} x_{16k},\ k \in U$$

MGREG assisted by linear mixed model

Model fitted by REML

Predicted values

$$\hat{y}_k = \hat{\beta}_0 + \hat{u}_d + \hat{\beta}_1 x_k + \ldots + \hat{\beta}_{16} x_{16k},\ k \in U_d,\ d = 1,\ldots,D$$

Estimators of regional means

Population mean for domain d: $\bar{y}_d = t_d / N_d, \;\; d = 1,...,70$

HT estimator for domain means $\bar{y}_d$

$$\hat{t}_{dHT} = \sum_{k \in s_d} w_k y_k, \;\; d = 1,...,70$$

$$\hat{\bar{y}}_{dHT} = \hat{t}_{dHT} / N_d$$

wnere N_d are known domain sizes in population

GREG estimators for domain means $\bar{y}_d$

$$\hat{t}_{dGREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} w_k (y_k - \hat{y}_k) \;, \;\; d = 1,...,70$$

$$\hat{\bar{y}}_{dGREG} = \hat{t}_{dGREG} / N_d$$

where $w_k = a_k g_k$ are final calibrated weights (g-weights)

Variance estimators
(unplanned domains)

HT estimator for domain means

$$\hat{V}_U\left(\hat{\bar{y}}_{dHT}\right) = \hat{V}_U\left(\hat{t}_{dHT}\right) / N_d^2$$

$$= \frac{n}{N_d^2(n-1)} \sum_{k \in S} \left(w_k y_{dk} - \hat{t}_{dHT} / n \right)$$

where $y_{dk} = I\{k \in U_d\} y_k$ are extended y-variables

GREG estimators for domain means

$$\hat{V}_U\left(\hat{\bar{y}}_{dGREG}\right) = \hat{V}_U\left(\hat{t}_{dGREG}\right) / N_d^2$$

$$= \frac{n}{N_d^2(n-1)} \sum_{k \in S} \left(w_k e_{dk} - \hat{t}_{dHTe} / n \right)^2$$

where $e_{dk} = I\{k \in U_d\} y_k - \hat{y}_k$ are extended residuals

Quality indicators

Standard error of domain mean estimate $\hat{\bar{y}}_d$

$$\text{s.e}\left(\hat{\bar{y}}_d\right) = \sqrt{\hat{V}\left(\hat{\bar{y}}_d\right)}, \quad d = 1,...,70$$

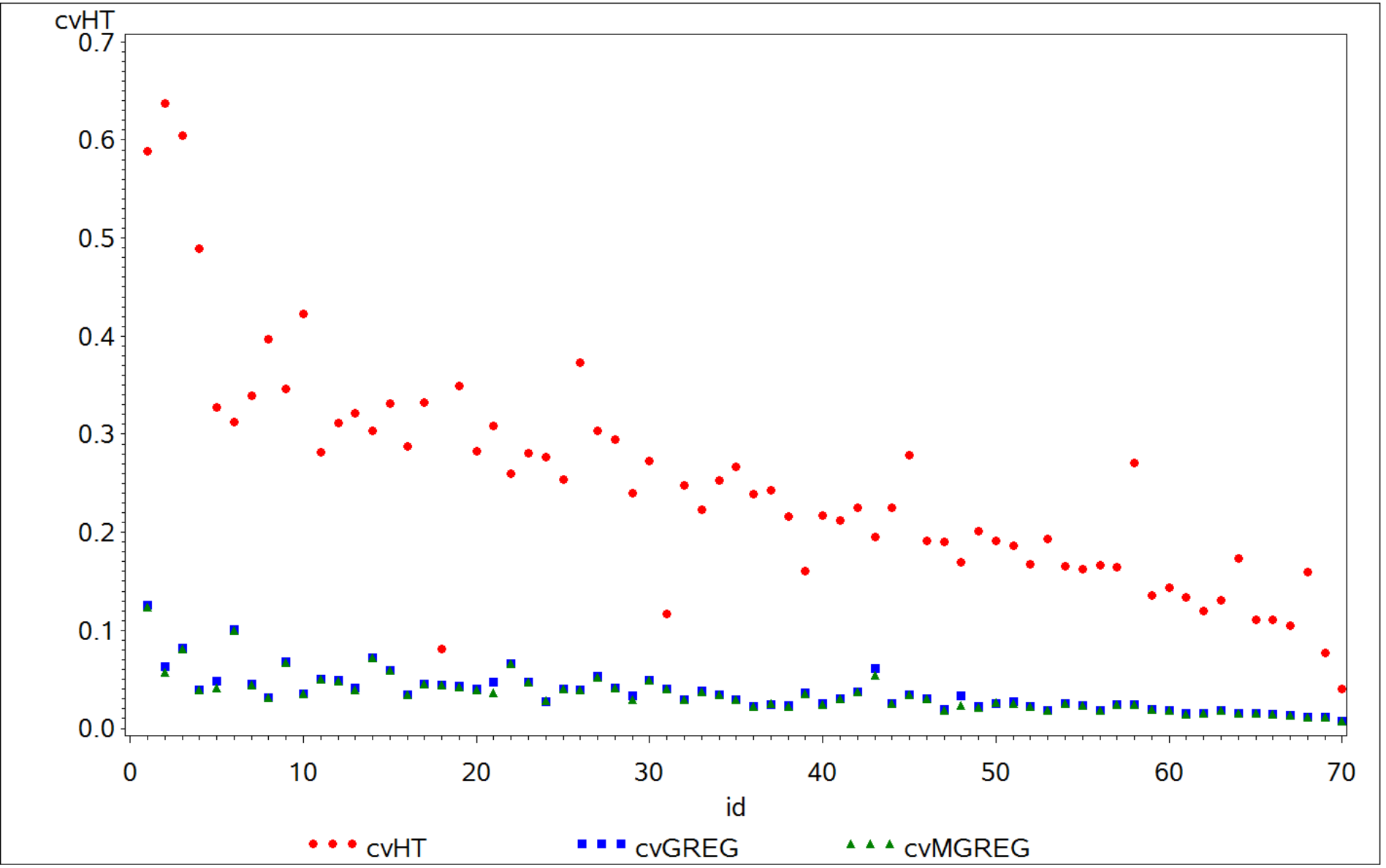
Coefficient of variation of domain mean estimate $\hat{\bar{y}}_d$

$$\text{cv}\left(\hat{\bar{y}}_d\right) = \frac{\text{s.e}\left(\hat{\bar{y}}_d\right)}{\hat{\bar{y}}_d} \quad d = 1,...,70$$

Mean cv calculated in three domain size groups

Table 5. Average coefficient of variation of HT, GREG and MGREG estimates of domain totals by domain sample size class. Sample size n = 11,000, D=70 NUTS3 unplanned domains.

	Domain sample size class			
	Minor	Medium-sized	Major	All
	Average domain sample size			
	34	72	325	152
Direct estimator				
Design-based HT	37.2	24.9	15.4	24.8
Indirect estimators				
Model-assisted				
GREG	5.7	3.7	1.9	3.6
MGREG	5.5	3.6	1.9	3.5


cvHT
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
10
20
30
40
50
60
70
id
cvHT
cvGREG
cvMGREG

CASE STUDY 2: Model-based EBLUP and weighted EBLUP

- QUESTION: How to account for unequal probability sampling and weighting in model-based EBLUP estimation?
- For example:
  - Stratified sampling with non-proportional allocation
  - PPS type sampling designs
- The role of survey weights?
- The role of design variables in the model?

Multilevel models with survey weights

- Rabe-Hesketh (2006) Multilevel modelling of complex survey data. JRSS-A 169 (805–827).
- http://www.gllamm.org/JRSSAsurvey_06.pdf

- Carle A.C. (2009) Fitting multilevel models in complex survey data with design weights: Recommendations. BMC Medical Research Methodology.
- https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-9-49

- Example of using survey design weights in lmer: This is an example of how to use survey design weights with linear mixed models using the lmer() function. It follows the logic of[Carle 2009
- https://rpubs.com/corey_sparks/27276

- West B. (2016) Fitting Weighted Multilevel Models to Complex Sample Survey Data in SAS: A Case Study
  http://www.misug.org/uploads/8/1/9/1/8191072/bwest_weighted_multilevel_models.pdf

Options considered

- PPS-WOR sampling design (πPS design)
- Continuous study variable *y*
- Linear mixed model with random intercepts
- **Model-based EBLUP**
  - Inclusion of PPS size variable in the model
- **Pseudo model-based EBLUP = EBLUPW**
  - Incorporation of design weights in the estimation procedure of the linear mixed model

Simulation experiments - 1

Population N = 1 million elements

D = 100 domains

Size of domain U_d is proportional to $\exp(q_d)$

where q_d is simulated from Uniform(0,2.9)

47 minor domains (-69 elements)
19 medium-sized domains (70-119)
34 major domains (120-)

Simulation experiments – 2

PPS size variable x_1: Uniform(1,11)

Variable x_2 (unrelated to the sampling design):
Uniform(-5,5)

Random intercept u_{0d} and random slopes u_{1d} and u_{2d}:
Multinormal distribution

$$Var(u_{0d}) = 1,\ Var(u_{1d}) = Var(u_{2d}) = 0.125$$

$$Corr(u_{0d}, u_{1d}) = Corr(u_{0d}, u_{2d}) = -0.5,\ Corr(u_{1d}, u_{2d}) = 0$$

Residual ε followed $N(0,100)$

Simulation experiments – 3

Values of the *y*-variable were generated as

$$y_k = (\beta_0 + u_{0d}) + (\beta_1 + u_{1d})x_{1k} + (\beta_2 + u_{2d})x_{2k} + \varepsilon_k$$
$$\beta_0 = \beta_1 = \beta_2 = 1$$

Note: Both random intercepts and random slopes

Correlations of the variables in the population

$$corr(y, x_1) = 0.441$$
$$corr(y, x_2) = 0.446$$

Simulation experiments - 4

Population N = 1,000,000

Sample n = 10,000

Monte Carlo experiments

K = 1000 independent PPS-WOR samples

Inclusion probabilities: $\pi_k = nx_{1k} / \sum_{k \in U} x_{1k}$

Weights $a_k = 1/\pi_k$ varied between 54.5 and 599.8

Models and estimators

EBLUP estimator of domain totals - basic form

$$\hat{t}_{dEBLUP} = \sum_{k \in U_d} \hat{y}_k, \quad d = 1,...,100$$

Fitted models:

Special cases of linear mixed models with random intercepts: $y_k = \beta_0 + u_{0d} + \beta_1 x_k + \varepsilon_k, k \in U_d$

Models fitted by REML or pseudo REML (REML-W)

Predicted values: $\hat{y}_k = \hat{\beta}_0 + \hat{u}_{0d} + \hat{\beta}_1 x_k,$

$k \in U_d, \; d = 1,...,100$

Pseudo EBLUP: mixed model with survey weights

Linear mixed model (matrix form) $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}$

Pseudo EBLUP (EBLUPW) estimators are derived by incorporating design weights a_k in ML-W and REML-W estimation procedures of model parameters by using HT estimators for certain matrix products (Domest and RDomest programs of Ari Veijanen)

Modification of matrix products of $\mathbf{X}$, $\mathbf{y}$, $\mathbf{Z}$ matrix (whose columns are domain indicators), and $\mathbf{e}$ (the vector of residuals):
Matrix product $\mathbf{A'B}$ ($\mathbf{A},\mathbf{B} = \mathbf{X},\mathbf{Z},\mathbf{y},\mathbf{e}$) was replaced by
$\mathbf{A'WB}$, where $\mathbf{W}$ is the diagonal matrix of design weights a_k

Quality measures

Absolute relative bias (ARB) for domain *d*

$$\mathrm{ARB}(\hat{t}_d) = \left| \frac{1}{1000} \sum_{v=1}^{1000} \hat{t}_d(s_v) - t_d \right| / t_d$$

Relative root mean squared error (RRMSE)

$$\mathrm{RRMSE}(\hat{t}_d) = \sqrt{\frac{1}{1000} \sum_{v=1}^{1000} (\hat{t}_d(s_v) - t_d)^2} \; / \, t_d$$

Average ARB and average RRMSE is computed in each domain sample size class

Table 1. Average ARB (%) and average RRMSE (%) of EBLUP estimators.

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 1 $y_k = \beta_0 + u_d + \varepsilon_k$						
EBLUP	19.7	19.5	20.3	19.9	19.8	20.6
EBLUPW	3.7	3.1	2.1	6.8	6.8	6.1
Model 2 $y_k = \beta_0 + u_d + \beta_1 x_{1k} + \varepsilon_k$						
EBLUP	4.0	3.6	2.3	5.4	5.2	4.5
EBLUPW	3.6	3.0	1.9	6.3	6.1	5.5
Model 3 $y_k = \beta_0 + u_d + \beta_2 x_{2k} + \varepsilon_k$						
EBLUP	19.6	19.6	20.2	19.9	19.9	20.5
EBLUPW	3.4	2.9	1.9	6.5	6.4	5.7
NOTE: Variable x_1 is the PPS size variable						

Lessons learned – CASE STUDY 2

- Mean ARB results: Bias can be large for misspecified model
- Mean RRMSE results:
  Squared bias component can still dominate the MSE
  - Can be difficult to obtain proper confidence intervals
- Unequal probability sampling of PPS type can be successfully accounted for in EBLUP with two options:
  - Inclusion of the size variable into the model for model-based EBLUP (Model 2)
  - Use of pseudo EBLUP (EBLUPW) by incorporating design weights in the estimation procedure of the model (all models considered here)

Selected literature

- Battese, G.E., Harter, R.M., and Fuller, W.A. (1988), An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data, *JASA* 80, 28–36.
- Berger, Y.G. (2004). A simple variance estimator for unequal probability sampling without replacement. Journal of Applied Statistics 31, 305-315.
- Berger, Y.G. (2005). Variance estimation with highly stratified sampling designs with unequal probabilities. Australian & New Zealand Journal of Statistics 47, 365-373.
- Berger, Y.G. and C.J. Skinner (2005). A jackknife variance estimator for unequal probability sampling. Journal of the Royal Statistical Society, Series B, 67, 79-89.
- Datta G. (2009). Model-based approach to small area estimation. Chapter 32 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Sample Surveys: Inference and Analysis. Vol. 29B.* New York: Elsevier.

Selected literature (contd.)

- Deville, J.-C. and C.-E. Särndal (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* 87, 376-382.
- Estevao V. M. and Särndal C.-E. (1999) The use of auxiliary information in design-based estimation for domains. Survey Methodology 2, 213-221.
- Fay, R.E., and Herriot, R.A. (1979). Estimates of income for small places: an application of James-Stein procedure to census data. *JASA* 74, 269–277.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science* 9, 55–93.
- Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. Annals of Mathematical Statistics 35, 1491-1523.
- Jiang J. and Lahiri P. (2006). Mixed model prediction and small area estimation. *TEST* 15, 1–96.

Selected literature (contd.)

- Kott, P.S. (2006). Delete-a-group variance estimation for the general regression estimator under Poisson sampling. Journal of Official Statistics 22, 759-767.-14.
- 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. and Pahkinen E. (2004). *Practical Methods for Design and Analysis of Complex Surveys.* Second Edition. Chichester: John Wiley & Sons. Chapter 6.
- Lehtonen R. and Veijanen A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeffermann D. (Eds.). *Handbook of Statistics. Sample Surveys: Inference and Analysis. Vol. 29B.* New York: Elsevier.

Selected literature (contd.)

- 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. IASS Satellite Conference on Small Area Estimation, Riga: Latvian Council of Science, 121-128.
- Lehtonen, R. and Veijanen, A. (2012). Small area poverty estimation by model calibration. *Journal of the Indian Society of Agricultural Statistics*, 66, 125-133.
- Lehtonen and Veijanen (2014). Small area estimation of poverty rate by model calibration and "hybrid" calibration. NORDSTAT 2014 Conference, June 2014, Turku.
- Lehtonen R. and Veijanen A. (2016a) Design-based methods to small area estimation and calibration approach. In: Pratesi M. (Ed.) Analysis of Poverty Data by Small Area Estimation. Chichester: Wiley.

Selected literature (contd.)

- Lehtonen R. and Veijanen A. (2016b) Estimation of poverty rate and quintile share ratio for domains and small areas. In: Alleva G. and Giommi A. (Eds.) Topics in Theoretical and Applied Statistics. New York: Springer.
- Lehtonen, R., Veijanen, A., Myrskylä, M. and Valaste, M. (2011). Small Area Estimation of Indicators on Poverty and Social Exclusion. AMELI WP2 Deliverable 2.2. Available at: http://www.uni-trier.de/index.php?id=24676&L=2
- Molina, I. and Rao, J.N.K. (2010). Small area estimation of poverty indicators. *Canadian Journal of Statistics*, Volume 38, Issue 3, 369–385.
- Montanari, G. E. and M. G. Ranalli (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association* 100, 1429-1442.

Selected literature (contd.)

- Montanari, G.E. and Ranalli, M.G. (2009). Multiple and ridge model calibration. Proceedings of Workshop on Calibration and Estimation in Surveys 2009. Statistics Canada.
- Münnich, R., Zins, S., Alfons, A., Bruch, C., Filzmoser, P., Graf, M., Hulliger, B., Kolb, J.-P., Lehtonen, R., Lussman, D., Meraner, A., Myrskylä, M., Nedyalkova, D., Shoch, T., Templ, M., Valaste, M. and Veijanen, A. (2011): Policy Recommendations and Methodological Report. Research Project Report WP10 of the EU/FP7 AMELI Project.
- Münnich, R., Burgard J.P. and Vogt M. (2009). Small area estimation for population counts in the German Census 2011. JSM 2009, Section on Survey Research Methods.
- Pfeffermann D. (2013). New important developments in small area estimation. *Statistical Science* 28, 40–68.
- Rao J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons.

Selected literature (contd.)

- Särndal, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. Journal of the American Statistical Association 91, 1289-1300.
- Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* 33, 99–119.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.
- Singh, M.P., J. Gambino and H.J. Mantel (1994). Issues and strategies for small area data. Survey Methodology 20, 3-14.
- Torabi, M. and J.N.K. Rao (2008). Small area estimation under a two-level model. *Survey Methodology* 34, 11-17.
- Wu, C. and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *JASA* 96, 185-193.
- Wu C. (2003) Optimal calibration estimators in survey sampling. Biometrika 90, 937–9

Selected literature (contd.)

- Zardetto D. (2015). ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys. JOS 31, 177–203.
  http://dx.doi.org/10.1515/JOS-2015-0013