



HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI

## SOFTWARE DOMEST FOR SMALL AREA ESTIMATION

Risto Lehtonen (University of Helsinki)

Ari Veijanen (Statistics Finland)

**(SAE-kurssi HY 12.4.2011)**

B-N-U Workshop, 23-27 August 2010, Vilnius



## Outline

- Preliminaries
- Key concepts and definitions
- Domest - Technical documentation
- Domest - Worked examples
  
- NOTE: Examples will be worked out during PC session



## Estimation for domains and SAE

### ■ *Domain estimation*

- The estimation of population quantities for the desired population subgroups called domains (large or small)
  - Totals , Means, Proportions
  - Medians, Quantiles, Percentiles
  - More complex indicators...

### ■ *Small area estimation, SAE*

- Estimation for domains whose **sample size** is small or very small (even zero)

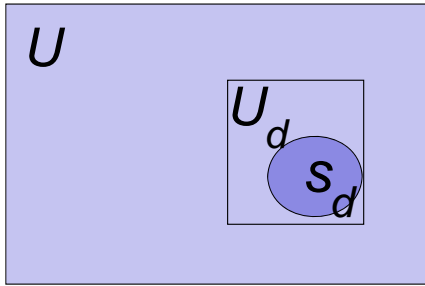
Risto Lehtonen

3



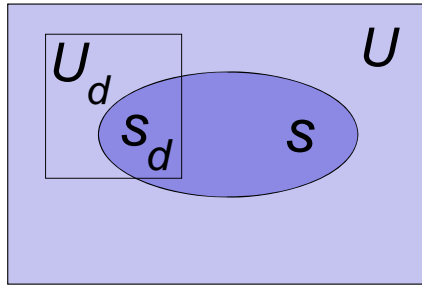
## Domain estimation design

- Types of domains of interest
  - Planned domains / Unplanned domains
- Type of domain estimator
  - Direct / Indirect
  - Design-based / Model-based
- Availability of auxiliary (population) data
  - Unit-level / Aggregated (area-level)
- Type of model
  - Linear / Non-linear
  - Fixed-effects model / Mixed model
  - Generalized linear mixed models (GLMM)
- Accuracy measures
  - Variance estimators / MSE estimators



**Planned domains**

$U$  Population  
 $U_d$  Population domain  $d$   
 Domains = Strata  
 $s_d \subset U_d$  Sample in domain  $d$   
 Sample size  $n_d$  in domain  $d$  is **fixed**  
 $d = 1, \dots, D$



**Unplanned domains**

$U$  Population  
 $s$  Sample  
 $U_d$  Population domain  $d$   
 $s_d = s \cap U_d$  Sample in domain  $d$   
 Sample size  $n_d$  in domain  $d$  is **random**  
 $d = 1, \dots, D$



**Domain type and estimator type**

Domain type	Estimator type	
	Direct	Indirect
<b>Planned</b>	Typical set-up	More rarely
<b>Unplanned</b>	More rarely	Typical set-up

## Design-based properties of estimators

	Design-based methods HT, GREG, MC	Model-based methods SYN, EBLUP, EB
<b>Bias</b>	<b>Design unbiased</b> (approximately) by the construction principle	<b>Design biased</b> Bias does not necessarily approach zero with increasing sample size
<b>Precision (Variance)</b>	<b>Large variance for small domains</b> Variance decreases with increasing sample size	<b>Small variance for small domains</b> Variance decreases with increasing sample size
<b>Accuracy (Mean Squared Error, MSE)</b>	<b>MSE = Variance (or nearly so)</b>	<b>MSE = Variance + squared Bias</b> Accuracy can be poor if the bias is substantial
<b>Confidence intervals</b>	Valid design-based CI can be constructed	Valid design-based CI not necessarily obtained

Risto Lehtonen

7

## Domest software

- Stand-alone interactive Java application
- GUI
- Estimation
  - Domain totals and means
  - Variance estimation
  - MSE estimation
  - Random effects estimation
- Domest documentation (2010)
  - [Part 1](#): Technical documentation
  - [Part 2](#): Worked examples

Risto Lehtonen

8



## Estimation procedure with Domest

- Domest Pages
  - Introduction
  - Session
  - Data
  - Transformations
  - Fixed Effects Model
  - Linear Mixed Model
  - Domain Estimation



## Page Data

- Page Data
  - Import population data set
  - Import sample data set
  - Define level of population data
  - Identify weight variable
  - Define stratum variable (if any)
  - Define sampling design
  - Display data
  - Describe data
- Input data formats
  - Text data set e.g. data.txt
  - SAS data set e.g. data.sas7bdat

Domest - Data

File Help Report Internet

Introduction Session Data Transformations Fixed Effects Model Linear Mixed Model Domain Estimation

**Data**

Import Help

Imported Data Sets

pop

str\_srswor\_sample

Description

Display Data

Delete

Population Data: pop

Sample: str\_srswor\_sample

Subpopulation Size: Unit level data

Design Weights: SamplingWeight

Strata: domain

Sampling Design: S(T)RSWOR

To Transformations

**str\_srswor\_sample**

5 variables, 100 records and 100 observations.

Original data set unknown

Variable	Type	Min	Max	Mean	Missing
SamplingWeig	Number	4.000	20.400	9.660	0
SelectionProb	Number	0.048	0.250	0.135	0
domain	Integer	1.000	10.000	5.500	0
id	Integer	11.000	953.000	475.410	0
r	Number	10.997	31.277	21.200	0
r	Number	14.710	28.663	20.888	0
Weights	Number	4.000	20.400	9.660	0

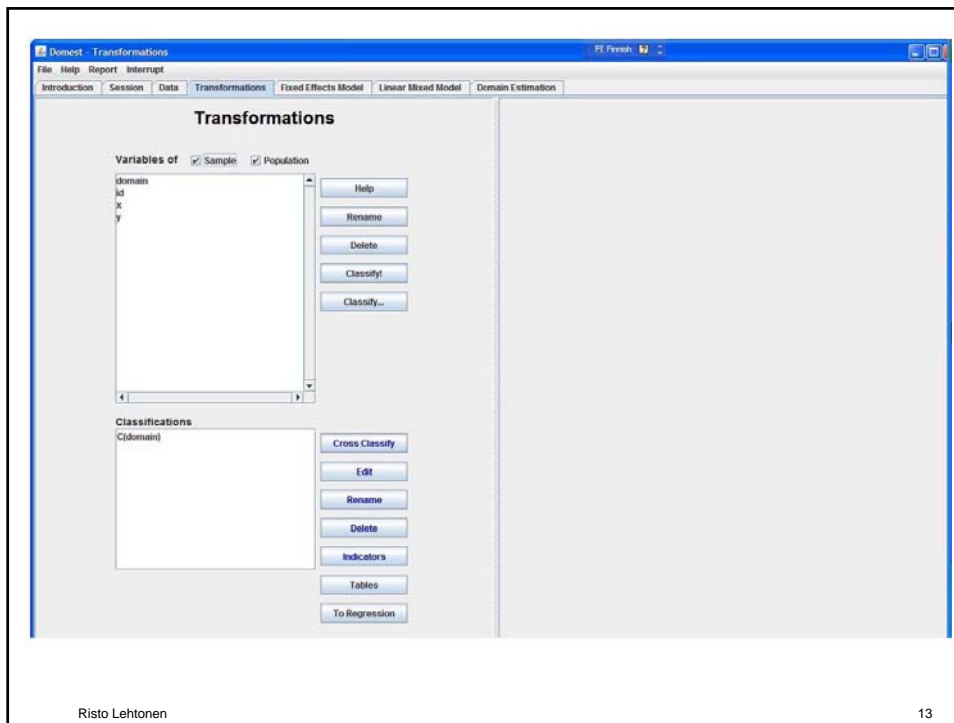
three decimals Scientific notation Report Print Export

Risto Lehtonen 11



## Page Transformations

- Classification of variables
- Renaming variables
- Creating of indicator variables
  
- Function Classify!
  - Creates classified domain variable
  
- Function Indicators
  - Create domain indicators for inclusion of domain-specific intercepts in the fixed-effects model



## Page Fixed Effects Model

- Interactive model fitting
- Identify y-variable
- Include explanatory x-variables in the model
- Use of weights
  - OLS or WLS estimation
- Fit the model
- Show results

Domest - Fixed Effects Model

File Help Report Interrupt

Introduction Session Data Transformations Fixed Effects Model Linear Mixed Model Domain Estimation

**Fixed Effects Model**

Y Variable:   Use Weights

Update t, corr

**Linear Regression Model of y**

Variable	In	Corr. with residual	$\beta$	t
1	<input checked="" type="checkbox"/>	0.000	8.816	11.701
domain	<input type="checkbox"/>	0.048	0.000	1.090
id	<input type="checkbox"/>	0.964	0.000	1.520
x	<input checked="" type="checkbox"/>	-0.143	0.558	16.426

three decimals  Scientific notation


**Linear Regression Model of y**

Linear fixed effects regression model  
 $y = 8.816 + 0.558x + e$ ,  $e^2 = 24.387$   
 Fitted by WLS with weights defined as SamplingWeight.  
 $R^2 = 0.734$   
 $F(\beta_1 = 0) = 269.865$ ,  $df = (1, 98)$

Parameter	Variable	Value	std. error	t-test
$\beta_0$	1	8.816	0.753	-
$\beta_1$	x	0.558	0.034	16.428
$e^2$		24.387	-	-

three decimals  Scientific notation

Risto Lehtonen 15

 **Page Linear Mixed Model**

- Specify random effects
  - Random intercepts
  - Area effects
  - Time effects
  - Time varying effects
- Select estimation method
  - REML, REML-W (weighted REML)
  - ML, ML-W (weighted ML)
- Fit the model
- Show results

Risto Lehtonen 16



**Linear Mixed Model**

Time Effects: [dropdown]  
 Random Intercept: [Select Classification]  
 Area Effects: [dropdown]  
 Time Varying: [Select Classification]  
 Models of Effects...  
 Fit By: REML  
 Fit  
 Show Table  
 To Domain Estimation

**Mixed Model of y**

Linear mixed model  
 $y = 20.588 + u(C(\text{domain})) + \epsilon$   
 $\text{Vari}(u) = 0.837$ ;  $\text{Vari}(\epsilon) = 7.959$   
 Area effects (C(domain)): Independent  
 Fitted by REML. Algorithm converged.

Parameter	Variable	Value	std. error
$\beta_0$	1	20.588	0.404
$\beta_u$	C(domain)	0.105	-
$\sigma^2$		7.959	-

three decimals Scientific notation Report Print Export

Risto Lehtonen 17

## Page Domain Estimation

- Domains
  - Identify domain variable
- Select Model
  - Linear fixed-effects model
  - Linear mixed model
  - Identify weight variable (if any)
- Select Estimators
  - HT, Hajek, GREG, MGREG, EBLUPs, SYN
- Select Statistics
  - Domain totals
  - Domain means
  - Variance and MSE

Risto Lehtonen 18



## Design-based estimators

- Horvitz-Thompson HT estimator
  - Options:
    - Ordinary variance estimator
    - Hájek variance estimator
- Hájek estimator
- GREG estimators
  - Fixed-effects assisting model
  - Options variance estimation
- MGREG estimators
  - Mixed-effects assisting model
  - Options for variance estimation



## Model-based estimators

- EBLUP estimators
  - Empirical Best Linear Unbiased Predictor
- Pseudo EBLUP estimator
- MSE estimation
  - Components  $g_1, g_2, g_3, g_4$  of Mean Cross Product Error matrix MCPE
- (Synthetic estimator)
- NOTE: Weights can be incorporated into the mixed model estimation procedure

**Table 1.** Summary of estimators for domains: Estimator and model types and the use of weights in the estimation procedure.

Domain estimator	Weights in domain estimator	Assisting / Underlying model	Weights in model fitting	Model fitting method
<b>Design-based estimators</b>				
HT	Yes	None	-	-
Hájek	Yes	None	-	-
GREG	Yes	Linear fixed-effects model	Yes	WLS
GREG-OLS	Yes	Linear fixed-effects model	No	OLS
MGREGW	Yes	Linear mixed model	Yes	ML-W or REML-W
MGREG	Yes	Linear mixed model	No	ML or REML

Risto Lehtonen

21

**Table 1.** Summary of estimators for domains: Estimator and model types and the use of weights in the estimation procedure.

Domain estimator	Weights in domain estimator	Assisting / Underlying model	Weights in model fitting	Model fitting method
<b>Model-based estimators</b>				
SYN	No	Linear fixed-effects model	No	OLS
SYNW	No	Linear fixed-effects model	Yes	WLS
EBLUP(Y)	No	Linear mixed model	No	ML of REML
EBLUP(Y)W	No	Linear mixed model	Yes	ML-W or REML-W
EBLUP( $\mu$ )	No	Linear mixed model	No	ML or REML
EBLUP( $\mu$ )W	No	Linear mixed model	Yes	ML-W or REML-W
Pseudo-EBLUP	Yes	Linear mixed model	No	ML or REML
Pseudo-EBLUPW	Yes	Linear mixed model	Yes	ML-W or REML-W

Risto Lehtonen

22

Risto Lehtonen

23

## References

- 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.
- 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.
- Rao J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons.
- Saei, A. and R. Chambers (2004). Small area estimation under linear and generalized linear mixed models with time and area effects. EURAREA Consortium 2004, *Project Reference Volume*, [www.statistics.gov.uk/eurarea](http://www.statistics.gov.uk/eurarea).
- Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* 33, 99–119.