

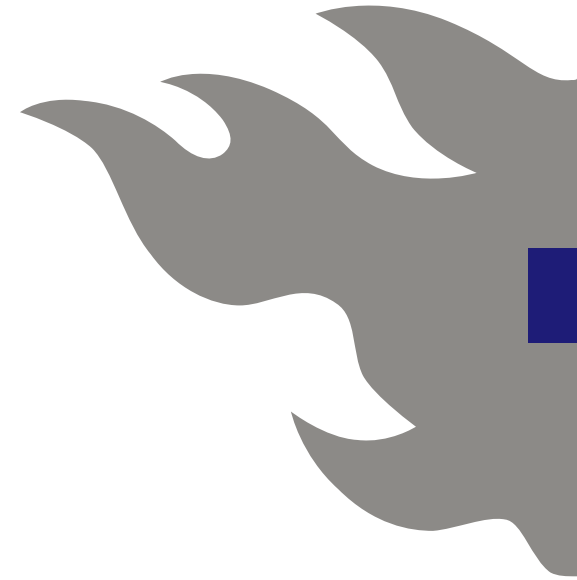


HELSINGIN YLIOPISTO  
HELSINGFORS UNIVERSITET  
UNIVERSITY OF HELSINKI

# **Pienalue-estimointi (78189)**

## **Kevät 2011**

Risto Lehtonen  
Helsingin yliopisto





# Pienalue-estimointi

■ Kurssin kotisivu

<http://wiki.helsinki.fi/pages/viewpage.action?pageId=62430039>



# Hyödyllisiä taustatietoja

## ■ Otantamenetelmät

- Lehtonen R. and Djerf K. (2008). *Survey sampling reference guidelines*. Luxembourg: Eurostat Methodologies and Working papers
- Saatavilla vapaasti osoitteessa:  
[http://epp.eurostat.ec.europa.eu/cache/ITY\\_OFFPUB/KS-RA-08-003/EN/KS-RA-08-003-EN.PDF](http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-RA-08-003/EN/KS-RA-08-003-EN.PDF)

## ■ Tilastollisten mallien perusteita

- [Lineaariset mallit](#)
- [Yleistetyt lineaariset mallit](#)



# Maailmanlaajuinen trendi

- Yhteiskunnassa on lisääntyvä tarve tuottaa luotettavia tietoja alueellisille ja muille populaation (perusjoukon) osajoukoille
  - *Estimation for domains*
  - *Small area estimation*
- EU:n tutkimuksen puiteohjelmien projekteja
  - EURAREA Project (2001-2004)  
<http://www.statistics.gov.uk/eurarea/>
  - AMELI Project (2008-2011)  
[http://www.uni-trier.de/index.php?id=24474&no\\_cache=1](http://www.uni-trier.de/index.php?id=24474&no_cache=1)
  - SAMPLE Project (2008-2011)  
<http://www.sample-project.eu/>



# SAE Conferences

## ■ EWORSAE

European Working Group on Small Area Estimation <http://sae.wzr.pl/>

SAE2005 (University of Jyväskylä)

SAE2007 (University of Pisa)

SAE2009 (University of M. Hernandez, Elche)

Forthcoming:

SAE2011 (University of Trier)

SAE2013 (Poland)



# Kirjallisuutta

■ [Rao J.N.K.](#) (2003). *Small Area Estimation*. New York: John Wiley & Sons.

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

<http://books.google.fi/books?id=Xb-m5Xg74F4C>

## Web extension:

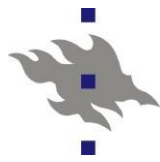
VLISS-Virtual Laboratory in Survey Sampling

<http://mathstat.helsinki.fi/VLISS/>



# Kirjallisuutta

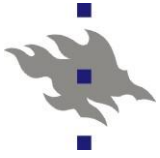
- Lehtonen R. and Djerf K. (eds.) (2001). *Lecture Notes on Estimation for Population Domains and Small Areas*. Statistics Finland: Reviews 2001/15.
- Lehtonen R. and Veijanen A. (2009). Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C.R. and Pfeiffermann D. (Eds.). *Handbook of Statistics. Vol. 29B. Sample Surveys: Inference and Analysis*. New York: Elsevier.



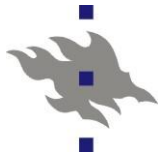
# SAE: Laskentatyökaluja

- SAS 9.2
  - Procedures SURVEYMEANS
  - Procedure SURVEYREG
    - DOMAIN-lause
- EURAREA-projekti
  - SAS-makro Standard estimators
  - SAS-makro EBLUPGREG
- Ohjelma [DOMEST](#)
  - Ari Veijanen & Risto Lehtonen
- R-kielisiä ohjelmia





# **JOHDANTO LÄHESTYMISTAPOJA**



# Käsitteitä ja määritelmiä

- Perusjoukon osajoukko *Domain*
  - Pienalue *Small area*
    - Lääni, maakunta, seutukunta, kunta...
    - Väestön demografiset ja sosioekonomiset osajoukot
    - Yritysten toimialakohtaiset osajoukot
  - Estimoidaan otosaineiston perusteella tunnuslukuja:
    - Kokonaismääriä *Totals*
    - Keskiaroja *Means*
    - Osuuksia *Proportions*
    - Mediaaneja...
- määritellyille perusjoukon osajoukoille  
(*domains, small areas*)



# Erikoistapaus - SAE

## ■ *Small area estimation, SAE*

- Estimointi tilanteessa, jossa osajoukkojen otoskoko on pieni

- Vaihtoehtoinen määritelmä (Partha Lahiri):

Small area = Domain of interest, for which the sample size is not adequate to produce reliable **direct estimates**



## Tyypillinen estimointitehtävä

**Määritellään ja identifioidaan osajoukot**

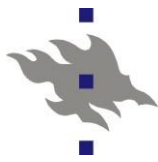
Osajoukkojen  $U_d$  lkm  $D$  on yleensä suuri

**Spesifioidaan tulosmuuttujan  $y$  parametrit**

Osajoukkototaalit  $t_d = \sum_{k \in U_d} y_k$

Keskiarvot  $\bar{Y}_d = t_d / N_d, d = 1, \dots, D$

missä  $N_d$  on osajoukon koko pj:ssa



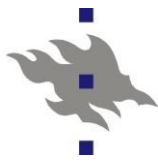
# Esimerkkejä

- Työttömien kokonaismäärän estimointi alueittain sukupuolen ja ikäryhmän mukaan muodostetuissa osajoukoissa
  - Tilastokeskuksen työvoimatutkimuksen aineisto
- Kotitalouksien käytettävissä olevien tulojen mediaanin estimointi kunnittain
  - EU:n SILC-tutkimusaineisto
- Alueellisten köyhyysasteiden estimointi
  - EU:n SILC-tutkimusaineisto



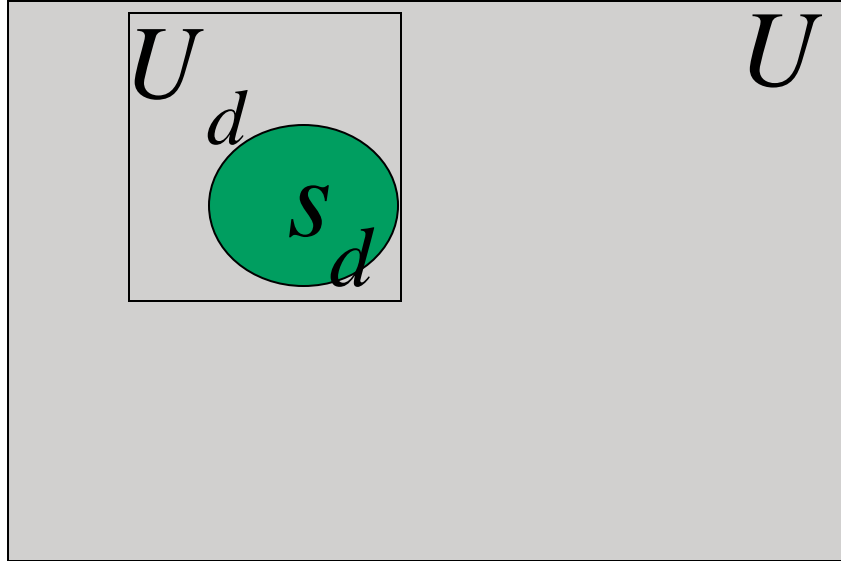
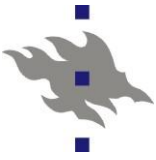
# Tärkeitä kysymyksiä

- Osajoukkojen tyyppi?
  - Suunnitellut / Ei-suunnitellut osajoukot
  - *Planned domains / Unplanned domains*
- Aineistolähteet?
  - Otosaineisto
  - Lisäinformaatio
- Estimaattorin tyyppi?
  - Suora / Epäsuora *Direct / Indirect*
  - Asetelmaperusteinen / Malliperusteinen
  - *Design based / Model-based*
- Mallin tyyppi?
  - Lineaarinen / Epälineaarinen *Linear / Non-linear*
  - Kiinteät vaikutukset / Sekamallit (*Mixed models*)



# Osajoukon tyyppi

- Suunnitellut osajoukot *Planned domains*
  - Tärkeimmät osajoukkotyypit pyritään määrittelemään otanta-asetelmassa **ositeiksi** (*strata*)
  - Osajoukkojen otoskoot on kiinnitetty
  - Otoskoot hallitaan kiintiöintimenetelmillä (*allocation*)
  - Liian pienet otoskoot voidaan välttää
  
- Ei-suunnitellut osajoukot *Unplanned domains*
  - Osajoukkojen otoskoot ovat satunnaismuuttujia
  - Voi tulla osajoukkoja joiden otoskoko pieni
  - Käytännössä yleinen tilanne (Miksi?)



## Planned domains

$U$  Population

$U_d$  Population domain  $d$

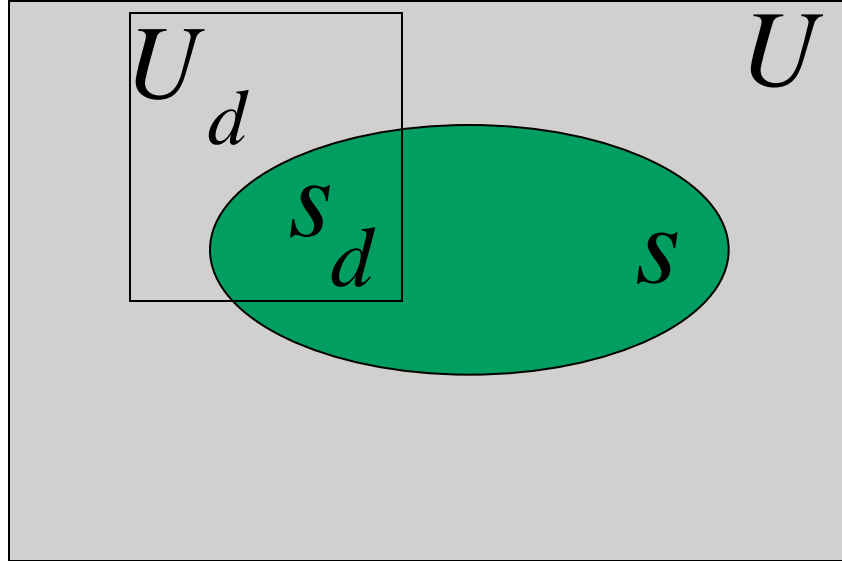
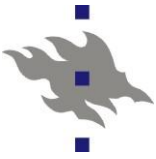
Domains = Strata

$s_d \subset U_d$  Sample in domain  $d$

Sample size  $n_d$  in domain  $d$  is fixed

$d = 1, \dots, D$





## Unplanned domains

$U$  Population

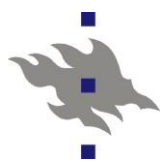
$s$  Sample

$U_d$  Population domain  $d$

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

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

$d = 1, \dots, D$



# Suora ja epäsuora estimaattori

- Suora estimaattori *Direct estimator*
  - *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)
  - **Suunniteltujen** osajoukkorakenteiden tilanne
  
- Epäsuora estimaattori *Indirect estimator*
  - *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
  - **Ei-suunniteltujen** osajoukkorakenteiden tilanne



# Esimerkki: Suora HT-estimaattori

Asetelma-aperusteinen Horvitz-Thompson  
estimaattori

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

käyttää  $y$ -arvoja vain osajoukosta  $s_d$

$\hat{t}_{dHT}$  on suora estimaattori

# Esimerkki:

## Suora GREG-estimaattori

Suora asetelmaperusteinen malliavusteinen GREG-estimaattori

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

käyttää lineaarisia malleja jotka on spesifioitu erikseen kullekin osajoukolle:

$$Y_k = \mathbf{x}'_k \boldsymbol{\beta}_d + \varepsilon_k, \quad k \in U_d, \quad d = 1, \dots, D$$

missä  $\boldsymbol{\beta}_d$  on osajoukkokohtainen, ja

$$\hat{y}_k = \mathbf{x}'_k \hat{\boldsymbol{\beta}}_d \text{ ovat sovitteita, laskettu jokaiselle } k \in U_d$$

# Esimerkki:

## Epäsuora GREG-estimaattori

Epäsuora asetelmaperusteinen  
GREG-estimaattori

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

käyttää lineaarista mallia

$$Y_k = \mathbf{x}'_k \boldsymbol{\beta} + \varepsilon_k, \quad k \in U$$

jossa vektori  $\boldsymbol{\beta}$  on yhteinen kaikille osajoukoille

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

# Esimerkki:

## Epäsuora SYN-estimaattori

Epäsuora malliperusteinen synteettinen SYN-estimaattori

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

käyttää lineaarista mallia

$$Y_k = \mathbf{x}'_k \boldsymbol{\beta} + \varepsilon_k, \quad k \in U$$

jossa  $\boldsymbol{\beta}$  on yhteinen kaikille osajoukoille

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



# “Voiman lainaaminen”

## “*Borrow strength*”

- Epäsuorat estimaattorit pyrkivät “lainaamaan voimaa”
  - Muista osajoukoista (spatiaalinen dimensio)
  - Saman osajoukon aikaisemmista mittauksista (temporaalinen dimensio)
- Tyypillistä erityisesti pienten osajoukkojen tilanteissa (pieni otoskoko)
- “Borrowing strength” käytetään usein malliperusteisissa SAE-tilanteissa



# Estimointitehtävä

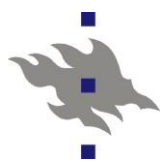
- **Suuri osajoukko – Large domain**
- Osajoukko jossa on mahdollista tuottaa riittäväällä tarkkuudella **asetelmaperusteinen suora** (*direct*) estimaatti
- **Tämän kurssin alue:**  
***Estimation for domains and small areas***
- **Pieni osajoukko – Small domain**
- Pieni osajoukko = Osajoukko jossa **ei ole** mahdollista tuottaa riittäväällä tarkkuudella **asetelmaperusteinen suora** (*direct*) estimaatti
- Tarvitaan **malliperusteisia epäsuoria** (*indirect*) estimaattoreita
  - ”Voiman lainaaminen”
  - *Borrowing strength*





# Lisäinformaatio

- Kaikissa tarkasteltavissa menetelmissä on olennaista:
  - Perusjoukkoa koskevan lisäinformaation hyvä saatavuus
    - *Auxiliary data, auxiliary information*
    - Rekistereistä saatavat lisätiedot, apumuuttujat
  - Lisäinformaation tuonti estimointiproseduriin tilanteeseen soveltuvien tilastollisten mallien avulla
    - Lineaariset mallit, logistiset mallit, sekamallit
    - Yleistetyt lineaariset sekamallit (Generalized linear mixed models)



# Asetelmaperusteiset menetelmät

## ■ Asetelmaperusteiset suorat estimaattorit

- Horvitz-Thompson (HT) –estimaattorit
- Hájek-tyyppinen estimaattori

## ■ Asetelmaperusteiset malliavusteiset estimaattorit

- Suoria tai epäsuoria estimaattoreita
- Yleistetyt regressioestimaattorit (generalized regression estimators) GREG
- Kalibrointiestimaattorit
  - Särndal, Swensson and Wretman (1992)
  - Lehtonen, Särndal and Veijanen (2003, 2005)
  - Lehtonen and Pahkinen (2004), luku 6
  - Lehtonen and Veijanen (2009)



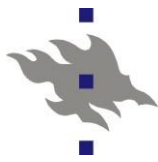
# Malliperusteiset menetelmät

- **Synteettiset estimaattorit SYN**
- **EBLUP- ja EBP-estimaattorit**
  - *Empirical Best Linear Unbiased Predictor*
  - *Empirical Best Predictor*
  - Rao (2003)
  - EURAREA-projekti, Domest-ohjelma
- **Bayes-menetelmät**
  - Empirical Bayes, Hierarchical Bayes
- **“Poverty mapping”**
  - World Bank, Peter Lanjouw, Chris Elbers,...
  - [PovMap Software](#)



# HUOM: Tilastollisen mallin rooli

- **Asetelmaperusteinen GREG**
  - Malleja käytetään avustavina työkaluina
  - GREG-estimaattorit ovat **malliavusteisia** (*model-assisted*)
  
- **Malliperusteinen SYN**
  - Nojautuu kokonaisuudessaan tilastolliseen malliin
  - SYN-estimaattorit ovat **malliperusteisia** (*model-based*) tai **mallisidonnaisia** (*model-dependent*)



# Estimaattoreiden ominaisuuksia

## ■ Table 1

- Asetelmaperusteiset estimaattorit HT, GREG
  - Likimain harhattomia
  - Varianssi voi kasvaa suureksi, jos osajoukon otoskoko on pieni
    - Varianssi pienenee otoskoon kasvaessa
- Malliperusteiset estimaattorit SYN, EBLUP, EBP
  - Harhaisia määritelmän mukaan (*All models are wrong but some are useful*).
    - Harha **ei pienene** otoskoon kasvaessa!
  - Varianssi voi olla pieni myös pienissä osajoukoissa
  - MSE voi olla suuri jos harha on dominoiva

# Estimaattoreiden tilastollisten ominaisuuksien vertailu

Osajoukkojen totaalien  $t_d$   
estimaattoreiden vertailu:

Harha:

*Bias*  $Bias(\hat{t}_d) = E(\hat{t}_d) - t_d$

Varianssi:

*Precision*  $Var(\hat{t}_d) = E(\hat{t}_d - E(\hat{t}_d))^2$

Keskineliövirhe:

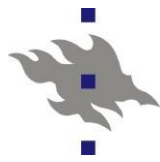
*Accuracy*

$$MSE(\hat{t}_d) = E(\hat{t}_d - t_d)^2 = Var(\hat{t}_d) + Bias^2(\hat{t}_d)$$

# 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



# Estimaattoreiden työnjako

- Asetelmaperusteisia estimaattoreita (HT, GREG) käytetään tyypillisesti suurille osajoukoille (suuri otoskoko, pieni varianssi).
- Malliperusteisia estimaattoreita (SYN, EBLUP) käytetään pienille osajoukoille, (pieni otoskoko, pieni varianssi) joissa asetelmaperusteiset estimaattorit toimivat huonosti (suuri varianssi).





# Natural application areas of estimation approaches by domain sample size

ESTIMATION APPROACH	DOMAIN SAMPLE SIZE		
	Minor	Medium	Major
Model-based			
<b>Synthetic SYN</b>	<b>++</b>	<b>+</b>	<b>0</b>
<b>EBLUP, EBP</b>	<b>+++</b>	<b>++</b>	<b>++</b>
Design-based			
<b>Horvitz-Thompson HT</b>	<b>0</b>	<b>+</b>	<b>++</b>
<b>GREG, MC</b>	<b>+</b>	<b>++</b>	<b>+++</b>

Applicability

0 Not at all

+ Low

++ Medium

+++ High



## Selected literature - Design-based

- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer.
- 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 - Design-based

- Lehtonen, R. and Veijanen, A. (1998). Logistic generalized regression estimators. *Survey Methodology* 24, 51–55.
- 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.
- Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology* 33, 99–119.



## Selected literature - Model-based

- Rao J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons.
- Longford N. (2005). *Missing Data and Small-area Estimation: Modern Analytical Equipment for the Survey Statistician*. New York: Springer.
- 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.



## Selected literature - Model-based

- 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.
- Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: an appraisal. *Statistical Science* 9, 55–93.
- Jiang J. and Lahiri P. (2006). Mixed model prediction and small area estimation. *TEST* 15, 1–96.