HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Small Area Estimation
Spring 2015

CASE STUDY 1
SAE in the SILC data

Risto Lehtonen, University of Helsinki

Estimation of mean of "Perceived income" for regional domains

- Source: 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 problem

- Estimation of mean perceived income for regions in Finland
- Regions: D = 70 NUTS3 areas
- Target population: N about 4,3 million

- Sizes of regions vary:
  - Smallest: about 2000 persons
  - Largest: about 1 million persons

EU-SILC data of Finland (2009)

- Sample size n = 11,000 households
- Interview data (CAPI)
- Respondent: Household head
- Stratified unequal probability sampling
- Reweighting to adjust for unit nonresponse
- Model-free calibration for final weights

- Domains are of unplanned type
  - Smallest domain sample size: 10
  - Largest domain sample size: 2425

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

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
  - Education
  - Actual (register) income
  - Socio-economic status
  - Stage in life of household-dwelling unit
- Categorical variables are transformed to indicator variables
- 16 predictor variables in the regression model
- All variables statistically significant
- R squared = 15%

Models

Linear fixed-effects model

$$y_k = \beta_0 + \beta_1 x_k + ... + \beta_J x_{Jk} + \varepsilon_k,\ \varepsilon_k \sim N(0, \sigma^2)$$

where beta coefficients are common for all domains

Linear mixed model

$$y_k = \beta_0 + u_d + \beta_1 x_k + ... + \beta_J x_{Jk} + \varepsilon_k$$

with domain-level random intercepts u_d

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

Estimators

HT estimator for domain means

$$\hat{t}_{dHT} = \sum_{k \hat{I} s_d} w_k y_k, \quad d = 1,...,D$$

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

GREG estimators for domain means

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

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

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

GREG estimators

GREG assisted by linear fixed-effects model

Model fitted by ML

Predicted values

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

MGREG assisted by linear mixed model

Model fitted by REML

Predicted values

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

Variance estimators (unplanned domains)

HT estimator for domain means

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

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

GREG estimators for domain means

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

where $e_{dk} = I\{k \in U_d\}e_k$ are extended residuals

Quality indicators

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

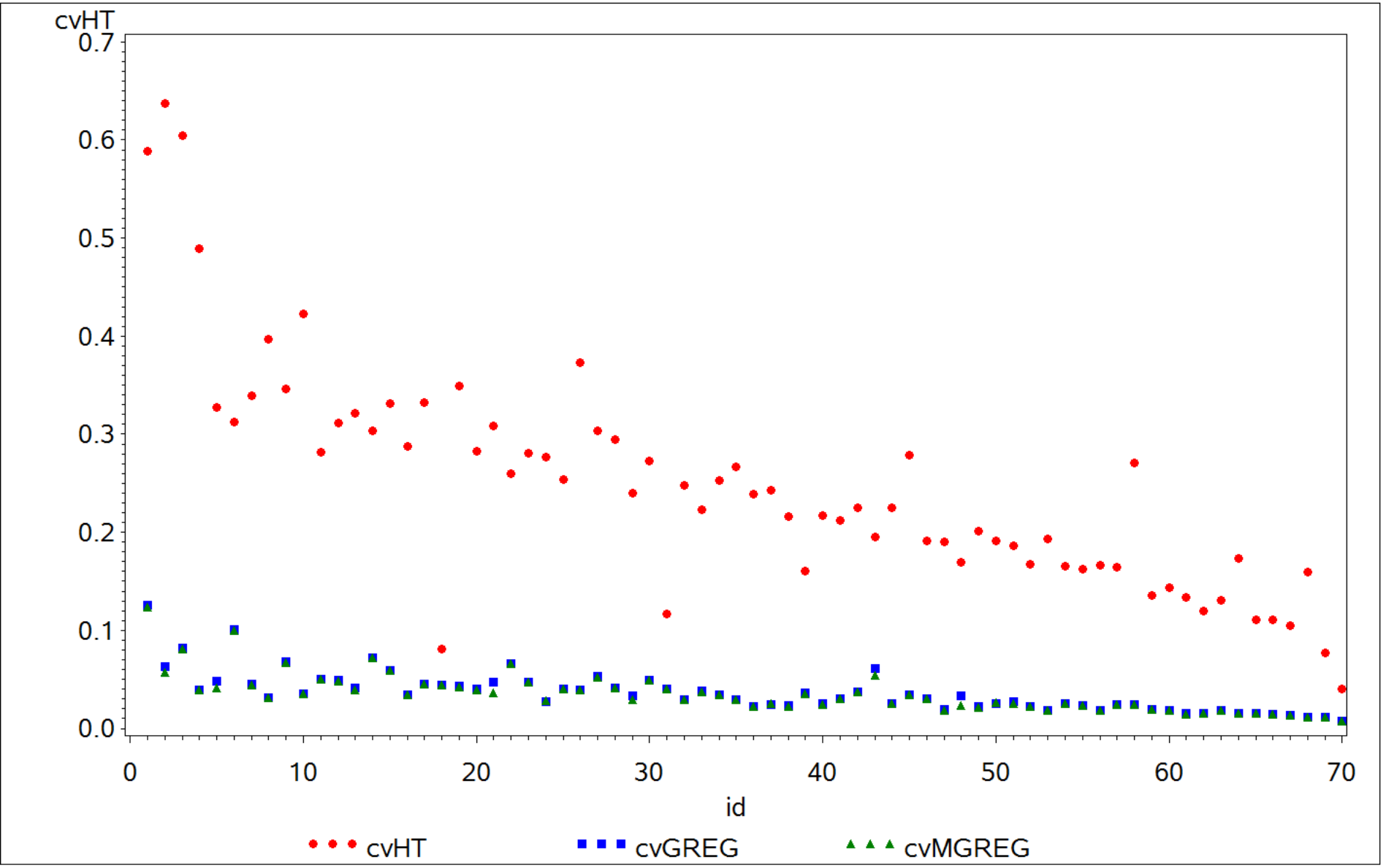
$$\text{s.e}\left(\hat{y}_d\right) = \sqrt{\hat{V}\left(\hat{y}_d\right)}$$

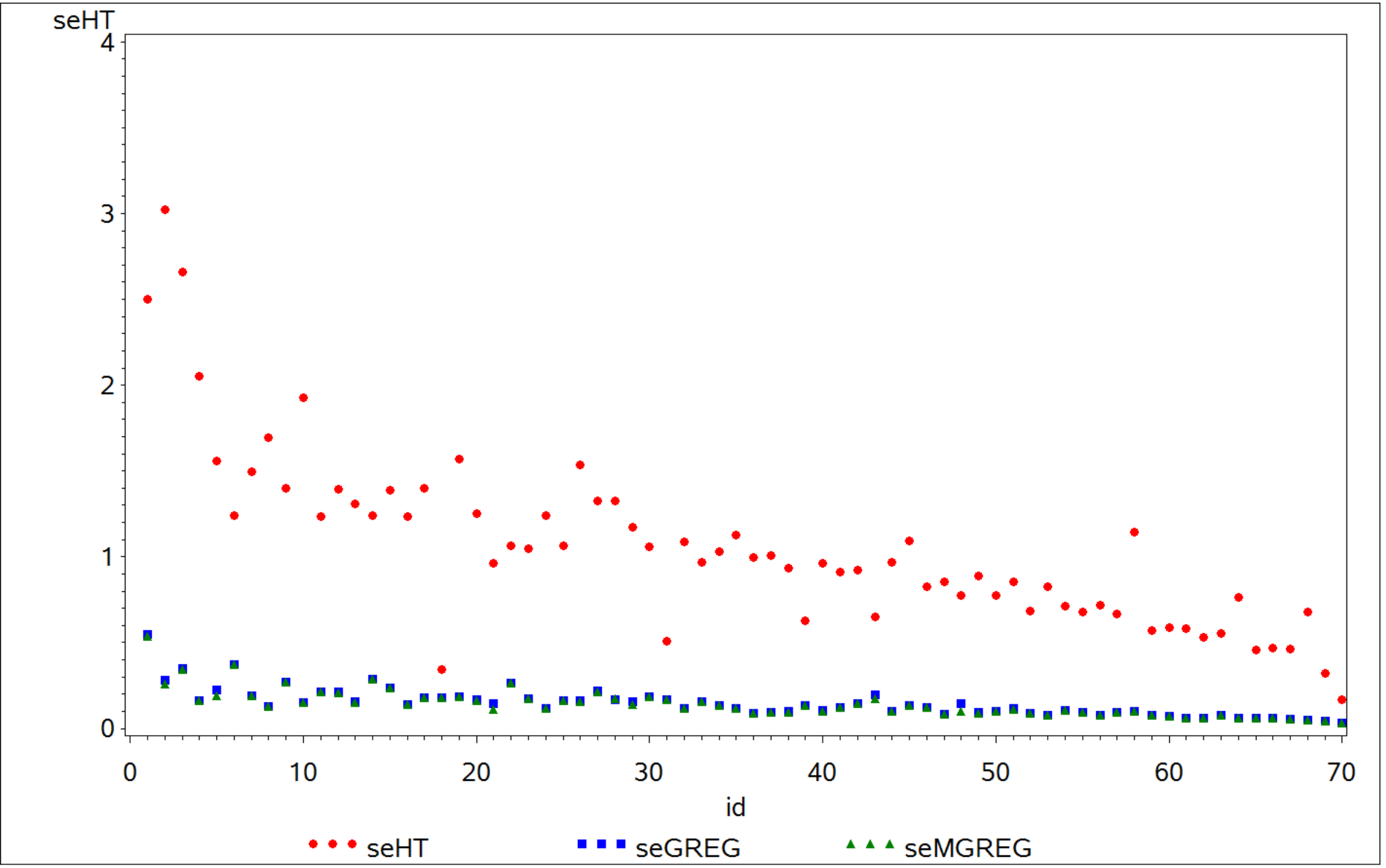
Coefficient of variation of domain mean estimate $\hat{y}_d$

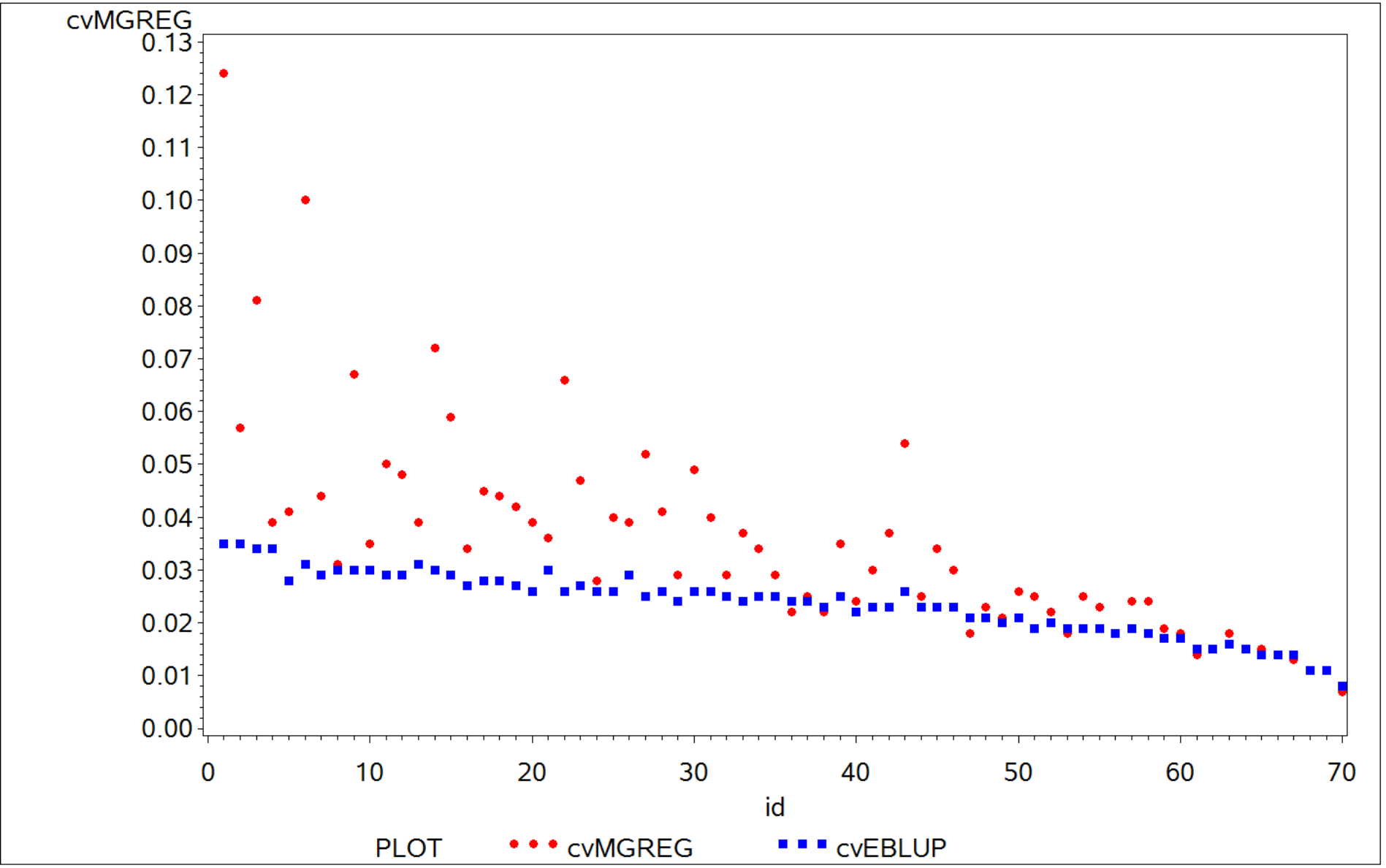
$$\text{cv}\left(\hat{y}_d\right) = \frac{\text{s.e}\left(\hat{y}_d\right)}{\hat{y}_d}$$

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


seHT
4
3
2
1
0
0
10
20
30
40
50
60
70
id
seHT
seGREG
seMGREG


cvMGREG
0.13
0.12
0.11
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50
60
70
id
PLOT
cvMGREG
cvEBLUP

Points for discussion

- Strategies in sampling design phase
- Strategies in estimation phase
- NOTE: Key feature: Clever use of auxiliary data and modelling!