Comparison of Energy Resource Survey Results of 2010 and 2011

Baiba Buceniece
Central Statistical Bureau of Latvia

About the survey

- Annual enterprise survey of energy resource acquisition and consumption.
- Stratified simple random sample.
- Sample size:
  - for 2010 - 5506;
  - for 2011 - 5508.
- Main variables of interest:
  - amount of received heat,
  - consumption of electricity,
  - consumption of petrol,
  - consumption of diesel oil,
  - consumption of natural gas.

Problems

- Lack of "good" auxiliary information:
  - in sample selection stage - weak correlation between survey variables and available stratification variables.
  - in estimation stage - auxiliary information for calibration available only in aggregated level;
- Large amount (674) of different very specific survey variables.
- Many "0" values of survey variables.
- Many domains of interest, which can not be planed in sample selection stage.

Sampling frame

- Consists of:
  - economically active merchants,
  - state and municipal budget authorities,
  - agricultural and fish farms with $>= 10$ employees.
- Not included:
  - individual merchants,
  - public organizations,
  - agricultural and fish farms with < 10 employees.
- Size of the frame:
  - for 2010 - 60321 units,
  - for 2011 - 63565 units.

Stratification

- Based on:
  - economic activity (NACE Rev.2),
  - turnover,
  - number of employees (in NACE section "O" - Public administration and defence; compulsory social security, where turnover is missing).

Sample allocation for 2010

- Calculated using Neyman allocation:

$$n_h^{Neyman} = n_d^{min} \times \frac{N_h S_h}{\sum_{h=1}^{L_d} N_h S_h}, \quad (1)$$

  - N_h^{Neyman} - population size of strata
  - n_d^{min} - minimum sample size of domain
  - L_d - number of stratas in domain
  - $S_h = \sqrt{\frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_i - \bar{Y}_h)^2}$ is standard deviation of enterprises **turnover** (or number of employees in stratas of NACE section "O").
- Optimization of sample allocation is done in order to achieve smallest sample size n_d of domain d, $(d = 1, \ldots, D)$ which allows not to exceed predefined sampling error (CV).
- For each domain $CV_{max} = 4\%$ for turnover (or number of employees).

Calculation of weights for 2010

- Several sets of weights are computed:
  - weights for variable "amount of received heat";
  - weights for variable "consumption of natural gas";
  - one set of weights for two variables "consumption of petrol" and "consumption of diesel fuel";
  - weights for variable "consumption of electricity" (weights are calibrated. Hence, the estimated variances are equal to 0 in domains, that match the domains used for calibration);
  - one set of weights for many variables associated with consumption of fuelwood;
  - one set of weights for other survey variables.
- All sets of weights, except for "electricity", are calculated as design weights adjusted by nonresponse and taking into account outliers defined for each set of variables separately.

Horvitz-Thompson (HT) estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{n^R} y_i w_i$$

- n^R - number of respondents
- N_h - population size of strata h
- y_i - value of study variable of unit i
- w_i - weight of unit i
- Variance estimator:

$$\hat{V}_{HT}(\hat{Y}) = \sum_{h=1}^{H}\left(1 - \frac{n_h^R}{N_h}\right)\frac{n_h^R}{n_h^R - 1}\sum_{i=1}^{n_h^R}\left(w_i y_i - \frac{1}{n_h^R}\sum_{i=1}^{n_h^R} w_i y_i\right)^2$$

Sample allocation for 2011

- Calculated using survey data from 2010.
  1. Stratification intended for 2011 is applied to sampling frame of 2010;
  2. g stratas are established $(g = 1, \ldots, G)$;
  3. Population size M_g of each strata g is calculated;
  4. Variances of 5 main variables of interest are estimated for each strata g:

$$s_g^2 = \frac{1}{M_g - 1}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi}^2 z_{hi} - \frac{M_g}{M_g - 1}\left(\left(\frac{1}{M_g}\sum_{h=1}^{H}\frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi} z_{hi}\right)^2 - \frac{1}{M_g^2}\sum_{h=1}^{H} N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right)\frac{1}{n_h - 1}\sum_{i=1}^{n_h}\left(y_{hi} z_{hi} - \frac{1}{n_h}\sum_{t=1}^{n_h} y_{ht} z_{ht}\right)^2\right), \quad (2)$$

  where $z_{hi} = \begin{cases} 1, hi \in \theta_g \\ 0, hi \notin \theta_g \end{cases}$; θ_g is set of indices of observed units in 2010 belonging to the strata g. To obtain the estimates s_g^2 following two conditions must be met: $n_h > 1$, $\forall h$ and $\theta_g \neq \emptyset$, $\forall g$.
  5. "Optimized" Neyman allocation is used for calculation of sample sizes for each of 5 main variables:
     - $CV_{max} = 25\%$ is set for domains;
     - in formula (1) S_h is replaced with $\sqrt{s_g^2}$ calculated by (2).
  6. 5 different sample sizes for each strata are obtained;
  7. The final sample size for each strata is calculated as an average of these 5.

Calculation of weights for 2011

- Only one set of weights is computed for all survey variables:
  1. design weights are adjusted taking into account nonresponse and frame changes;
  2. weights are calibrated using auxiliary information about delivered electricity and natural gas.

GREG estimator

- $\hat{Y}_{GREG} = \sum_{i=1}^{n^R} y_i w_i g_i$
- Variance estimator:

$$\hat{V}_{GREG}(\hat{Y}) = \sum_{h=1}^{H}\left(1 - \frac{n_h^R}{N_h}\right)\frac{n_h^R}{n_h^R - 1}\sum_{i=1}^{n_h^R}\left(w_i g_i e_i - \frac{1}{n_h^R}\sum_{i=1}^{n_h^R} w_i g_i e_i\right)^2$$

- Residual estimator:

$$\hat{e_i} = y_i - X_s \times \left(\left(X_s \times w_i\right)^T \times X_s\right)^{-1}\left(X_s \times w_i\right)^T \times y_i$$

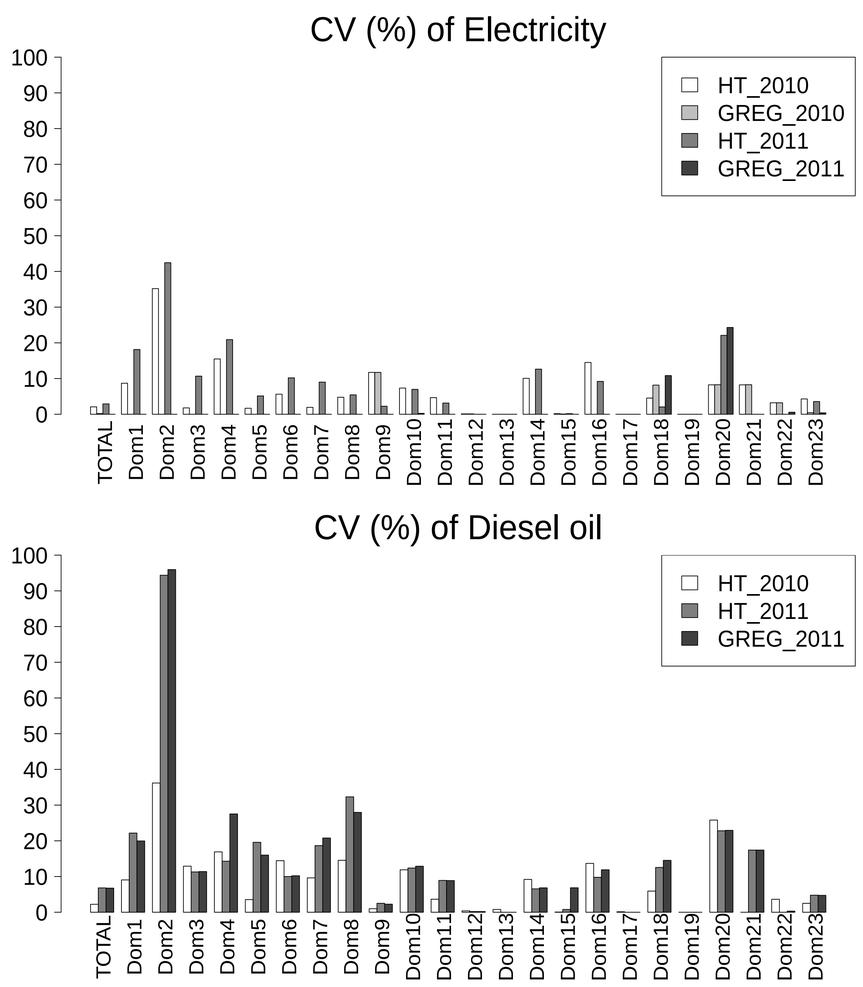
Conclusions - options of improvement

- In many cases (for domains and for totals) GREG estimator is less precise than HT estimator.
- Estimated sample size for 2011 is ≈12 000 to get CV (in domains) of main variables $<= 5\%$.
- Looking for better auxiliary information for calibration.
- Synchronisation of domains used for sample selection and domains used for publication of survey results.

Variances of "Petrol"

	2010	2011	
	V_{HT}	V_{HT}	V_{GREG}
TOTAL	7 129 774	4 424 207	5 460 166
Dom1	104 688	370 220	355 336
Dom2	55	22 623	20 723
Dom3	4 950	474	8 029
Dom4	1 458	167	203
Dom5	20 963	173 809	151 518
Dom6	9 493	18 899	18 433
Dom7	60 147	94 341	106 609
Dom8	12 731	1 413	2 422
Dom9	1 044	2 835	2 821
Dom10	8 600	6 492	6 028
Dom11	15 538	96	199
Dom12	4	24	21
Dom13	0	0	0
Dom14	14 918	68 000	67 894
Dom15	5	8 707	10 783
Dom16	944 449	393 381	1 086 036
Dom17	187	0	0
Dom18	1 105 890	114 792	184 583
Dom19	0	0	0
Dom20	7	1	1
Dom21	0	0	0
Dom22	0	0	0
Dom23	4 835 258	3 230 901	3 484 320

Coefficients of variation (%) of some variables



References

- Lapins, J. (1997). Sampling surveys in Latvia: Current situation, problems and future developments. *Statistics in Transition: Journal of the Polish Statistical Association* **3**, 281 - 292.
- Särndal, C. Lundström S. (2001). *"Estimation in the presence of nonresponse and frame imperfections"*. Statistics Sweden.
- Osier, G. The linearisation approach implemented by EUROSTAT for the first wave of EU-SILC: What could be done from the second wave onwards? Luxembourg Income Study (LIS), Institut National de la Statistique et des Etudes Economiques (STATEC Luxembourg) and Net-SILC2. Paper prepared for *the workshop on standard error estimation and other related sampling issues* (Eurostat, 29-30 March 2012)

e-mail: baiba.buceniece@csb.gov.lv