Combining samples of EU-SILC and HBS in Estonia

Julia Aru

Statistics Estonia, Tartu University

Workshop of BNU Network on Survey Statistics 2012

Outline

Rationale for combining samples

- Surveys focus on different topics but allways have an overlap in questions asked
- For common questions, more detailed and more precise estimates can be computed from the combined sample of two surveys
- Harmonisation: different estimates for one phenomenon confuse users
- Possibilities to improve precision of survey-specific estimates

HBS

Household Budget Survey

- Focus on expenditures
- Cross-sectional
- Diary-type survey, difficult for responders
- Low response rates
- In 2010: ca 3600 households responded

EU-SILC

Survey on Income and Living Conditions

- Focus on income and living conditions
- Rotational design: every household is asked to participate for 4 years, every year a new sample is added
- Sample of a single year consists of the new part and the repeated part
- Response models are very different for these parts (different behaviour and information avalaible)
- In 2010: ca 5000 hhs responded

Common features of HBS and EU-SILC

- Target population is all persons in private households
- Stratified systematic sampling of persons, whole household is interviewed
- Non-response correction with logistic regression (but different predictors)
- Calibration by county and sex-age group

Weighting methods

- Simple: adjust existing survey specific weights with a scaling factor
- Difficult: calculate the probabilities to be included into the combined sample from the beginning
- But do estimates differ, including variance estimates? Is difficult method worth the effort?

Adjusting existing weights

- multiply EU-SILC weights by α, HBS weights by $1-\alpha$, then re-calibrate
- There are many methods to calculate α: usually exploiting sample sizes, variance estimates, design effects etc
- I used:

$$\alpha = \frac{n_1/d_1}{n_1/d_1 + n_2/d_2},$$

where $d_1 = 1 + \frac{Var(w_i, i \in EU-SILC)}{\bar{w}^2_{SILC}}$, $d_2 = 1 + \frac{Var(w_i, i \in HBS)}{\bar{w}^2_{HBS}}$.

- d_1 and d_2 are design effects due to variable (non-constant) weights, and thus n_1/d_1 and n_2/d_2 are effective sample sizes of EU-SILC and HBS.

Cumulating probabilities (1)

- Calculate the probability to be included in the combined sample as a whole, irrespective on what survey the hh actually comes from
- Due to very different response models, EU-SILC is divided into two surveys: EU-SILC new part and EU-SILC repeated part

$$Pr(i \in R) = Pr(i \in R_{HBS}) + Pr(i \in R_{SILC-N}) + Pr(i \in R_{SILC-R})$$

- No intersection term due to negative coordination of samples

Cumulating probabilities (2)

- HBS

$$Pr(i \in R) = Pr(i \in S)Pr(i \in R|i \in S)$$

- EU-SILC new part

$$Pr(i \in R) = Pr(i \in S)Pr(i \in R|i \in S)$$

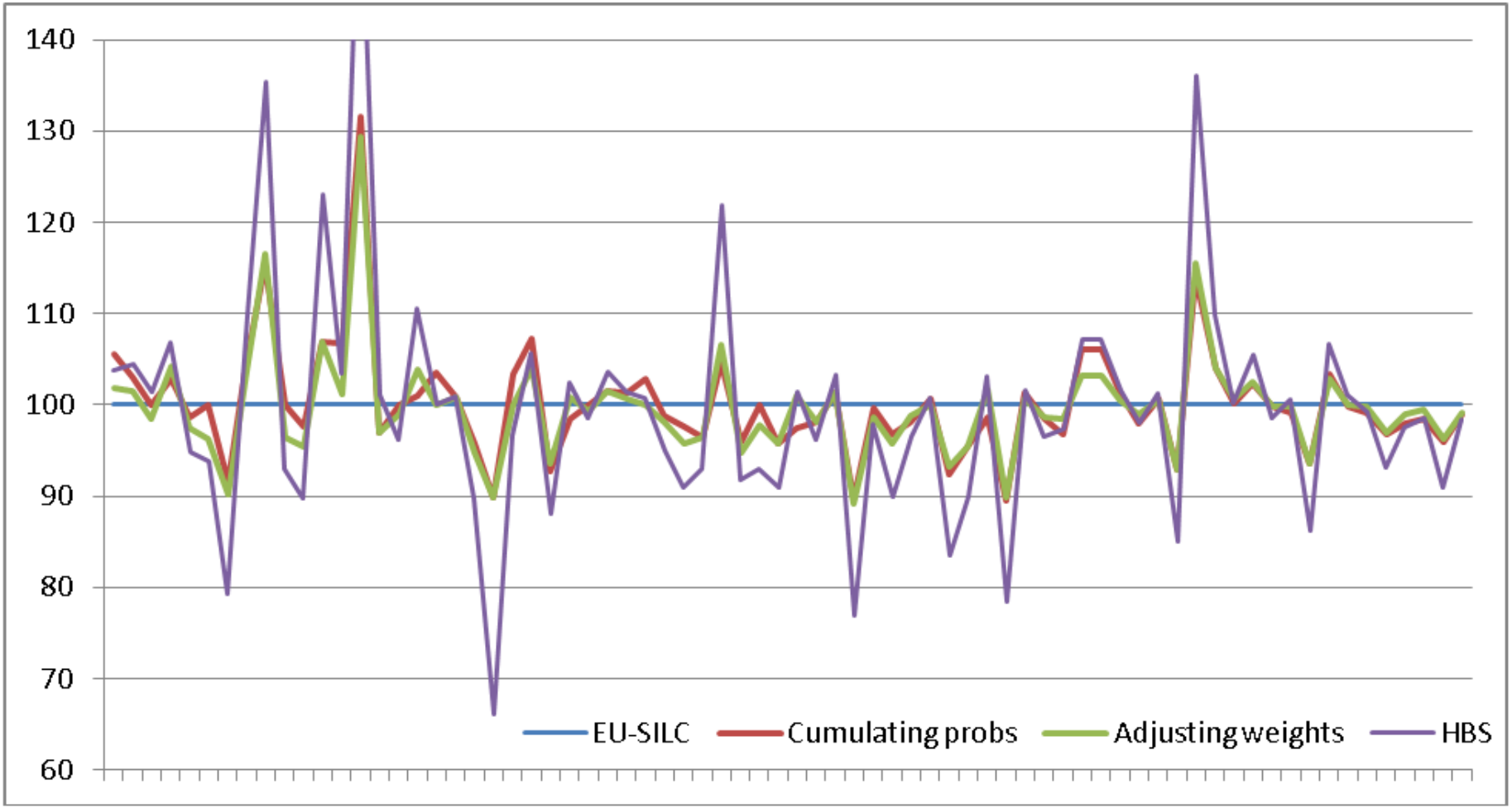
- EU-SILC repeated part, non-response + attrition

$$Pr(i \in R) = Pr(i \in S)Pr(i \in R'|i \in S)Pr(i \in R|i \in R')$$

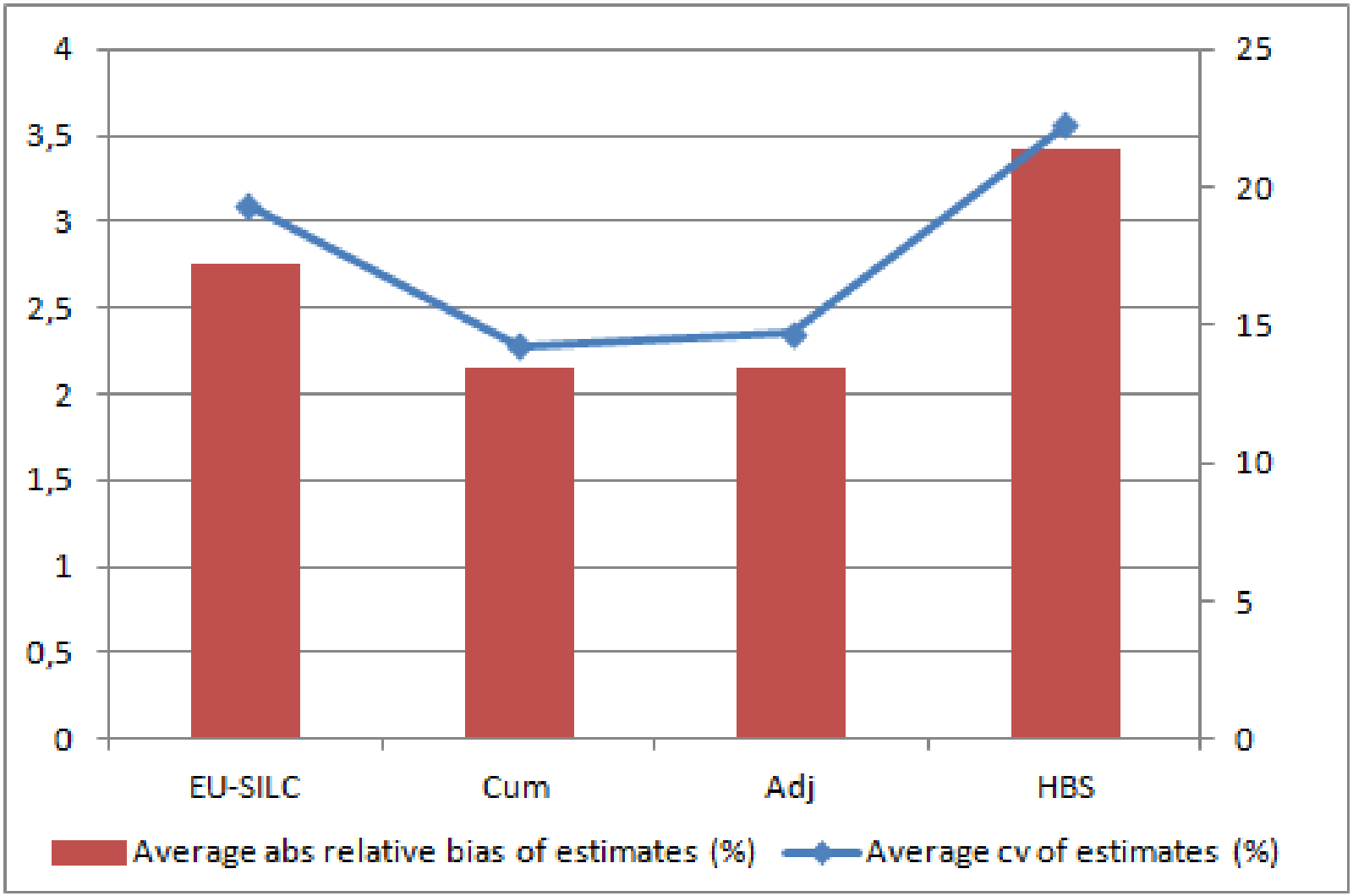
- E.g. for every household in HBS we need to calculated the inclusion probability and response probability *as if* it is included in EU-SILC new part and *as if* it is included in EU-SILC repeated part.
- Response probabilities were modelled by logistic regression with predictors actually used when calculating survey-specific weights.

Comparison of estimates

Estimates relative to corresponding EU-SILC figure



Variance estimates



Conclusions and future plans

- Both methods reduce bias and variance almost equally as compared to survey-specific estimates
- Adjusting weights: simple to calculate, satisfactory performance
- Cumulating probabilities: difficult to calculate, a lot of models involved, harder to explain to user, and just slightly better performance
- At least for combining EU-SILC and HBS, Statistics Estonia will use simpler method
- BUT: These two surveys had identical target populations and similar design! LFS has slightly different target population and different stratification, so we will repeat this analysis for the combination of three surveys: HBS, EU-SILC and LFS.
- We also plan to use estimates from the combined sample (e.g. education) to re-calibrate survey-specific weights and possibly improve their precision.

THANK YOU FOR ATTENTION!