Implementation of SAE to the Dutch Structural Business Survey

Marc Smeets (mset@cbs.nl) and Sabine Krieg (skrg@cbs.nl)

SAE2013, Bangkok, September 1-4, 2013
Introduction

- Research into application of small area estimation (SAE) to business surveys.
- Target variables:
  - continuous and skewly distributed,
  - large differences between enterprises and existence of outliers,
  - variables with many zeroes.
- Model specification:
  - random slope models, transformation of variables, unequal variance structure.
- In collaboration with University of Southampton (Nikos Tzavidis, Hukum Chandra): M-Quantile estimation, ...
Aims of current research

- Consideration of Dutch Structural Business Survey (SBS).
  - Measurement of annual total production and cost-benefit structure of enterprises in the Netherlands.
  - Focus on one sector: the retail trade.
- Getting reliable and consistent estimates
  - for a selection of 9 (related) structural variables,
  - at different publication levels,
  - satisfying preconditions imposed by production process.
- Investigating possibilities and (eventually) implementation of SAE.
Structural target variables

Variables and relations

\[
\text{results} = \text{returns} - \text{costs} \\
\text{returns} = \text{turnover} + \text{other returns} \\
\text{costs} = \text{costs of goods sold} + \text{personnel costs} + \text{depreciation} + \text{other costs}
\]

Abbreviation of variable names

\[
R = T - C \\
T = T_1 + T_2 \\
C = C_1 + C_2 + C_3 + C_4
\]
Publication levels

- Based on Standard Industrial Classification (SIC):
  - classification of enterprises according to economic activity,
  - represented by 5 digit SIC-code.
- Given by 5digit cells, industries, sectors and whole population
  - formed by combinations of SIC-codes,
  - publication levels are nested,
  - totals should add up to totals at higher level.
- Sampling design SBS stratified at the level of industries
  - sample sizes industries are fixed,
  - sample sizes 5digit cells are random and can be 0.
- Retail trade: 71 5digit cells and 27 industries.
Earlier results

Considered situations
- turnover per industry,
- results, returns and costs per 5digit cell.

Considered estimators
- EBLUP (J.N.K. Rao, 2003), SAEtrans (C. Chandra and R. Chambers, 2011)
- M-Quantile estimator (R. Chambers and N. Tzavidis, 2006)
- GREG, Survey Regression (C. Särndal et al, 1992)

Results
- SAE more accurate than GREG and Survey Regression,
- for industries M-Quantile most accurate, for 5digit cells EBLUP,
- SAEtrans most accurate if no strong covariate available (tax turnover).
Preconditions production process

- Totals of industries must be estimated by linear weighting
- turnover is replaced by tax turnover
  - totals of turnover equated with totals of tax turnover,
  - totals of other variables estimated with turnover as covariate and totals of tax turnover as population totals.
**Considered estimator**

- EBLUP based on following model (J.N.K. Rao, 2003):

\[ y_{ij} = x_{ij}^t \beta + z_{ij}^t \vartheta_j + e_{ij}, \text{ where} \]

\[ \vartheta_j \sim \mathcal{N}(0, \Theta), \]

\[ e_{ij} \sim \mathcal{N}(0, k_{ij}^2 \sigma_e^2), \text{ for 5digit cell } j \text{ and enterprise } i. \]

- Specification of \( k_{ij} \)
  - analysis of heteroscedasticity and skewness residuals \( e_{ij} \),
  - stratum standard deviations residuals of estimated regression model.

- Specification of \( x_{ij} \) and \( z_{ij} \)
  - analysis of AIC, point estimates, significance estimates of \( \beta \),
  - tax turnover and size of enterprise used as covariates,
  - random slopes for \( T_2, C_2, C_3 \) and \( C_4 \), otherwise \( z_{ij} = 1. \)
Consistency

- Consistency by Lagrange multiplier with absolute values of point estimates used as weights.
- Three versions of consistent EBLUPs
  1. EBLUPc1: consistent within the 5digit cells, between all variables,
  2. EBLUPc2: consistent between variables and publication levels,
  3. EBLUPc3: consistent between variables, publication levels and equated totals of turnover and tax turnover.
- Simulation based on response data 2006-2010,
  - \( N = 47127, \ n = 3036, \ m = 71, \) 10000 runs.
  - Means sample sizes 5digit cells vary from 0.1 to 436.
Effects of benchmarking

Consistency for Turnover (variable T1)

- EBLUP
- EBLUPc1 (var)
- EBLUPc2 (pub)
- EBLUPc3 (tax)
EBLUP vs Survey regr. (not consistent)
EBLUPc3 vs Survey regr. (consistent)
Conclusions

- SBS estimates 5digit cells can be improved by SAE for most variables, for other variables results are comparable.
- Equating turnover with tax turnover gives good results for turnover, returns, costs, but has not much effect for other variables.
- Benchmarking with direct estimates at industry level leads to instable estimates at level of 5digit cells for variable results.
- Estimates for variables with many zeroes (results, other returns, other costs) could possibly be further improved.