USING AUXILIARY DATA SOURCES IN THE STRUCTURAL SURVEY IN THE SERVICE SECTOR

Thomas Zimmermann (Destatis)

EESW17 - European Establishment Statistics Workshop 2017

Session 1: Sampling Design

Southampton, 30/08/2017

Background on the survey

- Provides relevant information on medium-term developments and structural changes for the service sector

- Sampling fraction $\leq 15\ \%$ of the total number of units in the population

- Stratification by NUTS1 regions, NACE4, and size classes determined by turnover (or # employees)

- Allocate sample sizes such that precise HT estimates for turnover are obtained for stratum groups (NUTS1 x NACE4)

Current Approach

Minimize the maximum weighted coefficient of variation in stratum groups, i.e.

$$F = \max_{g \in G} W_g^q \cdot CV\left(\hat{Y}_g\right) = \max_{g \in G} \frac{W_g^q}{Y_g} \sqrt{\sum_{h \in g} N_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right)}$$

subject to

$$m_h \leq n_h \leq M_h, \forall\, h$$

$$\sum n_h \leq n$$

We use $m_h = 3, M_h = N_h \;\forall\, h$ as well as $W_g = Y_g, q = 0.2.$

Comments

- Heterogeneity leads to highly unequal sampling fractions and large number of take-all strata
- Court decision: Spread response-burden more evenly and take-all strata only acceptable if imperative to quality
- Revision of the sample design is currently studied
- Additional idea: Exploit auxiliary information at the estimation stage

Alternative estimation methods

Requirements

- A single weight should be attached to each unit in the sample
- Good design-based properties
- Coherence with other statistics

→ Calibration estimators

- SAS macro CALMAR from INSEE
- GREG calibration and raking ratio approach considered

Potential sources of auxiliary information

- Sampling frame
- Business register
- Administrative data record

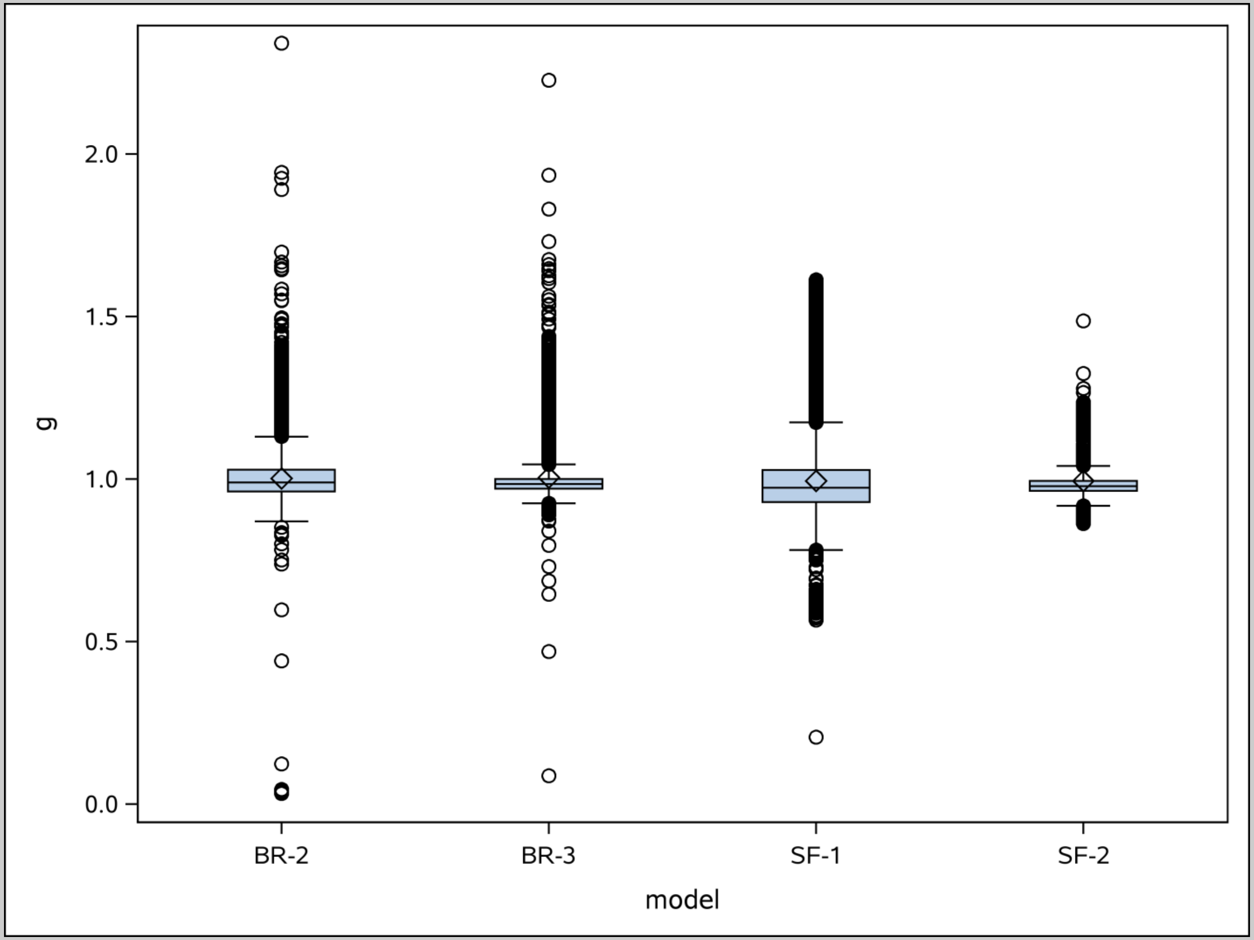
Variables: Turnover, Number of employees, number of enterprises

Account for misclassified units by logistic regression model

Comparison of the models

Model	Calibration constraints	n	R^2_{TUR}	R^2_{EMP}
SF-1	N for NUTS1, NACE2, SC Total TUR for NUTS1, NACE2, SC (all from SF)	153 699	0.951	0.837
SF-2	N for NUTS1, NACE2, SC Total TUR for NUTS1 (all from SF)	153 699	0.864	0.353
BR-1	N for NUTS1, NACE2, SC (from SF) Total TUR for NUTS1, NACE2, SC (from BR)	152 872	0.909	0.844
BR-2	N for NUTS1, NACE2, SC (from SF) Total TUR and EMP for NUTS1 (from BR)	152 342	0.905	0.967
BR-3	N for NUTS1, NACE2 (from SF) Total TUR and EMP for NUTS1 (from BR)	152 342	0.905	0.967

Variation of the g-weights

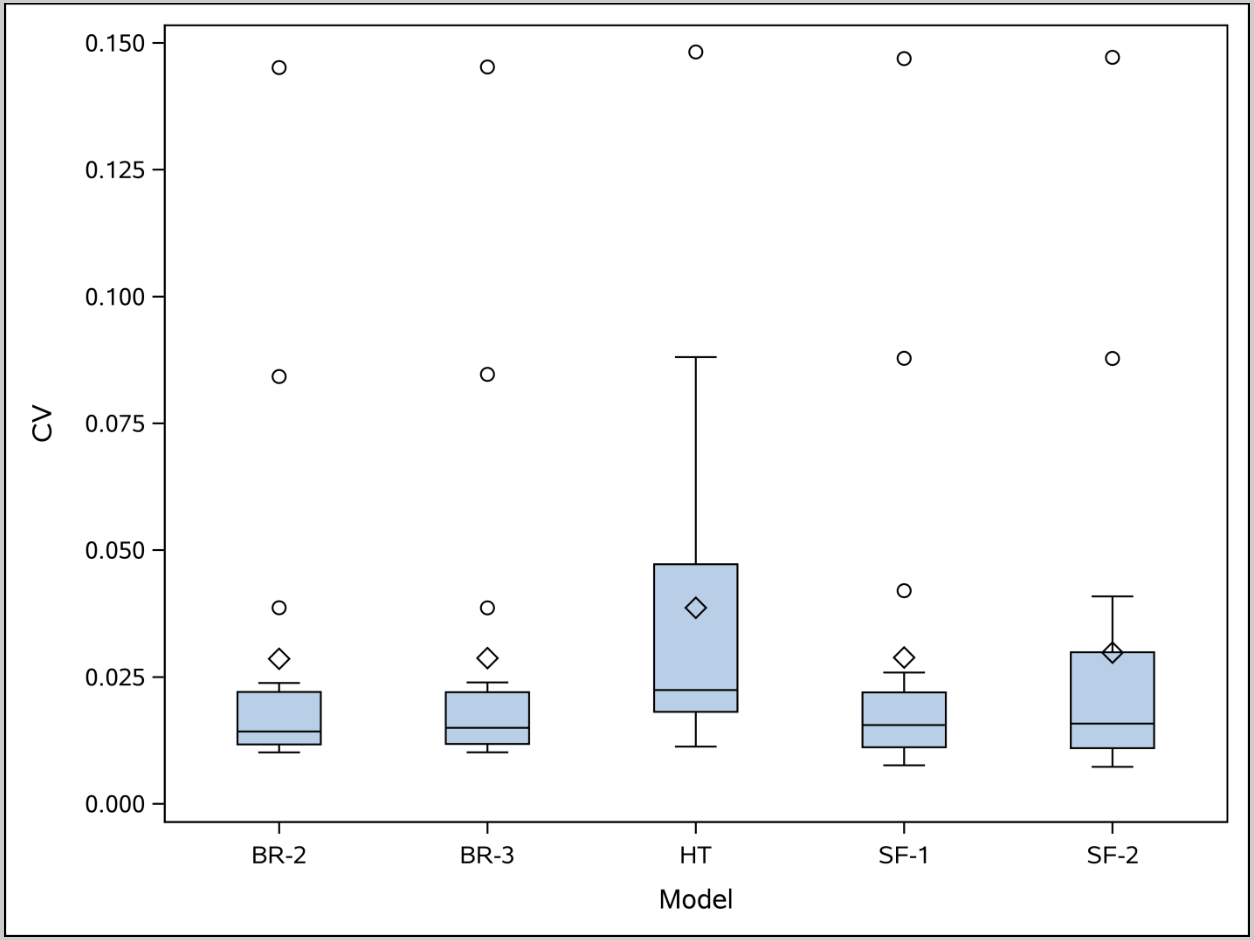


Deviations from HT estimate in %

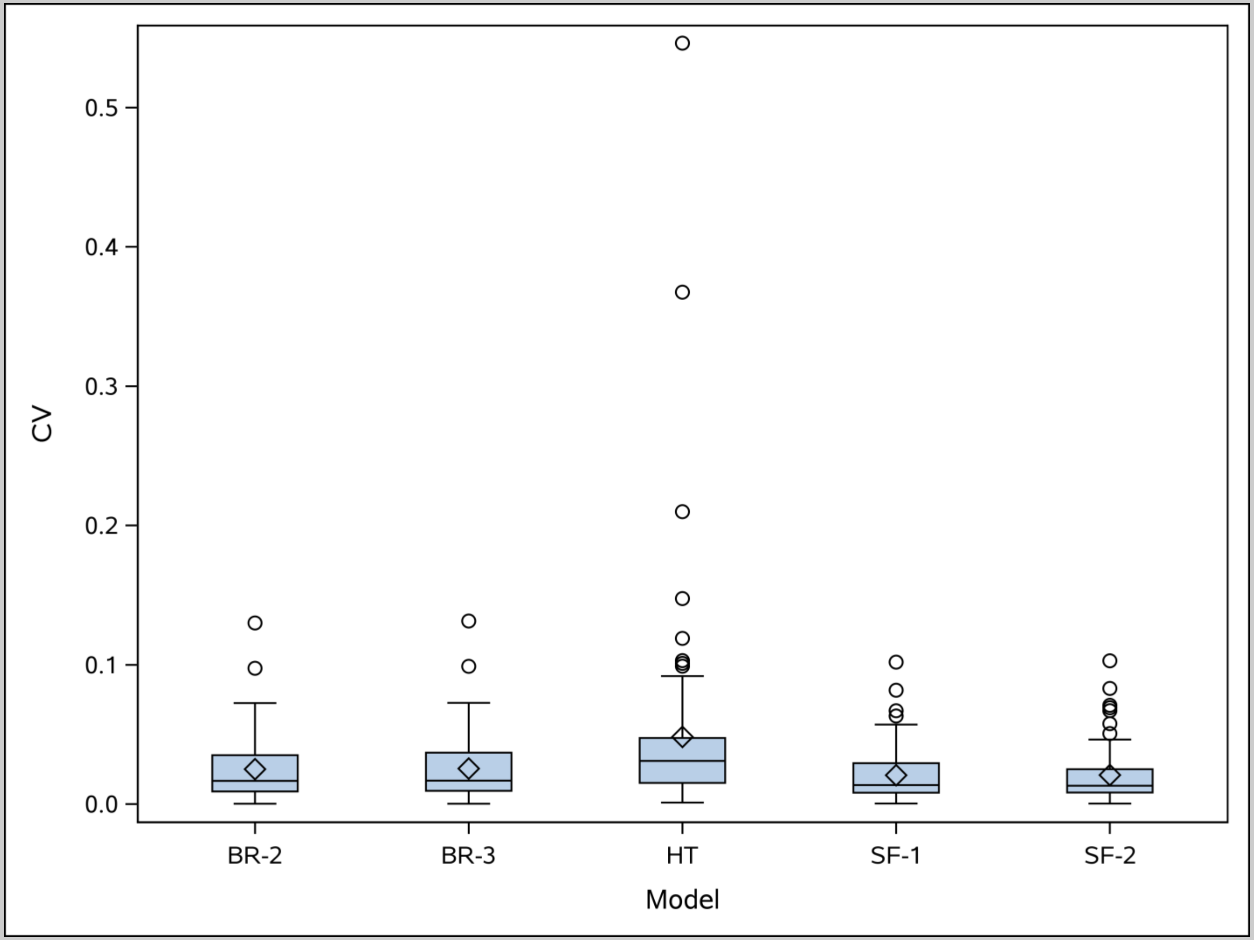
	SF-1	SF-2	BR-2	BR-3
TUR	-0,84	-0,88	0,38	0,60
EMP	-0,36	-1,38	1,95	2,02

- Estimates are comparable at the national level
- Differences up to 7 % for NUTS1-regions

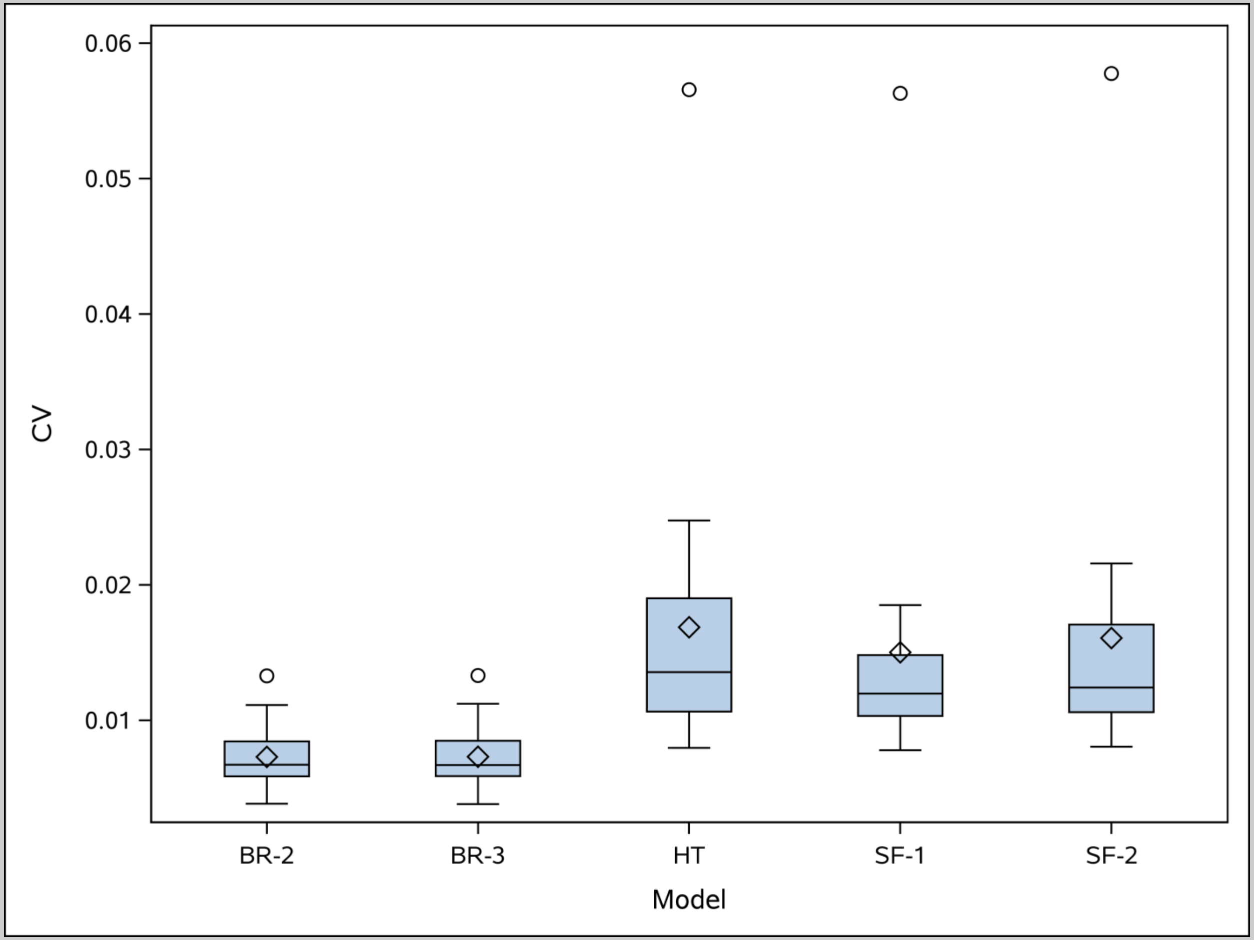
CVs on NUTS1 regions (TUR)



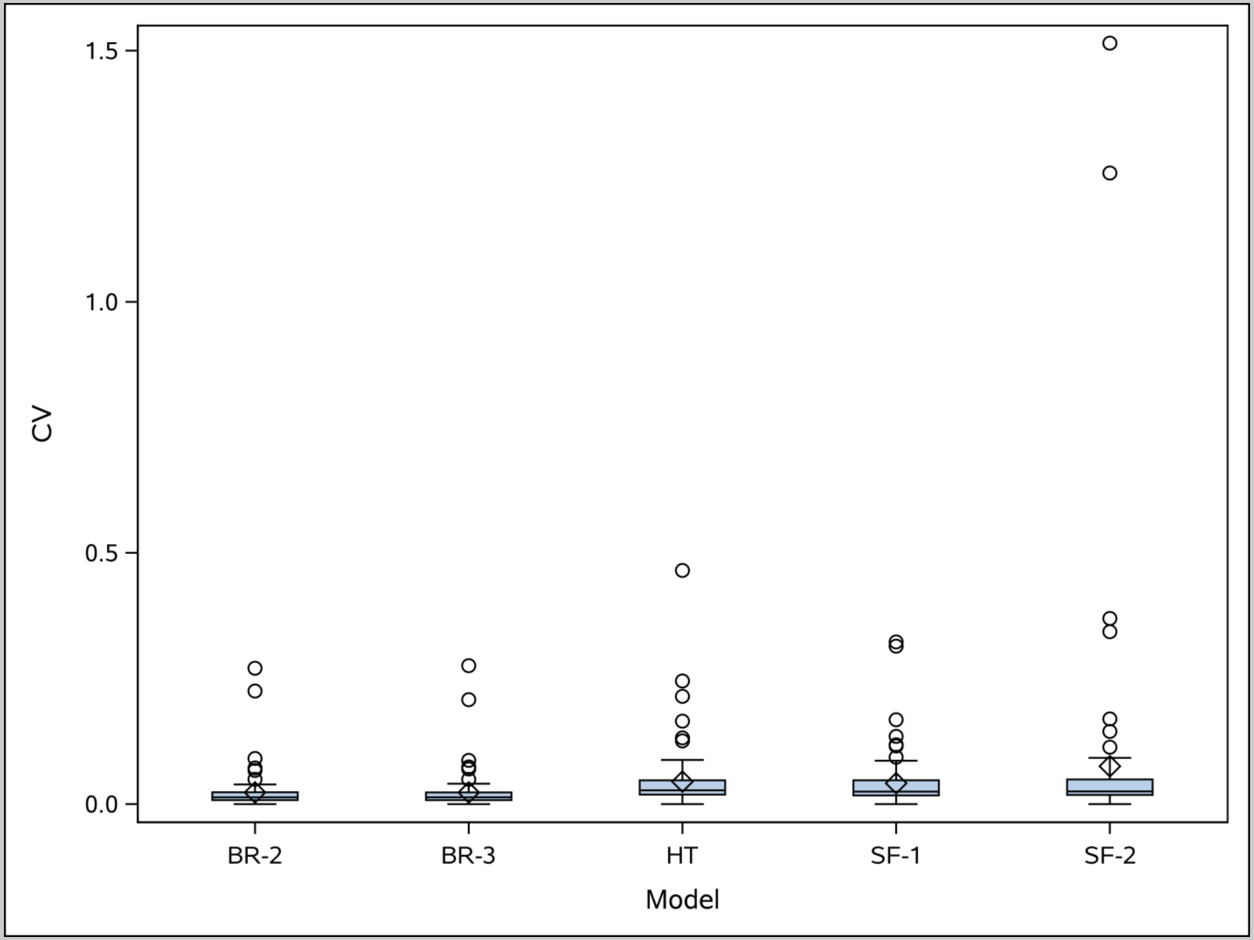
CVs on NACE4 classes (TUR)



CVs on NUTS1 regions (EMP)



CVs on NACE4 classes (EMP)



Questions for further discussion

1. Does your NSI apply some variant of regression / calibration estimation in business surveys?

2. If so, how do you cope with outliers and highly variable data in the covariates?

3. Do you have experience in smoothing your estimates across time or sectors?

THANK YOU FOR YOUR ATTENTION!

Thomas Zimmermann
Telefon: +49/(0) 611 / 75 38 41
thomas.zimmermann@destatis.de
www.destatis.de