The strategy to treat unique enterprises in the Swiss survey on production and value added

Daniel Assoulin
Statistical methods unit (METH)
Swiss Federal Statistical Office

Workshop on Statistical Data Editing, Neuchâtel, Switzerland
18-20 September 2018
The strategy to treat unique enterprises in the Swiss survey on production and value added
Illustration: Swiss survey on production and value added 2010

Distribution of number of employees in a certain activity sector

Restriction on enterprises with empl < 1000

The strategy to treat unique enterprises in the Swiss survey on production and value added | Introduction
Situation

Many variables with positive values and right skewed distributions.

- Units with strong influence on totals of target variables
- High variance in target variables

Sampling design 2010: Census of large enterprises by means of take-all strata based on proxy/auxiliary variable number of employees ($x=empl$).
Number of employees in the 69 enterprises of the take-all strata
Problem of nonresponse in take-all strata:

- experience shows: hard to avoid completely
- jeopardizes the desired variance elimination through exhaustive sample
- existence of single enterprises, for which nonresponse has a strong influence on estimates and make interpretations of evolutions difficult.
Strategy

- Identification of potentially influential enterprises based on register information
- Prevention of nonresponse is the best treatment $\rightarrow$ enforcing response within this group (selective editing, Luzi et al. (2007)).
- Treatment as response homogeneity group with response rate 100%, extrapolation weight 1.
Identification

- \( R_{TA_c} = \) set of \( m_c \) responding enterprises in \( TA_c \) (take-all strata branch \( c \))
- Treatment of \( R_{TA_c} \) as simple random sample of size \( m_c \) out of \( TA_c \)
- \( X_{TA_c} : \) number of employees (\( empl \)) in \( TA_c \)
- Conditional bias (Moreno-Rebollo et al. (1999)) of the Horvitz-Thompson estimator \( \hat{X}_{TA_c} \) given \( i \) has not responded / is not in the sample:

\[
\mathbb{E}(\hat{X}_{TA_c} \mid i \in TA_c \setminus R_{TA_c}) - \mathbb{E}(\hat{X}_{TA_c}) = \bar{X}_{TA_c \setminus i} - x_i
\]  

(1)
Identification (2)

Unique enterprises in the sense of „large“:

\[ U_{c,1} = \left\{ i \in TA_c; \frac{x_i - \bar{X}_{TA_c \setminus i}}{X_c} \geq k \right\} \]  \hspace{1cm} (2)

\( X_c \) = total of branch \( c \), \( k \) was set to 0.025. Alternative: branch specific choice for \( k \).
Identification (3)

- 100% response rate in $U_{c,1}$
- Enterprises in $U_{c,1}$ extrapolated with weight 1: identification has influence on estimation $\rightarrow$ iteration
- Put $TA_c^{(2)} := TA_c \setminus U_{c,1}$ and calculate

$$U_{c,2} = \left\{ i \in TA_c^{(2)}; \frac{x_i - \bar{X}_{TA_c^{(2)} \setminus i}}{X_c} \geq k \right\}$$

- Iteration until $U_{c,l} = \emptyset$, no more enterprises are identified.
- Create the union of all identified enterprises: $U_c = \bigcup_{1}^{l} U_{c,j}$
Illustration (3)

5 enterprises are identified in the considered branch: 3 in $U_1$, 2 in $U_2$
Illustration (4)

**Tabelle:** Number of identified enterprises in the 59 considered branches (2010).

<table>
<thead>
<tr>
<th>Set</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>149245</td>
</tr>
<tr>
<td>Take-All</td>
<td>7891</td>
</tr>
<tr>
<td>$U_1$</td>
<td>148</td>
</tr>
<tr>
<td>$U_2$</td>
<td>36</td>
</tr>
<tr>
<td>$U_3$</td>
<td>4</td>
</tr>
<tr>
<td>$U$</td>
<td>188</td>
</tr>
</tbody>
</table>
Remarks

- In final estimation re-weighting for non-response is followed by calibration (GREG; auxiliary variable $fte=$full-time-equivalent). Ideally, identification would be based on calibration residuals.

- This information is not available, but according experience, variability of residuals increases with $empl$.

- In recent years turnover data from VAT-register was integrated into the Business Register: enhancement of the identification of unique enterprises by means of the ratio $\text{turnover}/fte$ (productivity) which improves anticipation of large residuals.
Influence on quality of the estimates

For branches in which the 100% response rate among identified units is achieved:

- potential for considerable variance reduction of estimates
- the identified enterprises can be treated as self-representing (weight one) without introducing a bias in the estimation. (Assumption: no influence of enforcement on observed values.)
Conclusions

1. Right skewed distribution of target variables, single enterprises with large influence on interesting totals.

2. The presented strategy aims at avoiding sampling and nonresponse errors among unique enterprises and is therefore based on
   - sampling design (take all strata) and
   - enforcement of 100% response among identified enterprises (selective editing).
Conclusions (2)

3. The considered identification by means of the conditional bias of the estimation
   ▶ takes into account the influence of nonresponse of single enterprises on the estimate
   ▶ should be iterated in order to take into account the editing (follow-up) effect on the
     estimate.

4. Successful application reduces variance of estimation without introducing a bias. Efficiency
   depends on strong auxiliary variables.
Literature


