THE STRATEGY TO TREAT UNIQUE ENTERPRISES IN THE SWISS SURVEY ON PRODUCTION AND VALUE ADDED  
Daniel Assoulin, Statistical Methods Unit, Swiss Federal Statistical Office

I. INTRODUCTION

1. The Swiss survey on production and value added is an annual business survey. One goal of the survey is to estimate totals for different economic activity sectors and different variables (e.g., total production). The distributions of most quantitative variables collected are skewed to the right. It is common, that depending on the considered activity sector, some single (large) enterprises may contribute a considerable part to the totals of interest. Due to their size (according to register information) they are usually sampled exhaustively. But if such an enterprise fails to respond, this may have a considerable impact on the estimate, as it is questionable to what extend reweighting of the other enterprises can compensate for this nonresponse. An algorithm will be presented, which based on the conditional bias of the estimation, applied on available proxy variables, identifies units that are then considered as unique. This implies that they are treated with priority in the data collection and editing phase with the goal of achieving a 100% response. Under this condition, estimation treats these enterprises as an exhaustively sampled response homogeneity group with response probability one. The algorithm and the strategy are illustrated for the Swiss survey on production and value added 2010. While some aspects have evolved since then (due to new register data), the general concepts presented in this document are still applied and have proved to be a useful part of the survey process.

II. Situation

A. Population and sample design

2. Most of the important target variables of the survey have positive values and positively skewed distributions. Further characteristics are

- large enterprises with strong influence on the interesting totals
- large variance in target variables.

The sampling design for 2010 is based on the proxy variable number of employees (empl) which is contained in the business register and updated on a yearly basis. For the various illustrations in the document we often concentrate on one specific activity sector. The distribution of the proxy variable empl for this sector is illustrated in Figure (1).

Some details:

- The survey for the reference year 2010 is based on a sample of global size ($n$) of roughly 20,000 units out of a population of about 150,000 enterprises ($N$).
While the sample is drawn by Poisson sampling in the Federal Statistical Office’s coordination system for business surveys, the sampling probabilities have been designed as for the case of a stratified sample.

- Stratification is determined by economic activity sectors (NACE 2) and size categories (empl).
- The sample is designed as a rotating panel with a target rotation rate of 20% per year.
- Large enterprises are sampled with probability one. While there exist various methods for calculating take-all limits (e.g., Hidiroglou, M. A. [1986]) the limit was generally set to 50 employees (in some small branches below), as requested by the production unit. This leads to a rather large take-all strata containing 7891 units.

Detailed descriptions of the population, the sampling design and also the estimation can be found in Assoulin, D. [2013] and Assoulin, D. [2014].

B. Nonresponse in take-all strata

3. The issue of nonresponse in the take-all strata:

- experience shows that it is hard to avoid completely
- jeopardizes the desired effect on the precision (no variance contribution by take-all strata)
- existence of single enterprises, which have a strong influence on estimates. There is no efficient treatment for nonresponse among such enterprises, which is also a difficulty when interpreting observed evolutions.

Figure 2 illustrates that some enterprises may be considered as unique with respect to their size (empl) in the take-all strata of activity sectors (see Section A). These enterprises may also be unique with regard to their response behavior. In order to reduce the risk of bias and variance of the estimation, we aim to prevent nonresponse among such enterprises.
III. Strategy

4. The following approach has been put in place for the treatment of very large enterprises:
   - identification of potentially problematic (unique) enterprises ahead of the data collection phase
   - focus on obtaining complete response within this group
   - treatment as separate response homogeneity group with response and sampling probabilities of one.

The approach of identifying potential highly influential units to ensure their response can be viewed as a special case of selective editing Luzi, O. et al. [2007] or also a simple kind of static adaptive survey design Schouten, B., Calinescu, M. and Luiten, A. [2013] - creating a high priority group for follow-ups based on register information. The elimination of nonresponse among the identified enterprises needs high efforts and the success will depend on the number of concerned units. Hence, we need to look for a selection criteria which captures only a few enterprises that have the potential to strongly influence the results on the activity sector level.

IV. Identification

A. One step selection

5. For sake of simplicity we concentrate on one single activity sector. The considerations analogously hold also for other activity sectors.
   - $R_{TA_c}$ = set of $m_c$ responding enterprises in $TA_c$ (take-all strata activity sector $c$)
• Treatment of $R_{TA_c}$ as simple random sample of size $m_c$ out of $TA_c$. Assuming that the units in the take-all strata form a response homogeneity group, this is close to the approach used in estimation, where respondents are treated as simple random samples of fixed size within response homogeneity groups.

• $X_{TA_c}$: number of employees in $TA_c$

Then we consider the conditional bias, see Moreno-Rebollo, J.L., Munoz-Reyes, A., Munoz-Pichardo, J. [1999], of the Horvitz-Thompson estimator $\hat{X}_{TA_c}$ given $i^*$ did not answer:

$$E(\hat{X}_{TA_c}|i^* \in TA_c \setminus R_{TA_c}) - E(\hat{X}_{TA_c}) = \bar{X}_{TA_c \setminus i^*} - x_{i^*}$$

(1)

Calculation:

$$E(\hat{X}_{TA_c}|i^* \in TA_c \setminus R_{TA_c}) = \sum_{i \in TA_c} \frac{N_{TA_c}}{m_c} E(I(i \in R_{TA_c})|i^* \in TA_c \setminus R_{TA_c}) x_i$$

$$= \sum_{i \in TA_c \setminus i^*} \frac{N_{TA_c}}{m_c} E(I(i \in R_{TA_c})|i^* \in TA_c \setminus R_{TA_c}) x_i$$

$$= N_{TA_c} \bar{X}_{TA_c \setminus i^*}$$

(2)

$$X_{TA_c} = x_{i^*} + (N_{TA_c} - 1) \bar{X}_{TA_c \setminus i^*}$$

(3)

The difference between (2) and (3) leads to $\bar{X}_{TA_c \setminus i^*} - x_{i^*}$.

6. Unique enterprises in the sense of large are then identified as

$$U_{c,1} = \left\{ i \in TA_c; \frac{x_i - \bar{X}_{TA_c \setminus i}}{X_c} \geq k \right\}$$

(4)

where $X_c = \text{total of activity sector } c$. For the considered survey $k$ was set to 0.025. Alternatively different values could be used depending on the considered activity sector (and its size with respect to the number of enterprises).

7. Hence, in order to be selected, an enterprise needs on one hand to be large in its activity sector and on the other hand it needs to be unique in the sense, that it is questionable whether re-weighting of the other enterprises in the take-all strata can compensate it in case of nonresponse.

8. Remarks

• As for the final weighting nonresponse in the take-all strata is not only treated by simple re-weighting but also by calibration (GREG; auxiliary variable $fte=\text{full-time-equivalent}$), it would be appropriate to use this estimator, applied to the most important target variables, in the selection criteria. As can be seen in Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. [2013] this would be done by replacing the variable $empl$ with the corresponding calibration residuals for all units, an information that is not available. Nevertheless, if we assume that the variance of these residuals increases with our proxy variable $empl$, our approach tends to choose units with potentially large calibration residuals. (In Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. [2013] the conditional bias is considered with regard to selection / non-selection for the sample, but as in our case we consider the respondents in the take-all strata as a random sample, the situation is identical.)

• In recent years turnover data from VAT-register was integrated into the Business Register. While this information was so far not used in calibration / general sampling design, it was
used to enhance the identification of unique enterprises by means of the ratio \( \frac{\text{turnover}}{\text{fte}} \) (productivity) which improves anticipation of large residuals.

- Beside the conditional bias of the estimator „given the enterprise does not respond” it is also possible to consider the conditional bias of the estimator „given the enterprise did respond”. The value of this bias is different from (1), as can be seen in Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. [2013], where it is used to construct a robust estimator. But it is worth to note that while robustification treats influential respondents, here we are rather concerned about influential non-respondents, so (1) seems to be the appropriate measure for our purposes.

B. Iteration

9. For enterprises identified as unique by the selection criteria described above, response will be enforced. This means that the response probability for such entities becomes one. Hence, for the estimation stage it is appropriate to use a weight of one for these enterprises. This simple situation may also illustrate the importance of taking into account the effects of adaptive designs / selective editing on the response probabilities used for estimation. For the one step selection we assumed that all respondents would be re-weighted. As the special treatment of unique enterprises contradicts this condition, it seems reasonable to adapt the selection to the new situation. Therefore, we exclude the enterprises detected in the one step selection from the units in the take-all strata \( TA_{c1} \), which leads to a smaller set of enterprises that we call \( TA_{c2} \). Then we apply (4) on \( TA_{c2} \), which gives us a second set of important units, called \( U_{c2} \). In the same way we iterate the selection procedure until after \( l \) iterations we have \( U_{c,l} = \emptyset \) and the final set of important units is given by

\[
U_c := \bigcup_{j=1}^{l} U_{c,j}.
\]  

10. As the variable \( x \) is always positive, we know that \( \bar{X}_{TA_{c1}} \geq 0 \) and therefore we have

\[
U_c \subset \left\{ i \in TA_c ; \frac{x_i^*}{\bar{X}_c} \geq 2.5\% \right\}.
\]

This means that the potential candidates for being selected are limited to enterprises with a portion of at least 2.5% of the total number of employees in the activity sector and so the algorithm will usually stop after a moderate number of iterations.

11. In the considered activity sector s the algorithm selects 5 enterprises : 3 in \( U_1 \), 2 in \( U_2 \). The selection is illustrated in Figure 3.

12. Table 1 shows the total number of selected enterprises in the different iteration steps for the survey on production and value added.

<table>
<thead>
<tr>
<th>Table 1. Number of identified enterprises among 59 considered activity sectors (2010).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set</td>
</tr>
<tr>
<td>Population</td>
</tr>
<tr>
<td>Take-All</td>
</tr>
<tr>
<td>( U_1 )</td>
</tr>
<tr>
<td>( U_2 )</td>
</tr>
<tr>
<td>( U_3 )</td>
</tr>
<tr>
<td>( U   )</td>
</tr>
</tbody>
</table>
Figure 3. Identification of enterprises in activity sector s.

V. Impact on the quality of the estimation

A. Bias and variance

13. For economic sectors for which a 100% response rate for identified units can be reached we have the following situation:
   
   - the weighting with one does not introduce bias into the estimation
   - reduction of variance for the estimation.

B. Illustration

14. Assumption: Without special efforts the response rate in $TA_c$ is $\alpha < 1$ and the respondents are considered as a simple random sample in $TA_c$. Let further denote
   
   - $\text{Var}_1(\hat{X}_{TA_c}) = \text{Variance of the Horvitz-Thompson estimator in } TA_c$
   - $\text{Var}_2(\hat{X}_{TA_c})$: analog, but treatment of the $N_{U_c}$ enterprises in $U_c$ as separate strata / response homogeneity group with response rate 100%.

   \[
   \text{Var}_1(\hat{X}_{TA_c}) = \frac{N_{TA_c}^2}{\alpha N_{TA_c}} (1 - \frac{\alpha N_{TA_c}}{N_{TA_c}}) S^2_{x,TA_c} \tag{7}
   \]

   \[
   \text{Var}_2(\hat{X}_{TA_c}) = \frac{(N_{TA_c} - N_{U_c})^2}{\alpha (N_{TA_c} - N_{U_c})} (1 - \frac{\alpha (N_{TA_c} - N_{U_c})}{N_{TA_c} - N_{U_c}}) S^2_{x,TA_c \setminus U_c} \tag{8}
   \]
\( S^2_{x,TA_c}; \) Variance of \( x \) in \( TA_c \)

Relative reduction of the variance in \( TA_c \):

\[
1 - \frac{\text{Var}_2(\hat{X}_{TA_c})}{\text{Var}_1(\hat{X}_{TA_c})}
\]

with

\[
\frac{\text{Var}_2(\hat{X}_{TA_c})}{\text{Var}_1(\hat{X}_{TA_c})} = \frac{(N_{TA_c} - N_{U_c}) S^2_{x,TA_c \setminus U_c}}{N_{TA_c} S^2_{x,TA_c}}.
\]

From (9) follows that the relative variance reduction does not depend on the response rate \( \alpha \). The impact on the variance for estimation of empl in relationship to the portion of identified units is illustrated in Figure 4.

![Figure 4. Illustration: Impact on variance.](image-url)

VI. Conclusions

15. There exist single units with large influence on interesting totals.

16. 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).

17. The considered identification by means of conditional bias of the estimation

- takes into account the influence of nonresponse of single enterprises on the estimate
- can be adapted to different estimators
- should be iterated in order to take into account the editing (follow-up) effect on the estimate.
18. The presented strategy aims to reduce the variance of the estimation without introducing a bias. Its efficiency depends on strong auxiliary information that is ideally available for all units in the population.

References


