European Establishment Statistics Workshop
Efficient methodology for producing high quality establishment statistics

(Stockholm, Sweden, 6-9 September 2009)

CHOOSING AN IMPUTATION METHOD FOR LARGE FIRMS

Prepared by Daniel Assoulin (Daniel.Assoulin@bfs.admin.ch), Swiss Federal Statistical Office

I. INTRODUCTION

The national survey on production and value added is an annual business survey. One goal of the survey is to estimate total production values for different economic activity sectors. Experience shows that total production for an activity sector is often strongly influenced by a few large firms that are sampled with probability one. Despite efforts to eliminate non-response within the exhaustive stratum of large firms, non-response still occurs. Besides of strongly affecting precision of estimations, non-response of large firms leads to interpretation problems as subject matter specialists examine these firms' values very closely when commenting on evolutions of estimated totals. The paper explains why it was considered to treat unit non-response among large firms with imputations. Then it discusses the importance of estimation accuracy (comprising bias and variance) and predictive accuracy as criteria when choosing an imputation method for the situation, where imputed values should be appropriate for estimating totals and interpreting results. These imputation properties lead to a choice of a few imputation quality indicators. The calculation of these indicators requires true values along with the corresponding imputations. As this information was not available in practice, the indicators are used in combination with a simulation study to establish a quality ranking among four imputation methods. Results reflect the impact of good auxiliary information on estimation and predictive accuracy and may give a justification for gathering such information at an additional cost.

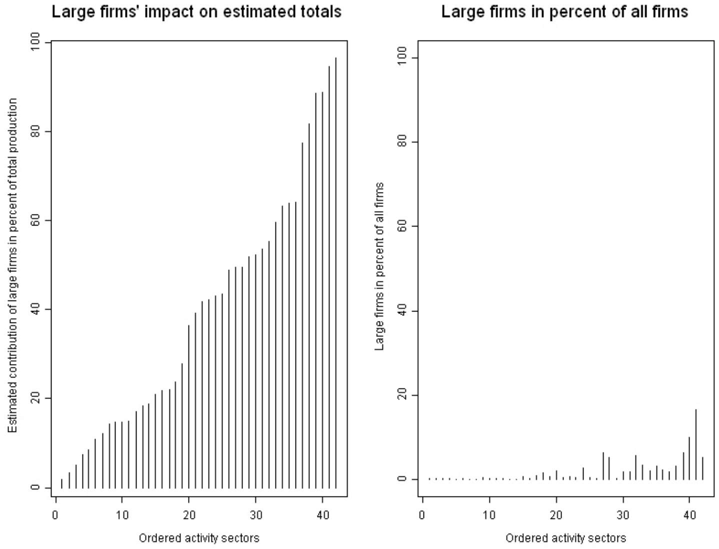
II. Situation

The survey is based on a stratified sample of global size $n = 11'533$ out of a population of $N = 151'514$ firms. Some characteristics of the sampling design are as follows:

- Stratification is determined by economic activity sectors (NACE 2) and size categories, where size is measured by the number of full time employees.
- Depending on the activity sector different limits of firm size have been established using the method described in (Hidiroglou 1986) in order to define exhaustive strata.
- So far the sample is renewed every 3-5 years and for estimation purposes population is assumed to stay stable over this time. In future it is planed to renew the sample every year according to a rotation scheme.

It can be observed that the larger firms distinguish themselves from other firms within the exhaustive stratum by better response rates and by higher but also more variable target variables. Hence, in order to reduce bias and variance of the estimation, extrapolation considers these (very) large firms within every activity sector as a separate substrata within the take-all strata. In this work, large firms are considered as firms belonging to these substrata, which the current extrapolation framework defines using a boundary of 300 full-time equivalents (FTE). A main goal of the survey

FIGURE 1. Portion of large firms ($FTE \geq 300$) vs. impact on estimated totals by activity sectors.



is the estimation of total production for different activity sectors. The focus within this work is on this target variable. In most activity sectors large firms are small sub populations. However, in many cases they have a large impact on the estimation of total production. This fact is illustrated in Figure 1, which charts the portion of large firms for selected activity sectors together with their relative impact on estimated totals of production for year 2004. For sake of simplicity further discussions concentrate on one single activity sector. The considerations analogously hold also for other activity sectors. Furthermore, we will restrict our view to imputations for large firms in that activity sector. Hence, we will deal with one (exhaustive) stratum of large firms and in general the notation given in Table 1 will be used.

TABLE 1. Notation.

Term	**Explication**
L	Stratum of large firms
N_L	Stratum size of large firms
$\widehat{y}_i(t)$	Value of firm i, year t, after imputation
$y_i^*(t)$	True production value of firm i, year t
NA	Denotes missing values
$m_L(t)$	Number of firms with $y_i^*(t) \neq NA$
$R_L(t)$	Large firms with $y_i^*(t) \neq NA$
$R_L^c(t)$	Large firms with $y_i^*(t) = NA$

In the statistical data preparation process (E&I - process) large firms go through interactive treatment, during which missing and erroneous production values are treated by case-to-case imputations based on firm specific information (tax data, recall, business reports). In cases of unit non-response the necessary information for the firm specific treatment is not always available. Therefore, missing production values due to unit non-response remain a problem after interactive treatment. The following assumptions enable the comparison of imputation methods based on the simulation study presented later:

- Data after interactive treatment correspond to true values ($y_i^*(t)$). Hence, after interactive treatment we are in the situation, where a production value is either true or missing. In further considerations non-response of a large firm is identified with $y_i^*(t) = NA$.
- Uniform response probabilities within the stratum of large firms belonging to the same activity sector.

In the case of full response among large firms they are extrapolated with weight one. For all other firms the estimation of total production is based on a compound robustified ratio estimator using auxiliary variable full time equivalent (of employment). The estimator is denoted by $\widehat{Y}_{L^c}$ and the robustification is based on the one-step ratio estimator presented in (Hulliger 1999). The estimation of total production during year t for a certain activity sector can then be written as

$$\widehat{Y}(t) = \widehat{Y}_{L^c}(t) + \sum_L y_i^*(t). \tag{1}$$

The survey on value added and production is a structural business survey. Nevertheless, for macro-editing purposes subject matter specialists are highly interested in evolutions of estimations with respect to the previous year. Based on (1) the estimation of the evolution between two consecutive years t and $t-1$ without renewal of the sample is:

$$\widehat{Y}(t) - \widehat{Y}(t-1) = \widehat{Y}_{L^c}(t) - \widehat{Y}_{L^c}(t-1) + \sum_L (y_i^*(t) - y_i^*(t-1)). \tag{2}$$

Relaxing the assumption of complete response among large firms one has to deal with situations, where large firms respond only in one or even in none of the considered years. Imputing missing values and using (2) after imputation leads to

$$\widehat{Y}(t) - \widehat{Y}(t-1) = \widehat{Y}_{L^c}(t) - \widehat{Y}_{L^c}(t-1) + \sum_L (\widehat{y}_i(t) - \widehat{y}_i(t-1)). \tag{3}$$

The variable $\widehat{y}_i(t)$ is defined as

$$\widehat{y}_i(t) = \begin{cases} y_i^*(t) & \text{if } y_i^*(t) \neq NA \\ y_{I,i}(t) & \text{if } y_i^*(t) = NA, \end{cases}$$

where $y_{I,i}$ is calculated with a statistical imputation method. As they distinguish evolutions that can be explained by economic reasons from evolutions that are due to changes in single firms subject matter specialists examine large firms' values very closely when commenting the results. The special interest in large firms is due to the fact that they are considered as unique with a high impact on estimated totals. In (3) an impact $\widehat{y}_i(t) - \widehat{y}_i(t-1)$ on the estimation is attributed to each firm i even if it did not respond for both of the considered years. In the current approach large units with $y_i^*(t) = NA$ are treated by re-weighting. Experience shows that the lack of an explicit value for large non-respondents leads to interpretation problems when subject matter specialists are discussing the results and may end up with erroneous comments in terms of non-response. Therefore, we consider the treatment of missing production values by an imputation method which explicitly attributes to each large firm a value that can be interpreted as its contribution to the estimation.

III. CONSIDERED IMPUTATION PROPERTIES AND APPROPRIATE INDICATORS

Due to the large impact of large firms on the estimated total, the absolute value of the error resulting from imputation should be small. Hence, the imputation method must have good estimation accuracy. As the imputed values are used to estimate the contributions of non responding large firms to $\widehat{Y}(t)$, the imputation method must be proven to have also good predictive accuracy.

Therefore, estimation and predictive accuracy have been judged as most important in our case and the paper is focused on these properties. An overview of desirable imputation properties given in (Chambers 2006) contains predictive accuracy, ranking accuracy, distributional accuracy, estimation accuracy and imputation plausibility with the descriptions of predictive and estimation accuracy given below.

- Predictive accuracy: The imputation procedure should maximize preservation of true values. That is, it should result in imputed values that are as "close" as possible to the true values.
- Estimation accuracy: The imputation should reproduce the lower order moments of the distribution of true values. In particular it should lead to unbiased and efficient inferences for parameters of the distribution of true values (given that these true values are unavailable).

As we focus on the estimation of totals, we use the term estimation accuracy in a restricted sense and limit the comparisons to first order moments of the distributions of true and imputed values. Dividing the imputation error

$$\sum_L \widehat{y}_i(t) - \sum_L y_i^*(t) \tag{4}$$

by $\sum y_i^*(t)$ leads to the weighted relative average imputation error (weight 1) as displayed in (Luzi, O. et al. 2007). Taking the weighted relative average imputation error's absolute value gives

$$I_1 = |\frac{1}{\sum_L y_i^*(t)}(\sum_L \widehat{y}_i(t) - \sum_L y_i^*(t))| \tag{5}$$

Due to its appealing interpretation (absolute value of imputation error relative to the estimated total) we choose that indicator for comparing estimation accuracy among different imputation methods. The weighted L1-distance between true and imputed values leads to (weight=1)

$$I_2 = \frac{1}{N_L} \sum_L |\widehat{y}_i(t) - y_i^*(t)| \tag{6}$$

This indicator is used to measure predictive accuracy of imputations and has been chosen among other indicators presented in (Luzi, O. et al. 2007) and (EUREDIT Project 2004) for its easy interpretation.

IV. CONSIDERED IMPUTATION METHODS

Four imputation methods have been evaluated for $k \in R_L^c(t)$

(1) mean: imputation by the mean of respondents:

$$y_{I,k}(t) = \frac{1}{m_L(t)} \sum_{i \in R_L(t)} y_i^*(t) \tag{7}$$

(2) nnbr1: imputation by the nearest neighbor d determined by the L1-distance between vectors x containing 7 standardized (mean = 0 and standard deviation = 1) auxiliary variables (the three variables production, intermediate consumption and personnel costs from the two previous surveys and variable fulltime equivalent according to the sampling frame). Hence

$$d = \text{argmin}_{i \in \cap_{t-2}^{t} R_L(t)} Dist_{L_1}(\boldsymbol{x}_i, \boldsymbol{x}_k) \tag{8}$$

where $Dist_{L_1}$ refers to the L_1-distance, calculated on dimensions with non missing values.

(3) nnbr2: imputation by the nearest neighbor d determined by the L1-distance between auxiliary variable full time equivalent.

TABLE 2. Simulation results.

meth	$\bar{I}_1$	$\bar{I}_2$	$\bar{R}_{I_1}$	$\bar{R}_{I_2}$
mean	5.92%	13’433	3.20	3.72
ratio	1.18%	2’415	1.46	1.02
nnbr1	2.76%	5’959	2.24	2.06
nnbr2	5.63%	12’532	3.10	3.20

(4) ratio: ratio imputation based on the auxiliary variable

$$x_i = \begin{cases} y_i^*(t-1) & \text{if } y_i^*(t-1) \neq NA \\ y_{C,i}(t-1) & \text{if } y_i^*(t-1) = NA \end{cases}$$

where $y_{C,i}(t-1)$ denotes a value for year $(t-1)$ based for example on updated tax information that was not available during the survey for year $t-1$. This leads to

$$y_{I,k}(t) = \frac{\sum_{R_L(t)} y_i^*(t)}{\sum_{R_L(t)} x_i} x_k, \; k \in R_L^c(t). \tag{9}$$

Note: Auxiliary variable x_i is constructed for all $i \in L$ to be used in the ratio imputation. In cases with $y_i^*(t-1) = NA$ the construction of x_i implies an additional data collection effort.

V. APPLICATION OF QUALITY INDICATORS

We focus on the quality assessment of the imputation methods in one economic sector. The same considerations apply to the other sectors. The calculation of the considered quality indicators requires true values along with the imputed ones. As this information was not available, the application of the quality indicators had to be based on simulations. In the considered sector we have $N_L = 45$ and $m_L = 39$. Hence, a uniform response rate of 87 % is assumed among large firms. In order to calculate the quality indicators 50 samples of size 5 out of the 39 available values have been selected by simple random sampling. The sampled firms are then imputed according to the four imputation methods. For each sample and each imputation method I_1 and I_2 where calculated. In order to get also a more outlier robust assessment the indicators resulting from the different imputation methods had been ranked in a descending order. This leads to 50 ranks between 1 and 4 for each method. As good quality is reflected in a small indicator, rank 1 points out the best, and rank 4 the worst method. Table 2 displays for each method the mean of the (50) indicator values $\bar{I}_q$ and the mean of the corresponding ranks $\bar{R}_{I_q}, \; q = 1, 2$.

A. Discussion of the simulation results

Ranking the means of indicators I_1 and I_2 leads to 1st ratio, 2nd nnbr1, 3rd nnbr2, 4th mean with regard to both, estimation and predictive accuracy. This ranking is confirmed by the mean ranks $\bar{R}_{I_1}$ and $\bar{R}_{I_2}$. In the case of ratio imputation the mean of the absolute relative imputation error is 1.18%, which is about five time less than for mean imputation, indicating the high impact of the auxiliary information used in ratio imputation. This is in accordance with the high correlation of production values observed between consecutive years (Spearman and Pearson correlations of around 0.95 for 2003/4) and may justify to a certain extend the additional collection effort for auxiliary data related to the considered ratio imputation. The nearest neighbor method which is on the second place also uses the information of previous years. The observation that ratio imputation leads to better results can be explained with the fact that even the nearest neighbor can be „far away" from the non-respondent.

Remark: Interpreting the results based on the simulation study it is important to keep in mind that they evaluate the quality of the imputation methods under the assumption of uniform response probabilities in the considered strata. The violation of this assumption may lead to a bias in the estimation that would not be reflected in this results.

VI. **CONCLUSIONS**

- Large firms may have a large impact on estimated totals and are often considered as unique.
- Subject matter specialists examine large firm's values closely when discussing results and the lack of an explicit value for large non-respondents may lead to erroneous comments in terms of non-response.
- An imputation method with good predictive accuracy leads to explicit values that can be interpreted as estimated contributions of large non-respondents to estimated totals.
- Due to their considerable impact on final results, estimation accuracy is another important quality criteria when choosing an imputation method for large firms.
- Quality assessment of the imputation methods can be based on appropriate quality indicators. Most imputation quality indicators use true values along with imputed values. As true values are often not available, a simulation study can be used in combination with the indicators in order to assess the quality of imputations. In this work, the true response probabilities have not been known and the simulation study was based on the assumption of uniform response probabilities within the considered stratum. Therefore, results must be interpreted with care.
- The quality assessment designed a ratio imputation, based on the previous year's production value, as the preferable choice among the four considered imputation methods. The auxiliary information used in ratio imputation would require an additional data collection effort for large firms that could be justified to a certain extend with the good performance of this method in the quality assessment.

References

Chambers, R. (2006). Evaluation Criteria for Editing and Imputation in Euredit. *Statistical Data Editing, Vol. 3, United Nations Publications*, 17–28.

EUREDIT Project (2004). *Towards Effective Statistical Editing and Imputation Strategies - Findings of the Euredit project*, Volume 1. http://www.cs.york.ac.uk/euredit/results/results.html.

Hidiroglou, M. A. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician 40*, 27–31.

Hulliger, B. (1999). Simple and robust estimators for sampling. In *Proceedings of the Section on Survey Research Methods*, pp. 54–63. American Statistical Association.

Luzi, O. et al. (2007, August). *EDIMBUS-RPM.*
http://edimbus.istat.it/EDIMBUS1/document/RPM_EDIMBUS/RPM_EDIMBUS.pdf.