WP. 37
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (vii): Indicators for measuring the quality impact of data editing and imputation

USING QUALITY INDICATORS TO CHOOSE AN IMPUTATION METHOD FOR LARGE FIRMS

Supporting paper

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

I. INTRODUCTION

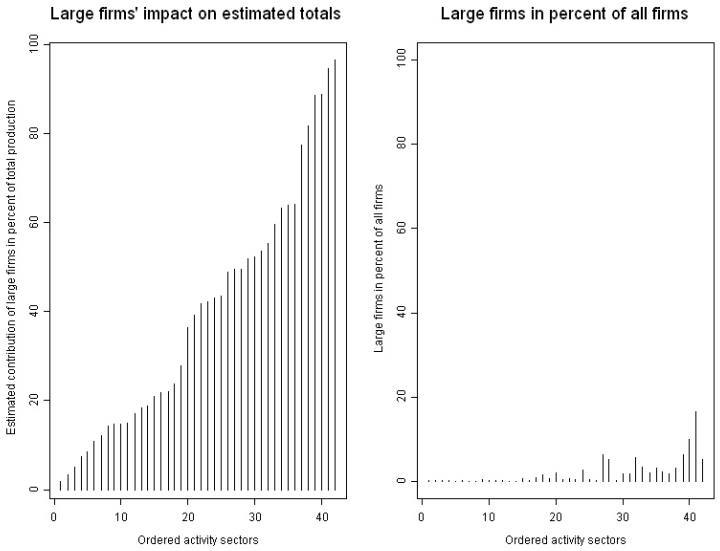
1. The national survey on production and value added is an annual business survey. One goal of the survey is to estimate total production values for different economic activity sectors. Experience shows that total production for an activity sector is often strongly influenced by a few large firms that 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.

II. SITUATION

2. 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.

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



- So far the sample is renewed every 3-5 years and for estimation purposes population size is assumed to stay stable over this time. In future it is planed to renew the sample every year according to a rotation scheme.

3. It can be observed that large 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 treats them within every activity sector as a separate exhaustive substrata. The current extrapolation framework uses a boundary of 300 full-time equivalents (FTE) in order to establish substrata of large firms.

4. A main goal of the survey is the estimation of total production for different activity sectors. The focus within this work is on this target variable.

5. 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.

6. 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.

7. In the statistical data preparation process (E&I - process) large firms go through interactive treatment, during which missing and erroneous values are treated by imputation if the necessary information for this case to case treatment is available. For erroneous data and item non-response, interactive treatment always leads to imputations.

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 $y_i^*(t) = NA$

8. 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.

9. 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:

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

10. 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.

11. 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. This is due to large firm's a priori known high impact on

the estimations and the availability of firm specific data. 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 general re-weighting for non-response will not provide such an information.

12. 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 it was decided to investigate the possibility to treat non-response of large firms by imputation which explicitly attributes to each large firm a value that can be interpreted as its contribution to the estimation.

III. CONSIDERED IMPUTATION PROPERTIES

13. Due to the large impact of

$$\sum_L \widehat{y}_i(t)$$

on $\widehat{Y}(t)$, the absolute value of the error resulting from imputation

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

should be small. Hence, the imputation method must have good estimation accuracy.

14. 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.

15. Therefore, estimation and predictive accuracy have been judged as most important in our case and the paper is focused on these properties.

16. 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).

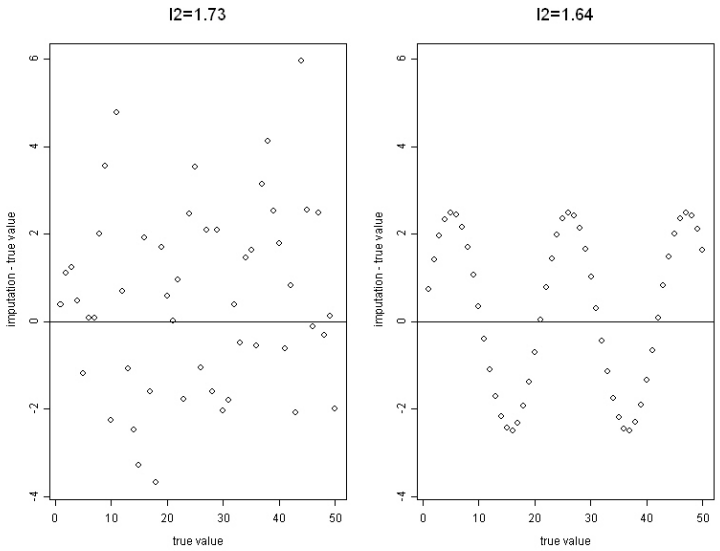
As we focus on the estimation of totals, we use the term estimation accuracy in a restricted sense and limit the comparisons to first order moments of the distributions of true and imputed values.

IV. CHOICE OF INDICATORS

17. After statistical imputation the estimator of the total for population L is

$$\widehat{Y}_L(t) = \sum_L \widehat{y}_i(t). \tag{4}$$

FIGURE 2. An example illustrating the additional value of graphical analysis when assessing predictive accuracy for imputations.



Hence, the estimation error is only due to the imputation error:

$$\sum_L \widehat{y}_i(t) - \sum_L y_i^*(t) \tag{5}$$

Dividing the expression 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's imputation errors absolute value gives

$$I_1 = |\frac{1}{\sum_L y_i^*(t)}(\sum_L \widehat{y}_i(t) - \sum_L y_i^*(t))| \tag{6}$$

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.

18. The weighted L1-distance between true and imputed values leads to (weight=1)

$$I_2 = \frac{1}{N_L}\sum_L |\widehat{y}_i(t) - y_i^*(t)| \tag{7}$$

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.

19. Indicators I_1 and I_2 enable to establish a ranking between imputation methods according to predictive and estimation accuracy. I_1 is directly derived from (5) and measures estimation accuracy with regard to the estimation of the total. There exist different variants of the indicator but they would not give additional value to our analysis. The case of predictive accuracy is not so clear. For example I_2 will not detect structural patterns in differences between true and imputed values. This is illustrated in Figure 2, where the two displayed imputation methods lead to similar values in I_2, but do obviously not have equal predictive properties. One may try to assess predictive accuracy by a carefully chosen set of indicators able to detect all behavior of $\widehat{y}_i(t) - y_i^*(t)$ that deviates from desired predictive properties (e.g. expectation unequal zero, structural patterns,

large variance). In this work a simpler approach was chosen: The evaluation of predictive accuracy is based on indicator I_2 but underpinned with a graphical comparison between imputed and true values.

V. CONSIDERED IMPUTATION METHODS

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

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

$$y_{I,k}(t) = \frac{1}{m_L(t)} \sum_{i \in R_L(t)} y_i^*(t) \tag{8}$$

(2) nnbr1: imputation by the nearest neighbor d determined by the L1-distance between vectors $\mathbf{x}$ 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

$$d = \text{argmin}_{i \in \cap_{t-2}^{t} R_L(t)} Dist_{L_1}(\boldsymbol{x}_i, \boldsymbol{x}_k) \tag{9}$$

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.

(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

$$y_{I,k}(t) = \frac{\sum_{R_L(t)} y_i^*(t)}{\sum_{R_L(t)} x_i} x_k, \; k \in R_L^c(t). \tag{10}$$

Note that x_i is constructed for all $i \in L$ to be used in the ratio imputation.

VI. APPLICATION OF QUALITY INDICATORS

21. We focus on the quality assessment of the indicators in one economic sector. The same considerations apply to the other sectors.

22. 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.

23. 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

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

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 $\bar{R}_{I_q},\ q = 1, 2$.

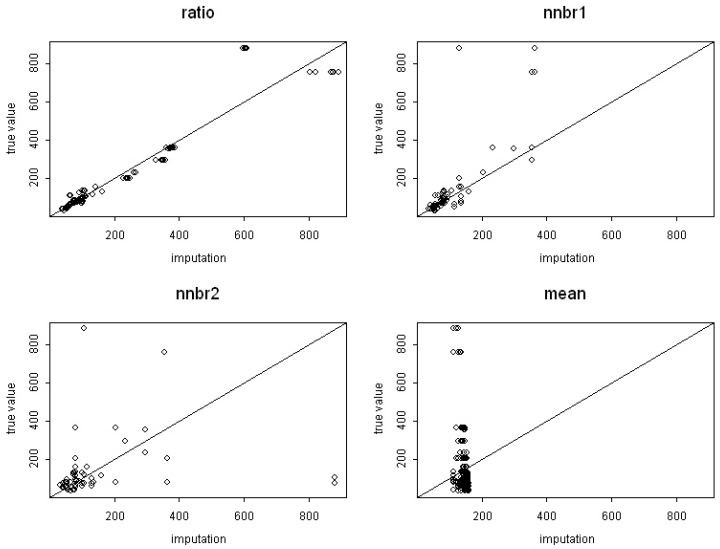
A. **Discussion of the simulation results**

24. Ranking the means of indicator I_1 leads to 1st ratio, 2nd nnbr1, 3rd nnbr2, 4th mean with regard to estimation accuracy. This ranking is confirmed by the mean ranks of the different imputation methods with respect to indicator I_1. 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 around 0.95 for 2003/4). 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. As the nearest neighbor only based on full time equivalent is on the third and mean imputation on the forth place, the ranking reflects the different amplitudes of auxiliary information included in the four imputation methods.

25. $\bar{I}_2$ and $\bar{R}_{I_2}$ had been calculated to assess predictive accuracy. Both criterias lead to the quality ranking 1st ratio, 2nd nnbr1, 3rd nnbr2, 4th mean. As mentioned in the discussion around choosing the indicators we underpin the judgement about predictive accuracy with a graphical representation. Figure 3 contains for each imputation method a plot of imputed vs. true values. The plots clearly confirm the quality ranking obtained on the basis of the indicator. Nevertheless, they show also that the risk of a large difference between an imputed and a true value can not be excluded in any of the considered methods.

26. 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.

FIGURE 3. True vs. imputed production values (million CHF).



B. VARIANCE IN THE CASE OF RATIO IMPUTATION

27. As we choose ratio imputation for treating non-response of large firms we get the following estimator for the total production value of large firms in the considered activity sector:

$$\begin{aligned} \sum_L \widehat{y}_i(t) &= \sum_{i\in R_L} y_i^*(t) + \sum_{i\in R_L^c} \frac{\sum_{R_L} y_k^*(t)}{\sum_{R_L} x_k} x_i \\ &= \frac{\sum_{R_L} y_i^*(t)}{\sum_{R_L} x_k} X_T, \end{aligned} \tag{11}$$

where X_T denotes the total of x in L. From (11) follows that the estimator is equivalent to the ratio estimator applied to the net sample of respondents. Hence the properties (variance, bias) of the ratio estimator as discussed for example in (Särndal, Swensson, and Wretman 1992) carry over to the estimator based on ratio imputation. Under the assumption of uniform response probabilities a standard variance estimation using the linear approximation of the ratio estimator leads to an estimated coefficient of variation of

$$\widehat{CV} = \frac{\sqrt{\widehat{V}(\sum_L \widehat{y}_i(t)|m_L(t)=39)}}{\sum_L \widehat{y}_i(t)} = \frac{123473}{6051935} \approx 2\%$$

This is in line with a precision goal of about $CV = 2.5\%$.

VII. CONCLUSIONS

28. In view of a quality assessment for imputation methods one has to define the main criteria for quality. A natural way seems to base these criteria on the users needs with regard to the imputations.

29. In the considered case, where imputations are intended to be used for the estimation of totals and as a support for subject matter specialists when discussing extrapolation results, predictive and estimation accuracy have been chosen to assess imputation quality.

30. While one indicator was considered as sufficient to measure estimation accuracy with regard to the estimation of a total, the assessment of predictive accuracy should be based on several indicators reflecting the desired predictive properties of the imputation methods or complemented with a graphical evaluation.

31. 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.

32. Due to their huge impact on final results, the non-response treatment for large firms is a critical factor with regard to precision goals. In the considered simulation study the indicator for estimation accuracy led on average to relative imputation errors of 1.18% for ratio and 5.93% for mean imputation. This shows the importance of good auxiliary information and a careful choice of the imputation method when imputing for large firms.

33. For simplicity this paper focused on the quality assessment of imputation methods within one activity sector. The same procedure can be repeated in the other sectors, which in general leads to similar conclusions as in the considered case.

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.

Särndal, C., B. Swensson, and J. Wretman (1992). *Model Assisted Survey Sampling*. Springer.