WP.34
English only

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Madrid, Spain, 20-22 October 2003)

Topic (iv): Data editing by respondents and data suppliers

Non-response Recovery by Imputation using Temporal Extrapolation of (administrative) Profit & loss account data in the Structural Business Survey.

Supporting paper
Submitted by Statistics Belgium[1]

I. Introduction

1. The Structural Business Survey aims to provide all data necessary to fulfil the different requirements of the structural business statistics as described in the EU regulation 58/97.

2. The Survey uses a stratified simple random sampling selection scheme. Companies employing 50 persons and more constitute an open-end take-all stratum.

3. Grossing up using 'post stratification' to compensate for non-response of companies in that stratum may yield biased totals and high variation coefficients.

4. Two procedures were developed to impute missing data, using information from administrative records. Profit & loss account data are summarized in the balance sheet deposited yearly at the Belgian National Bank. These accounting totals are the key values that are inserted into the survey questionnaire. Any breakdown of those totals needs to be estimated. The two imputation systems considered, consist in:

- Using ratio estimation on grossed up data of similar respondent companies (same NACE class if possible and same size band)
- Using ratio estimation on data of that same respondent from the previous survey

5. The pro's and contra's of both approaches are discussed. We finally favoured the second approach whenever possible.

II. Obtaining comparable data

6. In order to obtain comparable data within the EU, variable definitions need to be uniform. However, the standardization of accounting is not yet finalized, nor is the actual accounting practice. Questionnaires are reflecting the nationally defined financial standards and therefore differ between the member states.

7. For example: in order to obtain the turnover as defined by Eurostat D2 (**EU12110**), several questionnaire items need to be added or subtracted. Some of them may be totals used in accounting, others are merely details of a breakdown of those totals, used only on the questionnaire. The Belgian translation of

[1] Prepared by Guy Vekeman, Methodology Department of the National Statistics Institute of Belgium, Rue de Louvain 44, 1000 Brussels, guy.vekeman@statbel.mineco.fgov.be

the Eurostat definition of turnover, implies two accounting totals (underlined) and four questionnaire items that need to be subtracted from the sum of both totals:

EU12110 = (Turnover from normal operations)
- (Subsidies booked in turnover)
+ (Miscellaneous recurrent income)
- (Compensations and subsidies)
- (Profits realized on sales of material assets, as compared to their accounting value)
- (Indemnities received)

8. As a first approximation, one could estimate how good both administrative result sheet totals are as a proxy for the EU12110. The correlation between EU12110 and the turnover from normal operations, taken over the whole sample, is very good, and it may be tempting to presume that the five other contributions can be neglected. However, a good proxy for the total of the sectors covered by the survey is not nearly well enough as a proxy for the individual domains of interest.

9. The correlation below was calculated for 25552 companies, taken from the 2001 sample, that yearly need to deposit their balance and result accounts at the National Bank of Belgium.

Correlations

		OUTTOCATOT	V01_12110
OUTTOCATOT	Pearson Correlation	1	,999**
	Sig. (2-tailed)	,	,000
	N	25552	25552
V01_12110	Pearson Correlation	,999**	1
	Sig. (2-tailed)	,000	,
	N	25552	25552

**. Correlation is significant at the 0.01 level (2-tailed).

(2)

10. The miscellaneous recurrent income (notation OUTOTTOT), on the other hand, correlates only weakly with the EU12110 turnover, as is shown in the second table:

Correlations

		V01_12110	OUTOTTOT
V01_12110	Pearson Correlation	1	,332**
	Sig. (2-tailed)	,	,000
	N	25552	25552
OUTOTTOT	Pearson Correlation	,332**	1
	Sig. (2-tailed)	,000	,
	N	25552	25552

**. Correlation is significant at the 0.01 level (2-tailed).

11. The individual contributions that are subtracted do not correlate any better with the corresponding total turnover from normal operations or miscellaneous recurrent income. Correlation coefficients vary from 0,07 to 0,26.

12. This implies that a general ratio estimate will certainly lead to poor estimates for several domains of interest. Any additional knowledge may however improve the estimate rather dramatically.

[2] 'V01_12110' was taken as notation for EU12110 and 'outtocatot' for the turnover from normal operations.

III. Using additional individual data

13. As an example, let us first focus on **subsidies received**. If restricted to companies receiving subsidies booked in turnover, the amount received correlates rather well with the turnover from normal operations. A correlation coefficient of 0,966 was calculated for 389 selected companies, whereas the correlation is extremely poor (0,084) if all companies are considered. Whether or not a company may receive subsidies is determined mainly by its economic activity.

14. A top 5 of NACE divisions, obtained by crude addition of the amounts of subsidies received (irrespective of the weight of the sector in turnover, value added or employment) shows that gross trade (51), construction (45), miscellaneous services to companies (74), retail trade (52) and hotels, restaurants and catering (55) account for more than half of the total subsidies. Manufacturing industry is manifestly absent in this shortlist. Yet, NACE division is not a reliable indicator for determining whether to impute subsidies for a non-responding company. The shortlist above contains only NACE divisions with several thousands of companies, most of which do not receive any subsidies at all.

IV. Various imputation schemes

15. In an economic quantitative survey, one is restricted in the range of imputation schemes that can be applied. Hot-deck imputation, such as applied for many 'qualitative' or appreciation questions in many surveys, is inappropriate. An imputation procedure needs to yield a complete survey form that fulfils all consistency requirements imposed: sum of the details of a breakdown needs to add up to the total, total of income minus total expenditure needs to equal the operational result (profit or loss). Several inequalities need to hold: e.g. subsidies received cannot exceed turnover; total of exports cannot exceed turnover. Other restrictions are less absolute and merely indicate a confidence interval, such as for the average personnel cost per employee. Anomalies may exist and can generally be explained but should be exceptional.

16. Applying some kind of ratio estimation – using ratios drawn from one or several correct survey forms to calculate any breakdown of the totals obtained from the result account (the auxiliary information) of the reference year concerned – should lead to acceptable estimates.

17. Basically two methods can be used:

- If a correct form was obtained earlier, questionnaire items of the non-responding company from a prior survey can be used. (Temporal extrapolation)
- Questionnaire items of a set of responding companies similar to the non-responding one can be used. (Additional appropriate collective data)

V. Using temporal extrapolation on the breakdown of accounting totals

18. As was mentioned previously, the use of any additional individual data may improve the estimates considerably. The example of the correlation between subsidies booked in turnover and turnover from normal operations was given as an example. The simplest way to exploit the use of all knowledge about the individual company is to use a prior questionnaire that was completed correctly. This procedure assumes that the temporal correlation (between successive years) of the breakdown of accounting totals of a given company is on average better than the correlation of the breakdown calculated from a set of companies from the same or a neighbouring stratum in the same survey.

19. However, there are some restrictions. Although the previous survey data are generally convenient as a donor for the calculation of ratios, similar problems may rise as stated before. A zero total in the year for which data are available, can never be used to obtain meaningful ratios for an estimate. Suitable heuristics may in this case make use of the ratios calculated by use of additional appropriate collective data, as is described below. More problems rise whenever a questionnaire is altered. After adding breakdown items, no estimate can of course be provided from a questionnaire where that item did not exist yet. Collapsing breakdown items implies two modifications to the program in successive years.

VI. Using appropriate aggregated data

20. If there is a substantial amount of auxiliary information available at the time when the estimates are to be calculated, an imputation procedure may rely on this auxiliary information. A stratified survey does have a clear advantage whenever imputation is concerned. Any information of responding companies from the same stratum can be used to calculate ratios to be used for the breakdown of the totals of the result account of the non-responding company. This method seems pretty straightforward and can be implemented readily in a program that should provide the estimates (the non-response recovery). The clear advantage is that it can be used, given that the auxiliary information on the non-respondent is available for the reference year. The program can be tailored to take in account any additional data that may facilitate the calculation of the breakdown. There are however some unfortunate circumstances: not in all strata enough responding companies can be found, some strata are in fact empty or contain just one company. Non-response recovery using the above scheme is impossible in those circumstances. Secondly, not all ratios can be calculated using one single company as a 'donor'. If a certain accounting total equals zero, its breakdown is most likely all zero as well. (If negative items were to compensate positive contributions, no ratio can be calculated either.) Therefore, a union of strata pertaining to the same NACE-3 group/size band is observed, rather than the NACE-4/size band original strata. If the number of donors is still too low (three or less), more strata need to be aggregated. This does stabilize the ratios obtained, but it definitely makes them less suitable, since the 'donor' companies will be less similar to the non-responding one. Depending on what items are involved, different aggregations of strata may yield better estimates, however a decision tree to implement such a variable set of 'donor' companies implies the kind of heuristics that is rather difficult to implement in a program.

VII. Evaluation of the two procedures

21. There is no sound proof to demonstrate that the second procedure is better than the first one or the other way around. One clear indication in favour of the *temporal extrapolation* is that the variability of the ratios between breakdown items and their total is not affected. The latter procedure imputes the mean ratio and therefore concentrates its distribution around the mean.

22. One can evaluate the quality of the imputation procedure on any example by comparing the imputed values to true values. For a given (*responding*) company, the *two procedures were used to operate on a set of accounting totals*. Evaluating the quality then is equivalent to measuring in a proper way the 'distance' between the true values and both sets of imputed values.

23. A certain difference for a breakdown item of an accounting total (in euro) is clearly less severe than the same numerical difference for the total of hours worked, the number of employees, the number of local units or some variables that may be used for classification.

24. For that last type of variables the latter imputation method -using appropriate aggregated data- is not suitable, since no valid 'averages' can be calculated. Hot deck imputation should be used to circumvent this problem. In the former procedure, the prior value is copied and imputed. One could presume this to be equivalent to cold deck imputation using the prior questionnaire as a donor.

25. A heuristic measure of the distance can makes use of following intuitive formulas:

$$\mathrm{d}_{e,t} = \sqrt{\sum_{j=1}^{n} w_j^2 \cdot (e_j - t_j)^2} \quad \text{or alternatively:} \quad \mathrm{D}_{e,t} = \frac{2 \cdot \sqrt{\sum_{j=1}^{n} w_j^2 \cdot (e_j - t_j)^2}}{\sum_{j=1}^{n} w_j \, t_j}$$

The 'e' are estimates, the 't' are true values and the 'w' are weight factors for each of the breakdown elements or miscellaneous questionnaire items. The former measure $d_{e,t}$ is a weighted Euclidian distance between the estimate e and the true value t. The latter one $D_{e,t}$ is its relative counterpart. The index j runs through all n questionnaire items. It is clear that for all j-values referring to a total, the estimate coincides with the true value. Therefore an extra coefficient 2 was added in the relative distance formula to account for this. Implicit hypothesis: there is a breakdown for every accounting total.

26. For the weight factors following round numbers were used:

- Qualitative or classification variables: 10.000.000
- Number of employees, related breakdowns: 10.000
- Number of hours worked: 100
- Accounting item breakdown: 1 (monetary unit is 1 Euro)
- Accounting auxiliary value: 1

27. The procedures given were used on the *datasets of a major food processing company*. The temporal extrapolation of the breakdown of accounting totals yielded following distances:

$d_{e,t} = 36{,}0\ 10^{6}$

$D_{e,t} = 35{,}7\ 10^{-6}$

The use of aggregated data from the NACE-3 x size band stratum gave following distances:

$d_{e,t} = 274\ 10^{6}$

$D_{e,t} = 272\ 10^{-6}$

28. Last estimate was made using 23 companies from the same NACE-3 x size band stratum. No further stratum aggregation was required. The second estimate is still acceptably well, but its distance from the true values exceeds the former one by a factor of 7,5.

29. This ratio is by no means systematic. It is possible to find a company for which the method using aggregated data from the NACE-3 x size band stratum, would lead to a rather dramatic miscalculation. This is the case when the company to be estimated is by no means typical for the sector. For example: the urban public transport companies should not be approximated using the ratios calculated from average freight transport companies.

VIII. Implementation

30. Both imputation procedures are actually applied at Statistics Belgium. The temporal extrapolation of the breakdowns of accounting totals is used for data recovery with the larger non-respondents. It is implemented in a visual-basic module in MS-Excel. Several spreadsheets are opened, containing accounting information over several years and the data from the prior survey, which is exported from an application in MS-Access. This sheet is updated and the identifiers are adapted. Using a proper MS-Access query, the new values can then be appended to the database. The procedure takes about fifteen minutes to be completed by a qualified operator. This is less than half the time it requires to check a true questionnaire for consistency and input the data.

31. The use of ratios calculated from aggregated data is used for data recovery mainly for SME's above 10 employees[3] in industry and above 20 in other sectors, where sample sizes per stratum tend to be somewhat larger. Occasionally that procedure is also applied for data recovery of larger units when no comparable data existed in a prior survey. This procedure needs to be performed by a statistician on a PC. Due to the structure of the database and the use of ODBC to gain access to the database on a mainframe computer it is

[3] No imputation procedure is used to compensate for non-response of small SME's. Reweighting the respondents is thought not to affect the true totals in this case. (Assumption of 'missing completely at random')

intrinsically slow. However, after initialisation, it requires little input and can be operated as a 'background task' or on a dedicated PC. A discomforting collateral phenomenon is the propagation of small non-zero values: if just one or a few of the donor companies book a rather small amount under a specific code, a tiny ratio is calculated that will propagate small non-zero imputed values, whereas most entries ought to be zero.

32. For security reasons both initial values and adapted or final values are logged. Editing existing data only affects adapted values and the changes are flagged. A (different) flag is also set in the imputation procedure and distinction can even be made according to the procedure used. However, if the temporal extrapolation is printed out and values are registered using the normal data input program (as is often done for practical reasons) no flags are set with the values. In this case a comment is added to the input series, distinguishing true data from imputations. Follow-up of this information is however cumbersome, all comments are text fragments of varying nature.

33. The response rate has been declining as a result of budget cuts in the inspection department, which takes over data collection if a second reminder is not fruitful. This implies that imputation procedures are now used more frequently than what they were meant for initially. There is however an obvious danger in the cascaded use of imputation procedures, since imputed values gradually may diverge from true values.