

Quality of the data editing techniques in the annual business surveys : results of the analysis

Pascal Rivière, INSEE, France

The Annual Business Surveys are conducted each year on approximately 200 000 businesses. The sample is drawn in a population of 1 900 000 businesses belonging to six economic sectors (manufacturing, agricultural industry, trade, services, transportation, construction, but not finance).

Before 1996, there were six different production processes. A large project was launched in 1992 in order to modernize and to harmonize the surveys : common set of variables (mainly 75 quantitative variables), common metadata, common sampling tools, common data editing tools, common imputation methodologies, regular use of administrative sources, common tool to take into account restructurings.

The aim of this paper is to describe the results obtained in terms of quality. Indeed, the numerous metadata in the database allow to derive a certain number of quality indicators. These results were obtained on the services sector (63500 businesses).

Edits

We use three kinds of edits :

- ratio edits (e.g. value added / turnover)
- growth rate edits (e.g. number of employees (n) / number of employees (n-1))
- balance edits (any balance equation, for example on totals)

For each edit, interval bounds were computed by using Hidioglou-Berthelot approach. As there are approximately 90 edits (it changes each year because of improvements in the data editing process), 1000 editing strata (4 ranges of number of employees X 250 economic activities in the six sectors), and 4 bounds for each ratio (a small interval and a large interval), approximately 400 000 bounds have to be computed each year before starting the data editing process.

Imputation

In case of partial nonresponse or erroneous response, the data editing tool can apply four imputation techniques :

- deterministic
- auxiliary trend
- using an auxiliary variable and a ratio of means
- smoothing (using a balance equation)

In case of total non response, the system uses either random stratified hot deck or n-1 values (and a mean trend).

Metadata

For each variable, three values are kept : raw value, modified value (if manually modified after phoning the business), imputed value. If the initial value is not modified, these three values are equal.

For each variable, the system keeps an « imputation code » (0 if the variable is not imputed), which describes the imputation method that was applied.

For each unit (e.g. enterprise), there is a quality code, the highest value (19) meaning that the current set of values of the unit is acceptable to build statistics, the lowest values (1,2,3) meaning that the unit can be considered as a total nonresponse, and the intermediate values corresponding to different kinds of lacks or inconsistencies in the set of values. In practice, it gives priorities to manual editing : the units with the lowest qualities are checked first. Of course, after manual updatings, the quality code of the unit increases. That is why the system keeps two values : the current quality indicator, and the initial quality indicator.

For each edit, whenever the edit fails, there is a confirmation code, indicating if the variables of the edit were manually confirmed or not.

Indicators

We computed the following indicators :

- response rate
- manual checking rate
- repartition of non respondents
- % accepted (not manually checked) raw data
- % accepted imputed data
- % manually modified data (measures the impact of clerical work)
- among modified values :
 - % accepted
 - % imputed
- % confirmed raw data
- impact of imputed data (modified data, confirmed data) in the aggregates, and contribution to the difference between raw aggregates and final aggregates.
- concentration of manual modifications

General results

We will focus here on three variables (production, number of employees (in full time equivalent), investment), but some results are general (for example the response rate).

In the services sector, there were 63429 units in the initial sample. At the end of the process, 55776 were used (7653 were considered out of scope). Among these 55776, 44770 were respondents : the response rate is approximately 80%.

Among the nonrespondents :

- 32.4% were considered a posteriori out of scope

- 53.4% were imputed by random stratified hot deck
- 12.8% were imputed by using n-1 values (and a mean trend)
- 1.4 % were imputed but not used in final statistics

Among the respondents :

- **23%** were not manually edited. That means that approximately one fourth of the units were automatically accepted without any clerical work.
- **33.8%** had initially a low quality level (if the set of raw values had been kept without manual checking, these units would have been imputed as if they were total non responses)
- the others had non crucial problems as, for example : main activity not coded (still a text, not a code)

Impact of total nonresponse imputation in the aggregates :

- production 13.3%
- number of employees 10.6%
- investment 7.8%

That means, for example, that 7.8% of the total investment of the services sector has been obtained by total non response imputation.

Fraction of accepted raw data :

- production : 84.8% of the questionnaires (69.4% of the final aggregate)
- number of employees : 81.4% (70.3% of the final aggregate)
- investment : 87.6% (49.5% of the final aggregate)

Then we can see that the acceptance rate is high, but this rate is not enough to understand what happens, as the accepted values seem to have a smaller impact, particularly for investment.

Impact of partial nonresponse (or erroneous response) imputation

When the raw value is rejected by the system, it does not mean that the value will be imputed. Three situations can occur :

- the value is imputed
- the value is manually modified
- the raw value is confirmed

Let us analyse the results for our three variables :

Production (15.2% were rejected) : 4.6% were imputed, 7% modified, 3.6% confirmed.
 Number of employees (18.6% rejected) : 8.9% imputed, 4.3% modified, 5.3% confirmed.
 Investment (11.2% rejected) : 9.5% imputed, 1.5% modified, 0.2% confirmed.

Imputation methods used and their impact on aggregates :

- production : 4.6% imputed, which represents 3.2% of the aggregate ; deterministic 3.1% (2.9% of the aggregate), ratio of means 1.1% (0.2% of the aggregate) ; contribution of those imputations to the difference between raw aggregate and final aggregate : 1.9%.

- number of employees : 8.9% imputed, which represents 4.4% of the aggregate ; deterministic 5.3% (1.8% of the aggregate) ; contribution of those imputations to the difference between raw aggregate and final aggregate : 1.8%.

- investment : 9.5% imputed, which represents 12.7% of the aggregate ; ratio of means 5.4% (5% of the aggregate), smoothing 4% (7.6% of the aggregate) ; contribution of those imputations to the difference between raw aggregate and final aggregate : -1.9%.

Impact of clerical work

As we saw, the value can either be confirmed (the value remains unchanged, then final value = raw value in general) or modified by the clerk.

Production : 10.6% of the questionnaires were manually checked, which is 27.4% of the aggregate ; 3.6% confirmed (6.6% of the aggregate), 6.9% modified (20.8%); contribution to the difference between raw aggregate and final aggregate : -19.1%.

Number of employees : 9.7% manually checked (25.1% of the aggregate) ; 5.3% confirmed (4.5% of the aggregate), 4.3% modified (19.7%); contribution to the difference between raw aggregate and final aggregate : -36.5%.

Investment : 1.7% manually checked (37.7% of the aggregate) ; 0.2% confirmed (20.6% of the aggregate), 1.5% modified (17.1%); contribution to the difference between raw aggregate and final aggregate : -33.9%.

Concentration of modifications

The idea here is to look at the largest differences between raw value and modified value, in order to see if the modifications are concentrated or not. The absolute differences are sorted by decreasing value, then it is possible to compute the cumulated frequency and the cumulated sum of absolute differences, using this sorting.

For the variable « production », the 1% largest modifications represent 89.4% of the total (total sum of absolute differences), the 10% largest are 96.1% of the total.

Number of employees : the 1% largest modifications correspond to 92.7% of the total, and the 10% largest represent 96.9%.

Investment : the 1% largest modifications represent 98.4% of the total, and the 10% largest represent 99.8%.

Final remarks

The first conclusion is that the clerical modifications are unavoidable, highly concentrated, and that their impact on the results is far bigger than their impact of partial nonresponse (or erroneous response) imputation. This is not new : many NSIs arrived to the same kind of conclusion. Analyzing the kinds of errors, building more efficient edits are the new challenges.

Then, for the future, the main interest of these indicators is to give elements to improve the editing and imputation process (edits, edit bounds, imputation methods), and even the whole survey process (questionnaire design, ways of collecting data). Moreover, comparing these indicators from one year to another allow managers and statisticians to have a better understanding of the improvement of the « quality » of the process in general.