Improving the Efficiency of Data Editing and Imputation for a Large-Scale British Annual Business Survey

Alaa Al-Hamad(Office for National Statistics)
Pedro Luis do Nascimento Silva(Southampton University)
Gary Brown(Office for National Statistics)

Abstract

This paper reports results from a project to evaluate and improve the editing and imputation approach adopted in the Annual Business Inquiry – Part 2 (ABI/2). This is joint work carried out by the UK Office for National Statistics (ONS) and Southampton University's Statistical Sciences Research Institute. The ABI/2 is a large scale annual survey covering most sectors of the British economy, with an annual sample of around 60,000 businesses. We examined detailed specifications for the current editing and imputation processes, and their connections to data collection instruments, data capture, and estimation methods. A variety of quality indicators, impact measures, and statistical editing and imputation techniques were tested on three years of pre- and post-edited data. A number of alternative approaches to the overall editing and imputation process were investigated to maximise efficiency without impacting negatively on quality. Preliminary results suggest that these will yield increased benefits to the survey.

Introduction

In recent years the ONS has been undergoing a large efficiency programme to its business statistical surveys with a vision mainly concerned with increasing the efficiency of data processing. This programme has three aims:

1. to cope with the ever increasing demand for statistical information within a shorter time period, at lower cost and at a much higher level of detail and quality than before;
2. to meet financial pressures; and
3. to generate savings for reinvestment into improved outputs.

Currently one of the most costly components in the survey processing procedures in ONS is data cleaning. This is because to fulfil our role as a National statistical agency successfully, the organisation is always under pressure to produce high-quality data in order to maintain public confidence in official statistics. This leads to ONS business surveys having high cleaning costs – on average consuming around 40% of the total survey budget. Furthermore, most of the resources within data cleaning are consumed by data editing. In practice, editing consists mainly of two stages:

- Error localisation/detection;
- Error verification / correction.

The very high costs are generally due to use of labour intensive processes for error verification / correction, which often involve re-contacting respondents. The requirement for manual intervention is introduced through the nature of the errors

commonly found in business surveys. Some errors can be corrected through the application of logical editing rules, and some through recalculation of totals, but a large proportion are often simply "suspect data" that need to be queried directly with the businesses that provided the responses. These suspect data are rarely found to be in error - studies have found the "hit rate" to be as low as 20% (Rivière 2002) and this is found to be true for ONS surveys. Therefore, these suspect data are often simply confirmed by the respondents and hence pass through the data editing process unchanged. The benefits from querying these confirmed data are often intended to be informative rather than statistical, i.e., they help explain movements in the data.

In the ONS a number of methods have been used over the years to cut costs and achieve more efficient editing processes. Some of these methods prioritise suspect data by their importance, i.e. selective editing and (to a degree) (Hidiroglou and Berthelot 1986). Others correct these suspect data automatically through recalculation and the application of logic. However, although such methods are employed for some surveys it was thought they are not fully utilised to achieve the maximum efficiency across all business surveys. As a matter of fact, some surveys have not being reviewed for a number of years which resulted in them costing a considerable amount of resources and not achieving the optimum results required.

Therefore the aim of the study on which this paper is based is to review the current data editing process for a large ONS business survey, ABI/2, and to suggest ways in which efficiency can be improved. The approach adopted considers a holistic view of the editing process, rather than only some of its individual stages. Detailed specifications for the current editing processes will be examined, as well as their connections to data collection mode, data capture and the current editing methodology. The broader aim is mainly to learn from this exercise, and assess what possible improvements could be made to other ONS business surveys. However this paper will focus only on one aspect of this study, the consideration and use of selective editing procedures in an attempt to recommend more efficient alternative to the current editing procedure.

Basic Characteristics of the ABI/2

ABI/2 is the survey providing structural financial information about UK businesses on a yearly basis. This survey was formed in 1998 by pulling together several annual surveys previously carried out for specific sectors of the economy by ONS or its predecessors (Smith, Pont and Jones 2003). It provides estimates of totals for each section of the economy, with a coverage that includes sections A to O – see (ONS 2007c) – of the Standard Industrial Classification (SIC) of economic activities, for the variables listed below

Variable name / description

- Number of enterprises
- Total turnover
- Approximate gross value added (GVA) at basic prices
- Total purchases of goods, materials and services
- Total employment, point in time
- Total employment, average over the year
- Total employment costs

- Total net capital expenditure
- Total net capital expenditure, acquisitions
- Total net capital expenditure, disposals
- Total stocks and work in progress – value at end of year
- Total stocks and work in progress – value at beginning of year
- Total stocks and work in progress – increase during year.

These results are published on the internet, using tables that present time series for each section of the economy and each selected SIC, for all the above variables.

To collect the data, the ABI/2 survey in 2006 used 42 different questionnaires. These questionnaires are different for each section of the economy, and for several sections (like catering, retail, etc.) there are two questionnaires: a long and more detailed questionnaire, used to obtain data from all the large businesses and from a sample of the smaller ones; and a short questionnaire, which asks for less detailed information from the sampled businesses not getting the long questionnaire.

The sample for the ABI/2 is a stratified simple random sample. The stratification uses a combination of three variables:

- Region (England, Scotland and Wales);
- SIC activity codes (3 or 4 digit, depending on the region);
- Employment size band.

The sample allocation is highly disproportional, with all the large businesses (those with 250+ employees according to the survey frame extracted from the Inter-Departmental Business Register - IDBR) included in the sample with certainty, and sampling fractions that increase with the size of the business. The overall sample size for the 2005 survey edition was 73,955 businesses, with 54,123 respondents. Hence the total nonresponse rate was 26.8%. The main mode of data collection is through paper questionnaires.

For the purpose of the analysis in this paper, data will be used from only one sector of the economy and one questionnaire type: catering sector short form.

Assessing Options for Selective Editing

In this section we discuss the options being considered for tackling the efficiency of the editing on ABI/2 using an approach based on selective editing ideas (Hedlin 2003), (Lawrence and McKenzie 2000) (Latouche and Berthelot 1992).

In selective (or significance) editing, the key idea is that only a subset of the survey data are fully verified, or are verified using costly methods, while the editing of the remaining data is dealt with using cheap methods. Hence the survey records must be split between *critical* and *non-critical* sets. The critical records (those having an important expected impact on the final estimates) are submitted to all edit rules, and referred to reviewers whenever they fail any edits. The non-critical records are submitted to a smaller set of edit rules (or even no edit rules), and any edit failures are dealt with by automatic imputation.

The goals of using selective editing are to reduce survey costs, survey processing time respondent burden (by limiting re-contacts), and to avoid or reduce over-editing. It should also help staff focus their attention on cases with the highest likely impact on survey estimates.

To implement selective editing, a key decision is the choice of a score function, namely a function that is going to be used to split records between the critical and non-critical streams. There are two alternative approaches to setting up score functions:

- Estimate related – score functions computed taking account of target survey estimates;
- Edit related – score functions which depend on a specified set of edits.

For the ABI/2 we will test a number of alternatives developed under both approaches. For the estimate related score functions, there are two main options under consideration. The first is simply

$$d_k = w_k \sum_i |y_{ki} - \hat{y}_{ki}| \quad (1)$$

where y_{ki} is the a raw value of variable i for unit k, $\hat{y}_{ki}$ is a predicted or anticipated value for the same variable and observation, and w_k is the survey weight attached to unit k. Note that y_{ki} may or may not be suspect at this stage. Only records with values of d_k larger than a specified threshold c would be submitted to the current manual revision and possible respondent re-contact procedures. Some alternative versions of this score function would consider:

a) All variables for which total estimates are published;
b) A subset of variables for which estimates of totals are published.

In addition, the edit-related score functions RATIO, FLAG and DIFF proposed by (Latouche, et al. 1992) would also be considered.

Evidence from analysing edit failure rates for one sector using the current ABI/2 data suggest that there is good potential for savings by using selective editing. We examined the frequency distribution of the number of edit failures per record, considering both raw (unedited) data and final (clean, edited) data for two years. In both cases the responding businesses are the units of analysis. Tables 1 and 2 present the frequency distributions of number of edit failures per record for the catering short questionnaire in the two years considered. These tables provide a very brief summary of the editing process, but are useful because they allow us to assess how much editing takes place. The same set of edits applied in both years, and it contained 18 edit rules.

First, consider the proportion of records which are ‘clean’ on arrival, i.e., failed no edits – see the columns referring to unedited data sets. These proportions were 35-36% in 2003-2004. The corresponding complementary proportions were around 64-65% for 2003-2004, and represent the fraction of the ABI/2 records which were reviewed manually as part of the validation operation. These are large proportions, implying a substantial workload for the ONS validation unit. Examining the corresponding proportions computed from the final edited data, we observed that around 45-47% of the records would still be flagged by some edits, meaning that they

had been verified but either did not change, or if they did, the changes applied could not clear all edit failures initially detected.

Table 1 – Frequency distribution of number of edit failures per record – catering short questionnaire - 2003

Failures per record	Unedited dataset			Edited dataset		
	# of records	% of records	# of failures	# of records	% of records	# of failures
0	659	34.8	0	1,004	53.1	0
1	458	24.2	458	506	26.8	506
2	416	22.0	832	273	14.4	546
3	157	8.3	471	59	3.1	177
4	97	5.1	388	41	2.2	164
5	57	3.0	285	6	0.3	30
6	28	1.5	168	2	0.1	12
7	15	0.8	105	-	-	-
8	3	0.2	24	-	-	-
9	1	0.1	9	-	-	-
Total	**1,891**	**100.0**	**2,740**	**1,891**	**100.0**	**1,435**

Table 2 – Frequency distribution of number of edit failures per record – catering short questionnaire - 2004

Failures per record	Unedited dataset			Edited dataset		
	# of records	% of records	# of failures	# of records	% of records	# of failures
0	624	35.9	0	946	54.5	0
1	420	24.2	420	464	26.7	464
2	375	21.6	750	259	14.9	518
3	163	9.4	489	56	3.2	168
4	80	4.6	320	9	0.5	36
5	46	2.6	230	1	0.1	5
6	18	1.0	108	1	0.1	6
7	8	0.5	56	-	-	-
8	2	0.1	16	-	-	-
Total	**1,736**	**100.0**	**2,389**	**1,736**	**100.0**	**1,197**

A synthetic measure of the size of the editing operation can be obtained as the total number of edit failures each year, namely the sum of the cross products of the frequencies of records by the numbers of edit failures per record in tables 1 and 2. The validation unit handled 2,740 edit failures in 2003, and 2,389 in 2004, a reduction of 12.8%.

Another interesting analysis which can be derived from the calculations of total numbers of edit failures is the proportion of “false alarms”, namely the ratio between the total number of failures in the edited data divided by the number of failures in the unedited data. Assuming that the data are “error free” at the end of the editing process, this indicator summarizes the degree to which the edits flag false error situations. The values of this proportion were 52.4% for 2003, and 50.1% for 2004. This demonstrates that there is substantial potential to reduce the editing effort by revising the edits and/or the editing approach, because half of the editing effort did not result in changes that would remove the edit failure markers in 2003-2004.
So the apparent *false hit rate* of the editing operation was quite high (@ 50%).

In addition to testing some alternative score functions, a thorough review of the edits currently applied to the survey is under way. This review starts with assessing hit rates for each individual edit, and attempts to locate potential factors causing the highest hit rates, looking at questionnaire design, instructions for filling in questionnaires, methods used for data capture, etc. This review shall provide the basis for redesigning the edits applied to the data, also paving the way for use of automatic editing and imputation systems to handle the non-critical cases as indicated by the selective editing.

Concluding Remarks

The analysis in the previous section demonstrated that there is potential to improve efficiency of the editing without perhaps affecting the quality of the resulting output. At this stage, these findings are restricted to a single sector of the economy, and to the short form only. Our work proceeds in extending the analysis to other sectors, as well as to the testing of alternative score functions, using the raw and final datasets from the ABI/2 for 2003-2005. The final paper shall contain the results of this project.

References

Hedlin, D. (2003), "Score Functions to Reduce Business Survey Editing at the Uk Office for National Statistics," Journal of Official Statistics, 19, 177-199.

Hidiroglou, M. A., and Berthelot, J.-M. (1986), "Statistical Editing and Imputation for Periodic Business Surveys," Survey Methodology, 12, 73-83.

ONS. (2007c), "Sections a to O – Whole Economy. Press-Release About Abi/2 Results.," 2007, Tables with key outputs of Annual Business Inquiry Part 2.

Latouche, M., and Berthelot, J.-M. (1992), "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys," Journal of Official Statistics, 8, 389-400.

Lawrence, D., and McKenzie, R. (2000), "The General Application of Significance Editing," Journal of Official Statistics, 16, 243-253.

Rivière, P. (2002), "General Principles for Data Editing in Business Surveys and How to Optimise It," Conference of European Statisticians.

Smith, P., Pont, M., and Jones, T. (2003), "Developments in Business Survey Methodology in the Office for National Statistics, 1994-2000," The Statistician, 52, 257-295.