Proceedings of Q2006
European Conference on Quality in Survey Statistics

# Retaining Output Quality whilst Reducing Validation Costs in the ONS

Begoña Martín, Alaa Al-Hamad, Gary Brown[1]

Key words: validation, savings, quality

## 1. Introduction

Within a well designed and well performing survey, increased quality leads to increased cost. Likewise, decreased cost generally leads to decreased quality. However, financial pressure more often causes cost to decrease than increase, for example, as a result of implementing an efficiency drive, or in an attempt to achieve value for money.

One of the most costly components of ONS business surveys processes, approximately 40% of total survey costs (Granquist and Kovar, 1997), is data cleaning. A major element of data cleaning is recontacting the respondent. We only recontact if response data have failed validation rules.

In this paper, these validation rules will be split into two types: those that detect hard errors i.e. check for missing data, and those that check soft errors, i.e. check for unusual response. The parameters of soft validation rules determine which responses fail and which pass.

Good data failing means wasted effort – bad data passing means quality loss.

ONS accountability means we need to reduce wasted effort, and current funding limitations mean we need to reduce costs. To reduce costs, soft check parameters need to be changed to reduce the number of failures, but this lets through more bad data, and quality is lost.

This is the cost vs quality conundrum.

This paper focuses on how we choose validation rule parameters to achieve required reductions in validation failures whilst assessing the impact these reductions have on the quality of the survey output.

[1]Office for National Statistics, Cardiff Road, Newport, Gwent, NP10 8XG, UK
(Begona.Martin@ons.gov.uk, Alaa.Al-Hamad@ons.gov.uk, Gary.Brown@ons.gov.uk)

# 2. Background to Validation

## 2.1 Data Cleaning in ONS

In the ONS validation is defined as searching for suspect data, and subsequent verification of these suspect data. Once data are confirmed in error, they are edited. See Figure 1 for a simplified view of the ONS data cleaning process.

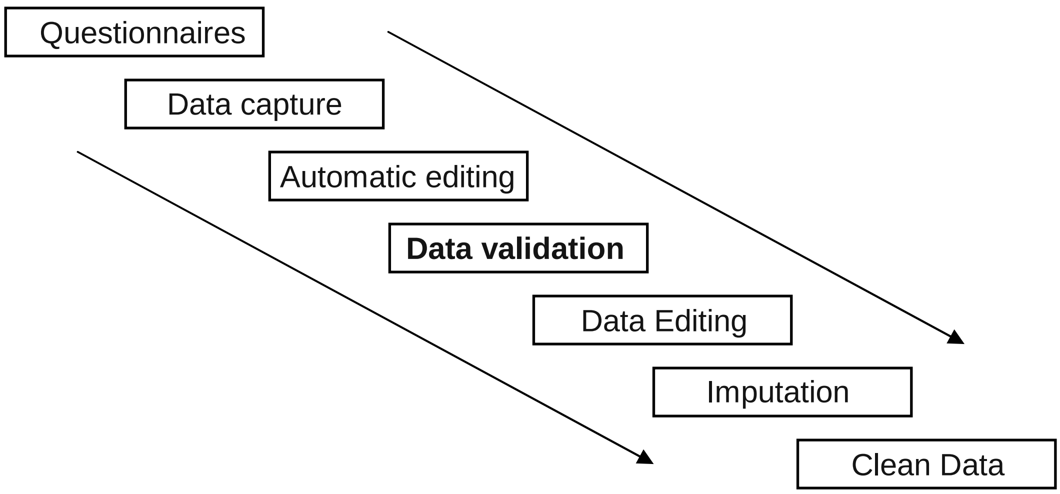


Figure 1: Data cleaning in ONS

Figure 1 represents the procedure for paper questionnaires for ONS business surveys. However, data also arrive through Telephone Data Entry (TDE), by FAX, by telephone, and through the internet.

- *Data capture*

When paper questionnaires arrive, they are scanned to turn the written responses into electronic data. Responses that cannot be scanned accurately (according to user specified confidence intervals) are subject to visual confirmation and data input.

- *Automatic editing*

Despite using of best practice techniques in the design of questionnaires for business surveys, errors still occur as respondents make mistakes. These include systematic errors such as quoting in pounds instead of in thousands of pounds, and basic addition errors. Many ONS surveys automatically identify and correct systematic errors.

- *Data validation*

The principle of data validation is to identify responses which are unlikely to be correct, based on rules. Some standard rules are used for all surveys. Rules usually compare returned or derived data with past data, or with constants.

- *(Selective) Data editing*

Data that fail validation are either confirmed as correct, or edited on the basis of auxiliary/prior knowledge (or by recontacting the respondent). However, for some surveys, selective editing prioritises failures (by their expected impact on final estimates), and only validates those failures above the chosen threshold.

- *Imputation*

After validation and editing, any missing or unusable data are imputed. Imputation in the ONS is usually based on average growth from the previous period.

## 2.2 Current Validation Procedure

### 2.2.1 Error Detection Rules

Rules take the form of simple equations, with the majority having parameters that define the acceptable region for a response. We classify validation rules as follows:

- inconsistencies – responses in the same questionnaire are inconsistent;
- missing values – no response exists;
- logical errors – the response is not possible;
- reasons for value changes – the reason is of interest, not the change; and
- value errors – the response falls outside likely bounds, so is suspect.

The only way to make savings on validation is to either drop rules, or change the parameters in value error rules. In this paper we focus on the latter.

Rules determine which responses are suspect and fail, and which are non-suspect and pass. If rules worked perfectly:

A. Data with error (bad data) would fail; and
B. Data without error (good data) would pass.

However, the truth is that two other less appealing outcomes also occur:

C. Some good data fail; and
D. Some bad data pass.

Therefore, the two objectives of validation are:

{I} maximise A and minimise C; and
{II} maximise B and minimise D.

### 2.2.2 Rule Adjustment and Implications

For value error rules, the parameters determine which responses fail and which pass. If parameters are changed, the numbers failing and passing change. The impact of changing parameters needs to be assessed in terms of the two validation objectives.

- Objective {I}
  The impact of change is quantifiable for this objective. All failures are validated, and so (assuming validation is a perfect science[2]) the numbers of A and C can be measured before and after parameter change.

- Objective {II}
  The impact of change is not immediately quantifiable for this objective. Before the change, the number B+D cannot be split (as there is no validation). Before parameters were changed, a sample of passes would need to be validated to estimate D (and hence by subtraction B). After the change, impact would be easy to calculate. Either the change would cause less passes and the new failures would be validated (i.e. change in D would be an output), or the change would cause more passes and a sample of the new passes would already have been validated (i.e. change in D could be estimated).

Given the aim of this paper, the focus is on objective {I}.

We need to reduce the amount of good data failing validation to make savings[3], but this has major implications. Data quality is reduced as less errors are detected and corrected, and if this leads to poor quality outputs it could be costly in terms of ONS reputation. This presents a challenge – how to retain quality whilst saving costs.

## 3. Cost vs Quality trade-off

### 3.1 Problem

Reducing survey costs, by reducing validation failures, has a negative impact on quality – as fewer errors will be detected and corrected. To reduce this impact, reductions need to be targeted at the poorest performing validation rules. The most obvious measurement of performance is the 'hit rate'.

The failure rate is the proportion of units that do not pass the edits, i.e. the proportion of suspect responses. The *hit rate* is the proportion of suspect responses that are erroneous. It represents, to some extent, the efficiency of identification. If the hit rate is small, it means that many responses were unnecessarily validated. In many surveys the failure rate is high, and the hit rate is low, which means that too many questionnaires are validated. Tate et al. (2001) indicates that in the major ONS business surveys, the failure rate is in the range of 30% to 80%, whereas the hit rate is generally around 20%.

---

[2]To take account of error in the validation process requires A and C to be split into: AY - Bad data fail and are corrected, AN - Bad data fail and are not corrected, CY - Good data fail and are not changed, and CN - Good data fail and are changed. However, to quantify these breakdowns would require intensive follow-up, and a second stage of gold standard validation, and is not attempted here.

[3]Although there is an overall saving, as the validation costs will be reduced, the analysis costs will be increased as the extra errors passed will impact on final results - requiring more drilling down to discover the truth.

The hit rate counts the number of errors detected, and if computed before and after parameter change it determines the number of errors missed. However, the raw count does not demonstrate their size, or their impact on the overall output quality.

In addition to the number of errors missed, this paper defines the relative change (in the survey estimates) as the main quality measure. We define it as follows.

Let *D* be a sector of activity for which estimates are published; *D* constitutes a domain of interest. Let $s_D$ be the set of records from domain *D* and $s_{ED}$ be the subset of records from $s_D$ that fail validation before change but pass after. Let $w_i$ be the weight of respondent *i*, $y_i$ be the response for respondent *i* and $\tilde{y}_i$ the edited value. If a record does not fail then the edited value is equal to the returned value.

The relative change (in percentage) in domain *D* is then given by

$$100 * \frac{\left| \sum_{S_{ED}} w_i (y_i - \tilde{y}_i) \right|}{\sum_{S_D} w_i \tilde{y}_i} \qquad (1)$$

Equation (1) measures the weighted errors missed relative to the weighted estimates under full validation.

Ideally, quality measures should also encompass bias and variance, but these have not yet been developed in this context. As variance is not known for all business surveys, this poses an additional problem.

### 3.2 Solution

Once you have measurement, you can solve the problem. Given a cost savings requirement of £*X*, then the solution is to:

- calculate the number of failures *Z* that £*X* equates to;
- simultaneously change parameters (i.e. permutations within set ranges) for all validation rules;
- calculate the quality measure for these changes;
- sort the permutations by failures and by quality measures within failures (in ascending order); and
- choose the parameter values that cause the smallest quality measure whilst achieving the required reduction in failures.

This is a simplistic view of an optimisation routine. The inputs are: the savings required, the ranges for parameter values, and the quality metrics to use for assessment. The outputs are: the changed parameter values, and the impact of these changes in terms of the quality metrics.

A generic version of this routine is under development. However, a single application has been achieved, on a real ONS business survey.

## 4. Implementation using ONS business survey

### 4.1 Quarterly Stocks Inquiry

The Quarterly Stocks Inquiry (QSI) is a sample based survey covering all of the UK. There are approximately 1.05 million businesses in the 8 sectors covered by the QSI. The overall sample size each quarter is approximately 21,500, i.e. 2% of the population. The response rate is 75%. The QSI collects data on the level of stocks at the beginning and end of each period (i.e. opening and closing stocks).

The primary use of the QSI is estimation of changes of stock (and of work in progress) for the expenditure and income measures of gross domestic product (GDP). In addition, the QSI is used in the compilation of the Index of Production (IOP) which forms part of the output measure of GDP.

The QSI has, on average, a validation failure rate of 50%: approximately 30% of these errors are in the Wholesale & Retail Sector (WS&R). This paper will therefore focus on testing proposed changes in validation rules for the WS&R, in the QSI.

### 4.2 Current and Proposed Rules

Three rules have been identified as potential sources for savings in the WS&R sector.
Rule 1: relative change in opening stocks;
Rule 2: relative change in closing stocks; and
Rule 3: comparison of current opening and previous closing stocks.

Let *K* denote thousands of pounds. Denoting the opening and closing stocks at time *t* for a given respondent *i* as $y_i^{op}(t)$ and $y_i^{cl}(t)$ respectively, then the current rules are as follows.

Rule 1. If $y_i^{op}(t) > 199K$, then fail if $100 * \left|\frac{y_i^{cl}(t) - y_i^{op}(t)}{y_i^{op}(t)}\right| > 40$ (2)

Rule 2. If $y_i^{cl}(t) > 199K$, then fail if $100 * \left|\frac{y_i^{op}(t) - y_i^{cl}(t)}{y_i^{cl}(t)}\right| > 40$ (3)

Rule 3. If $y_i^{cl}(t-1)$ was returned, then fail if $\left|y_i^{cl}(t-1) - y_i^{op}(t)\right| > 5K$ (4)

The proposed rules are based on replacing the existing thresholds with 5 new variable parameters. For example, in Rule 1, the threshold of 199K will be replaced by *Gate1* and the tolerance of 40% by *Tolerance1*. A new threshold Gate2 is then introduced such that if $y_i^{op}(t)$ is greater than *Gate2*, 40% is replaced by *Tolerance2*. The proposed rule resulting from these changes is illustrated below in Equation (5).

Rule 1*. If $Gate2 > y_i^{op}(t) > Gate1$, then fail if $100 * \left| \frac{y_i^{cl}(t) - y_i^{op}(t)}{y_i^{op}(t)} \right| > Tolerance2$

Or (5)

If $y_i^{op}(t) > Gate2$, then fail if $100 * \left| \frac{y_i^{cl}(t) - y_i^{op}(t)}{y_i^{op}(t)} \right| > Tolerance1$

The change to Rule 2 is the same, and for Rule 3 the 5K = “£5,000” threshold is changed to *Gate3*.

### 4.3 Simulation Study

We consider all validation rules existing for the survey simultaneously.

Data simulations use quarterly data for the period 2003-05. For each quarter, the first step is to split the data into two groups: $P_1$ corresponding to those records who failed validation and changed value, and $P_2$ corresponding to those records who failed but did not change value.

The simulation study consists of running the old validation rules on the data to determine the number of records that belong to group $P_1$. Validation is then rerun using the proposed rules, and records in $P_1$ that do not fail the new validation are counted. This count represents the number of genuine errors that will be missed under the new rules. We call it *missed errors*.

Conversely, we run the old validation rules on the data to determine the number of records belonging to group $P_2$. We then rerun the proposed validation rules and count the records from $P_2$ that no longer fail. This count represents the reduction in the number of unnecessary failures. We call it *unnecessary errors*.

This process is repeated for every possible permutation of parameter values in Table 1. At each repetition, values of the relative change and *Savings*, defined as the reduction on the total number of validation failures, i.e.

$$Savings = n_{missed\ errors} + n_{unnecessary\ errors} \qquad (6)$$

are calculated. Based on formula (1), the relative change is given by equation (7), where $y_i^{(ch)}$ denotes the response for respondent $i$ of the ‘change in stocks’ and $\tilde{y}_i^{(ch)}$ denotes the edited value. We define ‘change in stocks’ as the difference between the closing and opening stocks for that period.

$$100 * \frac{\left| \sum_{S_{ED}} w_i (y_i^{(ch)} - \tilde{y}_i^{(ch)}) \right|}{\sum_{S_D} w_i \tilde{y}_i^{(ch)}} \qquad (7)$$

| Gate 1 | Gate 2 | Tolerance1&2 | Gate 3 |
|---|---|---|---|
| 250,300 | 950,1000 | 40 & 50 | 10,20,30,40 |
| 400,500 | 1500,2000 | 40 & 55 | 50,75,100,130 |
| 600 | 2500,3000 | 40 & 60 | 150,175,200 |

Table 1: Parameter values tested

### 4.4 Results

Table 2 displays the savings, missed errors, and relative change for various sets of parameter values for *Gate1, Gate2, Tolerance1, Tolerance2* and *Gate3*. The data used were quarter 3 of 2005, the latest set of data available. Other quarters showed similar results. The table has been sorted by relative change.

| Gate1 | Gate2 | Tolerance1 | Tolerance2 | Gate3 | Savings | Missed Errors | Relative Change |
|---|---|---|---|---|---|---|---|
| 600 | 1000 | 40 | 50 | 10 | 111 | 77 | 0.56 |
| 600 | 1000 | 40 | 50 | 50 | 205 | 171 | 1.00 |
| 600 | 1000 | 40 | 50 | 40 | 192 | 158 | 1.28 |
| 600 | 1000 | 40 | 50 | 20 | 160 | 126 | 1.32 |
| 250 | 1000 | 40 | 50 | 100 | 243 | 209 | 2.96 |
| 300 | 1000 | 40 | 50 | 100 | 243 | 209 | 2.96 |
| 600 | 1000 | 40 | 50 | 100 | 243 | 209 | 2.96 |
| 600 | 950 | 40 | 50 | 200 | 274 | 240 | 3.93 |
| 600 | 1500 | 40 | 50 | 200 | 274 | 240 | 3.93 |
| 600 | 2000 | 40 | 50 | 200 | 274 | 240 | 3.93 |
| 600 | 1000 | 40 | 55 | 200 | 274 | 240 | 3.93 |
| 600 | 1000 | 40 | 60 | 200 | 274 | 240 | 3.93 |

Table 2: Results for quarter 3 of 2005

In Table 2 we observe that the relative change increases with the savings. The only parameter causing changes in savings, missed errors and relative change is *Gate3* in Rule 3. The results suggest that excluding Rules 1&2 from the validation system would not impact on the relative change and savings achieved.

By setting *Gate3* to 10 we achieved sensible relative change for all quarters (except one). However, the volatility across quarters suggests that setting parameters by period would be recommended for seasonal surveys.

The ultimate aim of this table is to provide the customer with a simple tool to decide on the loss of quality he would accept to achieve £*X* savings. Assuming that the customer seeks the smallest loss of quality, the optimum validation procedure would have *Gate1=600, Gate2=1000, Tolerance1=40, Tolerance2=50,* and *Gate3=10*. By using these validation gates, we would achieve an average saving of 111 failures per period, of which 69% were genuine errors, and we would only cause a relative change of 0.56% on the estimate of ‘change in stocks’.

## 5. Conclusions and Further Work

Until now, changes to validation rules have been made in isolation and without consideration of the impact of these changes will have on the quality of the survey output.

In this paper we define a simple decision support tool that quantifies the loss in quality resulting from reductions in validation costs. Customers are supplied with a table that shows savings alongside parameters (for these savings), numbers of errors that will be missed, and impact on final estimates. Sorting this table (by impact within savings) will identify the parameters that cause the minimum impact for the targeted savings.

Further work is dominated by development of a generic programmable solution to the problem considered in this paper. This will be a significant advance for data cleaning in the ONS, and will have positive implications for efficiency and quality throughout the survey process.

Other elements of further work are:
- investigating varying the parameters by domains (e.g. Standard Industrial Classification (SIC), employment sizeband);
- testing data driven methods, e.g. Hidiroglou-Berthelot (1986); and
- continuing work to improve other stages in the data cleaning process, as an alternative to validation.

## 6. References

Granquist, L. and Kovar, J. (1997), “Editing of Survey Data: How Much is Enough?”, *Survey Measurement and Process Quality*, eds. L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, New York: Wiley, pp. 415-435.

Tate, P., Underwood, C., Thomas, P. and Small, C. (2001), “Challenges in Developing and Implementing New Data Editing Methods for Business Surveys”. *Proceedings of Statistics Canada Symposium 2001, “Achieving Data Quality in a Statistical Agency: a Methodological Perspective”.*

Hidiroglou, M.A. and Berthelot, J.M. (1986), “Statistical Editing and Imputation for periodic business surveys”, *Survey Methodology*, June 1986, Vol. 12, N1, pp. 73-83.