Selective and Automatic editing with CADI-applications[5]

Frank van de Pol, Willem Molenaar, Statistics Netherlands

Summary

When designing a CADI-machine for a business survey, the implicit approach with Blaise is to edit every record with the same effort, even though some firms contribute vastly more to publication figures than others . Canadian, Swedish and Dutch research into the effectiveness of this approach has shown that in the cases under investigation at least 50% of the edits have virtually no effect on publication figures. This calls for methods to skip ineffective editing activities.

One way is to select records which have a high risk of containing influential errors, the critical stream, for Blaise-editing. The remaining records, the non-critical stream, may remain unedited. The part of the non-critical stream that has Blaise status 'dirty' or 'suspect' may also be 'cleaned' by a routine for automatic editing.

The art of selective editing is to devise a powerful and yet practical formula to determine the risk that a record contains influential errors. This involves taking account of inclusion probabilities, non-response probabilities, size of the publication cell and, most important of all, a benchmark to determine whether an observed score may be in error, like the cell-mean (or median) for that variable.

Partly drawing on experiences from official statistics in other countries like Canada, the US and Sweden, the principles of automatic editing will be briefly dealt with.

1. Introduction

This paper describes editing business survey data in a CADI (computer assisted data input) situation. Business survey data most often are collected with a survey by mail. The paper forms are returned and the data have to be entered and edited. With the introduction of Blaise-CADI, data entry and data editing have been integrated into one process, thus avoiding unnecessary wait queues due to division of labour. This advancement has a drawback, however, which is that with most Blaise data entry applications a copy of the original, unedited data is no longer saved, thus obstructing research on the effects of editing. For one survey, the Annual Construction Survey (ACS), we changed the production process to obtain unedited data.

Data were entered 'heads down', without changing any entry. Every day the newly entered records were saved and added to the file with unedited data, and their status was set to 'old'. Subject specialists would edit 'old' records only, using Blaise-CADI machines which contain almost 100 messages for errors or suspect conditions. According to these criteria, most records needed correction. In this way two versions of the data were obtained, one with ***un***edited data and one with edited data (euphemistically called: 'clean'). By comparing these two data files one can monitor data editing quality. We used these files to evaluate the effects of data editing.

2. Some edits have more effect than others

Boucher (1991) and Lindell (1994) compared unedited data with clean data and found that, for each variable, 50% - 80% of the edits had virtually no effect on the grand total. We reproduced their analysis for twelve ACS variables, two of which will be shown graphically: total number of employees and total production. Figure 1 shows these figures for three classes of business: contractors in residential and non-residential buildings (top), contractors in civil engineering (soil, roads, bridges; middle) and other building completion work (bottom). The lines go from no editing (on the left) to complete editing (on the right). Edits are sorted from the largest errors (left) to the smallest errors (right).

Figure 1 shows that data editing had little effect on the estimates. The level on the (unedited) left side of each graph is almost the same as the level on the (edited) right side. For only one of the 12 variables we looked into, a strong editing effect was found, namely trade benefits (not in a figure). Editing reduced this amount from about 2% of the production value to less than 1%.

Figure A. Edit effect on number of employees (left) and total production (right)

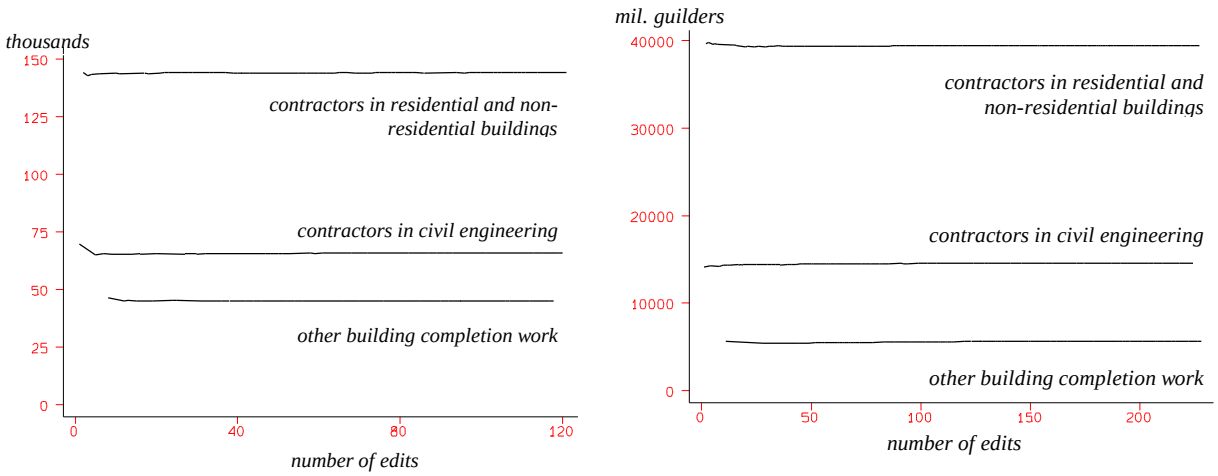


Figure 2. Edit effect on number of employees, vertical axis stretched

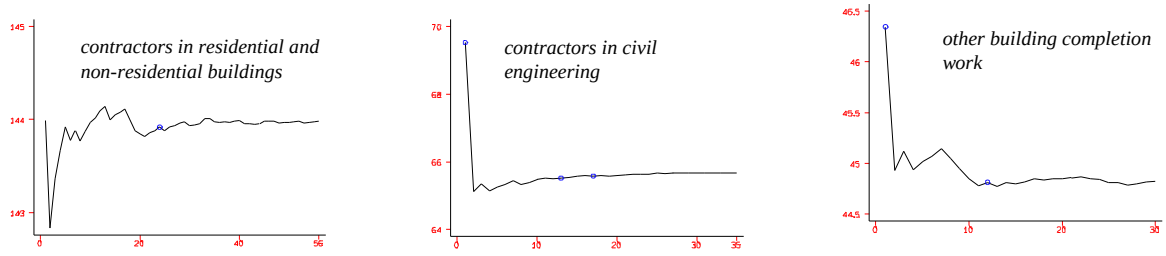
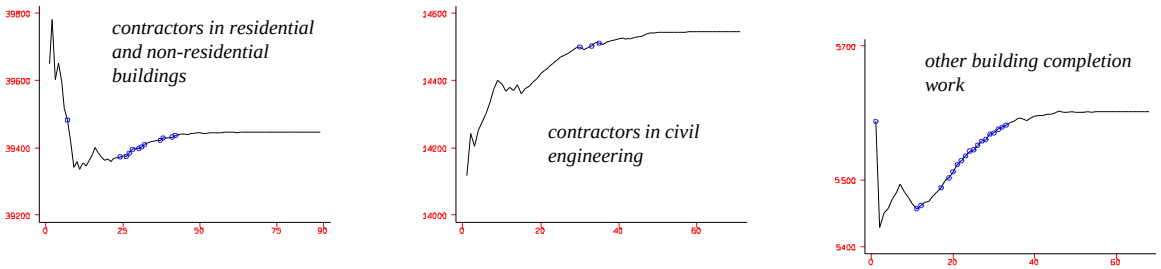


Figure 3. Edit effect on total production, vertical axis stretched



Zooming in to each of these lines (figures 3 and 4) it shows that the small effect that does exist is merely due to the first 10 to 50% of the edits, the largest ones. Like Boucher and Lindell, we find that at least the last 50% of the edits has virtually no effect. This does hold for all the 12 variables that we had at our disposal, even for trade benefits.

Can we know in advance which records hold the more influential errors? Some suggest that only large firms generate influential errors. To test this hypothesis edits on small firms (<10 employees) have been highlighted with a dot in figures 2 and 3, if they occurred in the left half of each line, that is the half of the biggest errors. If small firms would not have to be edited the lines in these figures would not show any dots, but these lines do show dots, sometimes even at the leftmost point, as the largest error. So it shows that not only large firms, but also small firms generate influential errors. Therefore we developed something more sophisticated to pinpoint the records with influential errors.

3. Which forms will have to be edited manually (with Blaise) ?

When a large part of the edits has little or no effect on the publication figures, the question arises: 'How can we avoid the effort of non-effective edits?'. For this aim a 'Safety-Index' is conceived which should predict whether a record needs editing or not. The index was inspired on a paper by Hidiroglou and Berthelot (1986), who worked on a survey with one target variable. The ACS, however, has several target variables.

Traditionally, important tools with ACS editing were ratios of key entries like 'number of employees', 'number of mandays worked', 'production', 'labour costs' and 'trade benefits'. For these ratios upper and lower limits were defined. Moreover, the ACS involves tests on the correctness of the sum of detail-entries. Such a test can also be written as the ratio of the computed sum and the reported total, which has to be 1.

For our safety-index we aim at summarizing these ratios (j = 1, ... J). The index should be sensitive for the effect of an error on publication estimates of totals. Hence it should also take the sample-inclusion probability and the response probability into account. In a previous paper it was argued that the difference of a ratio $r_j = y_{nj} / y_{dj}$ (n for numerator and d for denominator) from its median, $med(r_j)$, can be intepreted as a crude estimate of an error (Van de Pol, 1994). In the present paper we propose to make a symmetrical deviation measure by *dividing* each ratio by its median, $r_j / med(r_j)$, and taking the largest one of $r_j / med(r_j)$ and $med(r_j) / r_j$, that is

$d_j \equiv d(r_j) = \max\{[r_j / med(r_j)] , [med(r_j) / r_j)]\}$ or, equivalently,

$d_j \equiv d(r_j) = \exp\{abs[\log(r_j) - \log(med(r_j))]\}$.

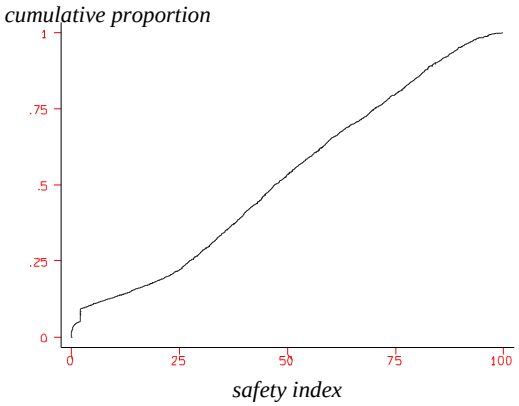
Latouche and Berthelot (1992) proposed other symmetrical measures for the case that more measurements, y_{nj} and y_{dj}, of the same firm are available from a previous time point.

The effect that an error in a ratio has on a publication total (numerator or denominator) depends on the contribution of that firm to the total concerned. Because we have to deal with unedited data, we considered several indicators for this contribution, in order to eliminate erroneous ones. We chose four highly correlated variables, x_1 = 'number of employees', x_2 = 'mandays worked', x_3 = 'net production' and x_4 = 'net production minus production via subcontractors'. All these indicators were transformed to standard deviation 1. Next, to obtain a robust mean of them, the smallest and largest value were eliminated and the average of the two middle values was taken as the contribution of the present firm, denoted by i (importance). If many entries were empty (the importance was zero) the importance was estimated from the firm size class, a 9-category variable which is known from other sources than the ACS.

Taking the product of the inclusion probability and the response probability, π, into account, a 'risk-index' can be formulated as

$$RI = \frac{i}{\pi} \sum_{j=1}^{J} e^{w_j \ln(d_j)} / J ,$$

with w_j a weight for deviation component j. For w_j we initially took the inverse of the standard deviation of $\ln(d_j)$. Weights were trimmed to improve the index's ability to predict all types of errors. The standard deviation of $\ln(d_j)$ can be obtained from last year's clean data. Median ratio's were obtained from unedited data, but can also be taken from last year's data, when most of current year's data is not yet received.



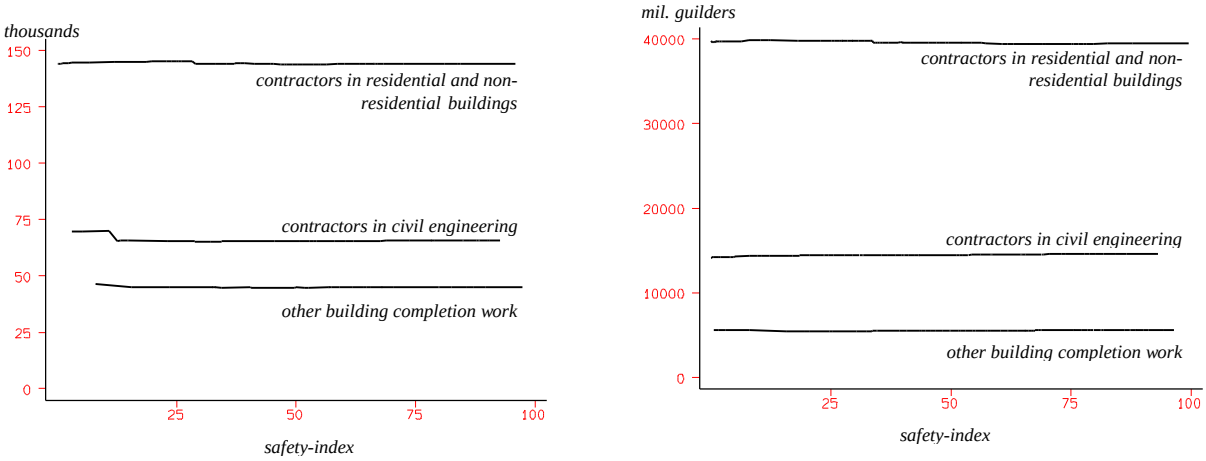
Next, because of peculiarities of Blaise 2, which will be described in section 4, we turned the risk-index into a 'safety-index', with value 0 indicating the highest possible risk, and value 100 the lowest possible risk, SI* = 100 - 100 RI / (1 + RI) . Moreover the index was transformed such that its value approaches a uniform distribution between 0 and 100. This has the advantage that editing all records between a certain value, say z, corresponds to editing approximately z% of the records. The safety-index is

SI = 100 - 100 c RI / (1 + c RI) .

Figure 4. Cumulative distribution of safety index

With $c = 1.6$ we got close to a uniform distribution, as is shown by the cumulative distribution plot in figure 4. With a uniform distribution it should show a straight diagonal line.

Figure 5. Edit effect on number of employees (left) and total production (right); edits



sorted by safety-index

Of course the size of each error is the best risk-index we could think of, but this size is unknown for unedited data. The safety-index however proves a sufficiently effective substitute in predicting which records have to be edited. This can be seen in figure 5, which is a repetition of figure 1, this time not with the number of edits (ordered by error size) on the horizontal axis, but with the safety-index instead. For brevity we do not show the same graphs for the other 10 variables that were inspected. The safety-index is as effective in predicting errors for those variables.

An overview of methods for rationalizing the editing of survey data is given by Granquist (1994).

4. How to give shape to selective editing with Blaise

Suppose we could assign the value of the safety-index to a key, with which Blaise gives access to records in order of the size of the key. Then subject specialists could edit records in order of error risk by requesting a record without any key-specification. The record with the highest error risk (lowest safety-index) would pop up first, the record with the one-but-highest error risk would be edited next, etc.

In Blaise version 2, records can be accessed with two keys, the primary 'key' and the 'selector' key(s). The primary key has to be unique and the safety-index is not, so this key cannot be used. Fortunately the selector key does not have to be unique, so we can use the selector key. When using the CADI machine, the subject specialist could use the 'locate forms' option. If he does not specify a selector value, Blaise will show the record with the lowest selector value. Data access in reverse order is not possible. This is why we formulated the risk-index as a safety-index: records with the lowest safety-index will show up first.

In order to access data in order of error risk, we must have the safety index computed first. When entering data in a CADI-machine, subject specialists should be able to tell the Blaise-program that all data of a record have been entered, and that the safety-index has to be computed. Ideally, we would like to have a function key for this (or a graphical button on the screen). Such functionality is not available, though. In Blaise 2, we used the <shift-F3> key instead, which invokes the evaluation of all error checks.

The computation of the safety-index has been linked to a dummy-variable, whose default-value is true. At the same time when Blaise error-messages are generated, the safety-index pops up on the screen. Records with a high safety-value, for instance larger than 50, should not be edited, even when the conventional Blaise-error messages tell that some details are

wrong. Correcting these details would not have any effect on publication figures, so it would be inefficient to correct them. After all, correction of 1000 records more would take another man-year. Only when the record is not safe, that is the index is lower than 50, the subject-specialist should edit the record.

Of course data entry can be carried out by people who are less highly qualified than subject specialist. In that case the data typist will evaluate the record with <shift-F3> and then save it, without editing. The subject specialist can access the most risky record later, by using the 'locate forms' entry of the main menu. As explained before, if he does not specify a selector key, the most risky record will pop up first.

At the Netherlands ACS forms used to be edited in order of day of reception. The most recently received record was edited first, because in case a respondent has to be called for clarification the respondent will bring back to memory more easily which figures he entered. Moreover respondents should not be demotivated by asking them questions many months after they returned the questionnaire. Therefore we presently try to take the best of both approaches. Recently entered forms are listed on paper in order of their safety-index and subject specialists will select high-risk records from that list. Then they will select the paper form from the pile of forms, which are sorted on day of reception, look up the firm-identification number, and call for the form on their CADI-machine via 'process old form'.

5. What to do with the forms that seem not to need editing

When a risk-index is used for prioritizing edits, some records will remain unedited, though they contain some slight inconsistencies. Moreover, in case of a long questionnaire, not all variables will be incorporated in the risk-index. One way to deal with remaining errors is to apply automatic editing. Automatic editing, not to be confused with computer assisted editing as is offered by Blaise, removes inconsistencies from the data. Fellegi and Holt (1976) developed a method to trace which field(s) cause a series of error messages. Their criterion is, in short, that as few fields as possible should be altered to remove the error messages. Once the erroneous fields have been designated, they can be replaced by imputed values.

In practice, efficient algorithms for this approach are hard to devise. A Windows- program which uses the Fellegi-Holt algorithm, like Lince (Informatica Comunidad de Madrid SA, 1993), will not handle much more than 30 error messages. An alternative algorithm, as is implemented in GEIS (Kovar and Whitridge, 1990), will only be fast for small datasets. If edit checks are restricted to ratios of two variables at a time, a simple and fast algorithm emerges. This algorithm has been implemented in SPEER (Winkler and Draper, 1994). At Statistics Netherlands we presently do research into these methods, which will hopefully become of use for all Blaise-users.

As these fully automated methods should not be used with very many error messages, we looked for simple additional methods. One possibility is to apply 'deterministic' corrections. Deterministic corrections can be applied in those cases where always the same correction is to be made in specific, well defined circumstances. In the ACS questionnaire for instance, some figures have to be subtracted from others to obtain their difference. The minus sign is printed in the questionaire, but some respondents will, erroneously, reproduce it in the answer box. Often these minus signs are, again erroneously, typed in. This will generate Blaise-error messages, which can easily be corrected automatically by testing whether deleting the minus sign will remove the error message.

Deterministic automatic corrections can be invoked in Blaise in the same way as the safety-index is computed. A dummy variable called 'auto' is specified, which has default value 'off'. The subject specialist can turn this variable 'on', and press <shift F3> to have automatic corrections.

Another deterministic correction was developed for those forms where part of the data is missing. In the ACS especially questions on production details are numerous. Some respondents only present figures on their production total, and skip the details. Verboon (1995) describes how automatic imputations perform, using cell means and last year's firm-specific data.

References

Boucher, L. (1991): *Micro-editing for the Annual Survey of Manufactures: What is the value added?* Proceedings of the U.S. Bureau of the Census 1991 Annual Research Conference. pp. 765-781

Fellegi, I.P., and D. Holt (1976): *A systematic approach to automatic edit and imputation*. **Journal of the American Statistical Association** 71, pp. 17-35

Granquist, L. (1994): *Macro-editing - A review of methods for rationalizing the editing of survey data*. In: United Nations Statistical Commission and Economic Commission for Europe: Statistical Data Editing, Vol. 1, Methods and Techniques (United Nations, Geneva)

Kovar, , J.G., and P. Whitridge (1990): Generalized edit and imputation system: overview and applications. **Revista Brasileira Estatística** 51, pp. 85-100

Hidiroglou, M.A., and J.-M. Berthelot (1986): *Statistical editing and imputation for periodic business surveys.* **Survey Methodology** 12, pp. 73-83

Informatica Comunidad de Madrid SA (1993): *Lince, Sistema de validación e imputation automatica de datos estadisticos*; manual de usario (ICM, Madrid)

Latouche, M., and J.-M. Berthelot (1992): *Use of a score function to prioritize and limit recontacts in business surveys*. **Journal of Official Statistics** 8, pp. 389-400

Lindell, K. (1994): *Evaluation of the editing process of the salary statistics for employees in country councils. paper presented at the UN congress on data editing in Cork, Ireland*. To appear in 'Statistical Data Editing, vol. 2' (UN Statistical Commission and Economic Commision for Europe, Geneva)

Van de Pol, F. (1994): *Selective editing in the Netherlands Annual Construction Survey. paper presented at the UN congress on data editing in Cork, Ireland.* To appear in 'Statistical Data Editing, vol. 2' (UN Statistical Commission and Economic Commision for Europe, Geneva)

Verboon (1995): *Editing subdivisions in the Annual Construction Survey*. Paper presented at the Bristol conference on Survey Measurement and Process Quality (Statistics Netherlands, Department of Statistical Methods)

Winkler, W., and L.R. Draper (1994): *Application of the SPEER edit system*. Research paper (US Bureau of the Census, Washington)