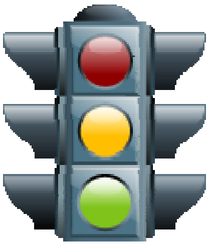
Statistiska centralbyrån
Statistics Sweden

A General Methodology for Selective Data Editing

Version 1.0



This report covers the methodology for selective editing which we propose for Statistics Sweden. The paper demonstrates the functionality we wish to find in a general box of tools for editing. It has been prepared by Anders Norberg, Chandra Adolfsson, Gunnar Arvidson, Peter Gidlund and Lennart Nordberg

Contents

1 Introduction

For various reasons measurement errors appear in survey data. Errors due to imperfect measurement tools; questionnaires, instructions, service to respondents etc., are clearly a task for the statistical bureau to handle. The identified problems for the respondents to deliver correct data must be considered as shortages of the measurement tools. Editing should be seen as the statistical quality control of the measurement process. A systematic improvement of the survey process includes collection of process data – analysis – change (improvement) of process – collection of process data, again and again (the PCA-cycle).

There is an end to how well questionnaires, instructions etc. can be designed. Some statistical information, requested by the customer of statistics, may not correspond to data that are available for the respondent in their accounting systems. Careless mistakes or deliberate fakes by the respondents should be prevented by follow-up to maintain confidence for the statistical bureau.

The role of editing is to:

- Find errors (Efficient controls)
- Identify sources of errors (Process data)
- Analyze process data – communicate with cognitive specialist
- Contribute to quality declaration
- Adjust significant errors (Adjust/correct)

Traditionally, the goal of editing has been to discover all errors in data and to “correct” them. We now realise that editing can result in errors (Granquist and Kovar(1997)). But it is not obvious that selective data editing is the way to meet this role, nor is it obvious that it does not. Selective editing meets the need for efficient search of significant errors. If and when the survey process is made cheaper, in many respects, resources can be laid on identifying so called “inliers”, i.e. identifying causes of mass-misunderstanding of questions resulting in small individual errors by respondents but systematic big errors in statistics.

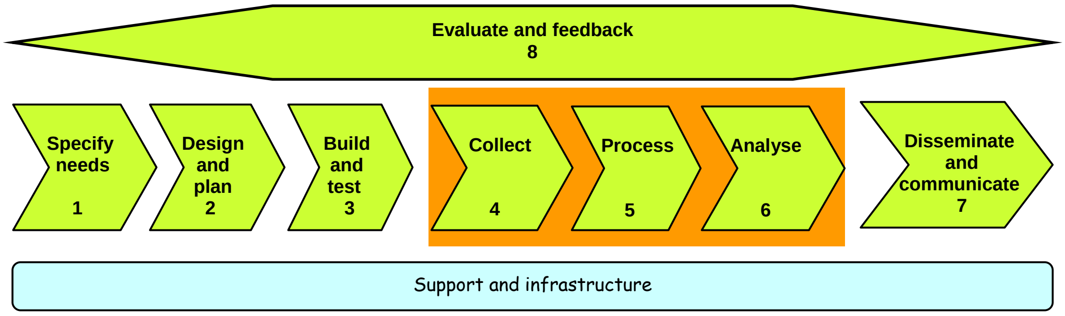
In chapter 1 different types of editing and some specific aspects of editing are briefly discussed.

Chapter 2 summarises case studies reviewing the editing methods used in nine large statistical surveys 2006. Chapter 3 gives an overview of the conditions, general processes and underlying principles for selective editing. In chapter 4 we present the general editing method to be implemented at Statistics Sweden. Chapter 4 plus chapter 5 are fundamental parts of the documented requirements for developing IT tools for data editing.

1.1 Different types of editing[1]

The production process of statistics is described at Statistics Sweden in terms of the following seven general sub-processes and two overall processes.

The statistics production process



The editing of data in statistical surveys takes place at several stages of the production process and communicates with several of the sub-processes referred to above. This applies to:

- **Respondent's editing**: Performed by the respondent, above all when filling in electronic questionnaires or jointly by the respondent and interviewer in interview studies.

- **Manual data review:** A manual process that takes place prior to data entry. This process can be necessary for re-profiling of survey units and classifying unit non-response versus over-coverage. Manual review may be justifiable as production editing in some cases, but a general recommendation is that its scope should be minimized. One reason is lack of cost effectiveness; another is that this review technique does not generate unedited or process data that makes it possible to analyse the data collection or the editing process. The first checking of incoming data files from respondents is also regarded as manual data review.

- **Data-registration editing:** Principally, the handling of fatal errors[2] identified by a data-registration program. The most important technique for the registration of data from paper forms is now scanning, where interpretation errors may arise.

- **Production editing** or **batch editing:** Follow-up[3] of the variables and units flagged by the editing program. Often, the program is run on batches of records, one at a time, but production editing by record can also take place.

- **Output editing:** Editing when all the material has been collected and converted into output tables and the like in order to check that no major errors have been made during the preceding processes.

In turn, the editing process can be divided into three sub-processes, which are fairly independent of each other in terms of the tools employed.

o Check / select data items and units / set error flags

[1] Links to glossary of terms on statistical data editing:
http://www.unece.org/stats/publications/editingglossary.pdf
http://stats.oecd.org/glossary/

[2] Obvious or evident errors, for example non-valid and missing values.

[3] The manual work performed, including re-contacts with respondents, which ends up with acceptance or change of data values.

- Follow-up
  - Use other data
  - Analyse
  - Re-contact respondents
  - Accept or alter
- Collect and record process data

This means that, from an IT-architectural perspective, it is of benefit if tools can be created separately for the three parts. Nevertheless, they need to be such that interfaces are clear, so that no problems arise when anyone of the three tools is revised or rebuilt.

Editing is a resource-demanding process, in particular for business surveys. In a study at Statistics Sweden of the total costs of 62 statistical surveys, a third of resources were found to be used for editing, see SCB (2005). In an earlier study, from around 1980, an estimated 40 percent of the total cost was found to be used for editing. Thus, we can note a cost reduction over 25 years. The proportion of resources invested in editing is larger for annual and periodic surveys than for monthly and quarterly surveys. The short-period surveys are subjected to relatively less production editing.

Table 1.1 Average proportions of the costs of sub-processes in the total cost of statistical surveys during 2003 and 2004.

Process	Proportion of total cost		
	All surveys	Short-period	Annual surveys and periodic
Respondent service	3.3	3.3	3.4
Manual data review	4.4	3.9	5.1
Data-registration editing	5.6	5.1	6.5
Production editing	15.3	12.7	18.9
Output editing	3.9	3.4	4.8
Total editing cost	32.6	28.3	38.6

This method report is largely concerned with production editing.

1.2 Goal of editing

A new role of editing is slowly being implemented at statistical institutes. It focus the editing on identifying and collecting process data on errors, problem areas and error causes in the measurement process to provide a basis for a continuous improvement of the process and the whole survey vehicle in general. The old paradigm – the more and tighter the edit checks and re-contacts, the better the quality – should be replaced (Granquist 1997).

The entire set of the query edit checks should be designed to focus on errors influencing the estimates, and be targeted on existing error types. The effects of the edit checks should be continuously evaluated by analysis of performance measures and other diagnostics, which the process should be designed to produce, i.e. process data is also used to improve on the editing process itself. Editing staff debriefings have been implemented, see Hartwig(2009).

When editing primarily is quality control of the measurement process, it is still needed to contribute to quality declaration and to adjust *(change/correct)* significant errors in the current survey round to avoid bias.

1.3 Goal of selective editing

The goal of selective editing is to minimize the amount of follow-up activities while maintaining an acceptable quality of the statistical output. We imagine the bias that is still in the estimates because not all data have been intensively edited[4]. If this so called pseudo-bias can be estimated, the acceptable editing process should generate a bias less than 20 percent relative to other survey errors, as a rule of thumb. This relative pseudo-bias can be allowed to vary for different parts of the output. We want to obtain high quality in output where it is most needed.

A reduction of follow-up activities liberates resources of which some are proposed to be used for reducing other sources of error in the statistical process. Optimally this leads not only to a cost reduction but also to an improved overall quality.

1.3.1 Exhaustion

To counteract the exhaustion effects of over-frequent re-contacts, which have a detrimental impact on response behaviour, the editing method should take into account the actions (acceptance or amendment) that were the result of the preceding re-contacts, i.e. make use of all the relevant information at hand.

1.3.2 New respondents

Especially in short-period surveys, respondents must recurrently supply data for certain surveys. Smaller enterprises end up in the sample, in accordance with the SAMU[5] sampling system or similar systems.

How shall Statistics Sweden treat new respondents in the editing process in comparison with respondents who have supplied data for nearly 60 months and are still error-flagged during the editing? It is often desirable in editing to have a smaller acceptance region for new respondents so as to quickly be able to resolve any problems on how the data should be supplied.

1.3.3 Confidence in National Statistical Institutes

It can have a negative effect on respondents and personnel if too many erroneous items of information pass through without any reaction. A reputation that *the Statistical Institute will accept whatever data is supplied* hardly promotes the will to supply high quality data. In order to maintain confidence it is desirable to identify and re-contact respondents who have supplied data that are strongly suspected to be erroneous. The use of efficient editing in order to maintain low costs for production may conflict with this quality. This is especially so if there are data with minor impacts on the output.

Suppose that a respondent supplies exactly the same information on several survey occasions in a survey where this may not be reasonable. Then this would be an indication of erroneous data. Edits for such errors, i.e. finding a type of inliers, can be difficult to construct. In case of small numbers, such as the number of employees reported sick for a small business, it is not implausible that a repeated value is reported month after month, whereas this is implausible for large enterprises with thousands of employees. The number of survey rounds where repeated values would be acceptable varies. An issue is how to explore historical data to find out. The general concept of selective editing, the method proposed in this paper, is not suitable for detection of inliers.

[4] In this line of thought we do not consider the potential introduction of errors in the follow-up process.

[5] SAMU is a system for coordinated and rotated sampling of businesses at Statistics Sweden

Statistics Sweden needs to do some general work on the issue of passivity edits.

Item non-response can have its roots in bad measurement tools but also in some respondents` easy way to keep away from respondents` duty. Item non-response is one type of fatal error, easy to identify. Either we make a re-contact for all respondents with non-response even if this is not justified from a short-term resource perspective or we treat them in selective editing with full suspicion and an arbitrary estimated impact on the statistics. If, in short term statistics, it is a repeated procedure by a respondent not to deliver information, the impact on statistics should be estimated not only for one survey occasion but for a series for which the respondent is supposed to be in the sample.

Pre-printed questionnaires is a case where the respondents can see if Statistics Sweden itself has generated erroneous data and accepted them. The errors can be introduced in the data registration process, most often with scanning.

Data from electronic questionnaires, where edits are implemented and processed by the respondents, should only sparsely be flagged for follow-up in the production editing process.

2 Case studies on editing

2.1 The studies

Editing has been a neglected process when it comes to introducing efficient and practical methods in statistical surveys. A project has carried out case studies for seven major statistical surveys, as a first step to produce general tools for editing. For two other surveys case studies more or less failed due to lack of unedited data.

From the 2006 edition of Structural Business Statistics (SBS) *{Företagens ekonomi (FEK)}* a new editing system will be used with focus on output regardless of source of data. Selective editing already exists in a part of the survey, but the method should be evaluated and implemented for other parts. A possible cost reduction has not been estimated.

Foreign trade – exports and imports of goods (IntraStat) *{Utrikeshandel med varor}* has several supplementary editing systems; non-valid entries, enterprise totals, price per quantity-editing, and a few to detect non-response. In the enterprise totals editing , about one percent of the incoming enterprises were manually flagged judgmentally. The hit rate is low. In the study a selective editing has been created that generates a ranked error list including about one percent of the respondents.

Price per quantity editing has applied score functions since 2004. The functions are built by separate indicators for suspicions and potential impacts for each record (the in- or outflow of a commodity to a country by an enterprise). Only 0,3 percent of the records are error flagged and the hit rate is about 60 percent. Minor improvements can be made by computing global scores for respondents rather than using scores for records only.

Business Activity Indicators (BAI) *{Kortperiodisk industristatistik (KortInd)}*. A traditional survey with traditional editing. The hit rate is very low. The editing staff look at and then accept most part of the error flagged units manually, they find most of their support for their actions in historical data. The number of editing controls can be reduced. Today 16 percent of the incoming units are error flagged and this fraction can be reduced to 8 percent using selective editing. The amount of the reduction of editing workload could not be estimated, however.

Wage and salary structures in the private sector *{Lönestrukturstatistik, privat sektor (SLP)}.* A two-stage survey of businesses and employees with several variables and diverse output of statistics. The fraction of employed with at least one fatal- or suspected error is 31 percent. This results in 89 percent of the businesses being error flagged for at least one employee. About half of the error flagged businesses are re-contacted. If a small decrease in quality is accepted, selective editing can reduce the portion of followed-up businesses with almost 25 percent.

Short-term statistics, wages and salaries in the private sector *{Konjunkturstatistik, löner för privat sektor (KLP)}*. The fraction of incoming units with at least one fatal- or suspected error is about 60-65 percent. Less than half of the respondents are re-contacted, most part of the errors are edited manually by the editing staff who are able to edit the errors with help from other sources of information. The number of errors that need manual attention can be reduced by approximately 20-40 percent using selective editing, but the amount of follow-ups that can be reduced has not been estimated.

Short-term employment, private sector and Job openings and unmet labour demand, private sector *{Kortperiodisk sysselsättningsstatistik, privat sektor (KSP) och Konjunkturstatistik över vakanser, privat sektor (KVP)}.* The current editing method being used for *Job openings and unmet labor demand* is inefficient, the hit rate is very low. Job openings and vacancies are variables that seldom take on a non-zero value and are therefore hard to handle in the statistical process and certainly in the editing process. In the surveys about 14-16 percent of the incoming units contain at least one fatal or suspected error. Currently all of these errors are handled manually by the editing staff.

The result of the case study regarding Short-term employment, private sector revealed that selective editing with score functions can reduce the number of errors that require manual attention by about 60 percent. A condition that must be fulfilled is that all the obvious measurement issues must be taken care of.

Producer and import price index (PPI) *{Prisindex i producent- och importled (PPI)}* The prioritization of the error flagged units is judgmental. Many seasonal commodities are flagged but accepted without contact with the respondent. Selective editing including time-series analysis was tested. The quality of the editing would be improved with selective editing, but a possible cost reduction has not been estimated.

Rents for dwellings *{Hyror i bostäder (HiB)}* Re-contact with the respondents is taken for about 30 percent of the dwellings. The case study could not successfully use unedited data from a back-up file because the material was incomplete. However the overall judgment of the study is that selective editing would work well in the survey. Editing should be focused on derived net rents instead of uncorrected rents.

The only editing of *Swedish national and international road goods transport {Inrikes och utrikes trafik med svenska lastbilar (SLIT)}* is manual pre-editing. The manual pre-editing includes coding of commodity groups and manual imputation. This means that there are no unedited data available in the survey and no evaluation of selective editing could be performed.

On the basis of the nine case studies the project team has proposed good editing methods with well documented methodology.

2.2 Varying conditions for editing

Statistical surveys differ in many respects, which may be of significance for how editing is performed.

2.2.1 Periodicity

We distinguish between three types of surveys in terms of how often they are conducted, which entails that we have variable access to earlier data for the construction of edits.

A. One-off surveys and also surveys that are conducted so seldom that there is no information from earlier observations that would provide a basis for finding reasonable edits. Here, the role of editing is to find significant errors rather than to contribute to survey improvement for the future.

B. Annual surveys and also intermittent surveys that, by contrast to A, have useful data from previous surveys rounds.

C. Monthly and quarterly surveys with access to extensive amounts of time series data.

At the level of the observed units, even in a monthly survey, there are units that are new in an annually updated sample. This is especially so for rotated samples. Units that are new lack earlier data for the unit itself and we can not produce as good predicted values as for units that have been observed for a long time series.

2.2.2 Survey design

We can distinguish between sample surveys and censuses. In the case of samples, weighting is always involved, which means that the units in the observation register have different impacts on the outcome. This must be considered during editing. The sampling method, whether it is stratified SRS or sampling with un-equal probabilities is of little concern for editing. Strata can be used as homogenous groups in the edits and in estimation of good predicted values which are needed.

A most significant aspect of design is whether we have a one-stage or a multi-stage sample. Efficient editing, focusing on minimising re-contacts with respondents, is much more complex for multistage samples, especially when the number of observed, secondary selected, units vary between primary selected units.

An example of a one-stage sampling design is found in the BAI survey, where the respondent for a Kind-of-Activity Unit (KAU[6]) makes monthly deliveries of data. Another example is found in the Short-term employment statistics, private sector *[KSP/KVP]* to which a Local unit (LU[7]) supplies information on the number of employees.

An example of a multi-stage sampling design is found in Foreign trade – exports and imports of goods (IntraStat) where the provider for a LeU[8] supplies information on one or more items of goods traded with other countries within the European Union.[9] The respondents to Wage and salary

[6] Kind of activity unit, *Verksamhetsenhet (VE) in Swedish*

[7] Local unit, *Arbetsställeenhet (AE) in Swedish*

[8] Legal unit, *Juridisk enhet (JE) in Swedish*

[9] A LeU may have one or more respondents, identified in IntraStat by the identifier TillNr.

structures in the private sector supply one record for each employee in their enterprise, which can be up to thousands of units.

2.2.3 Respondents

In principle, type of respondent/unit – individuals, enterprises, products, etc. – have no significance in terms of editing. Nevertheless, it is a fact that business populations generally show a much more skewed distribution on economic and other quantity variables than data about individuals. Surveys involving data about individuals and attitudinal questions cannot, for practical reasons, be edited retrospectively by means of re-contact.

2.2.4 Variables

From an editing perspective, various types of variables must be handled in different ways. Unit identification variables must be correct (at least have valid values); otherwise, there will be technical problems, e.g. in the matching of data. In the case of classification variables, there is a limited range of values, which could lead to invalid values for unedited data . These must always be remedied, possibly by imputation. In the case of quantitative variables, large deviation errors must be identified in the editing process. Smaller errors in the data can often be accepted without the statistical output in general suffering from any appreciable lack of quality.

In some surveys, data are collected on several variables that are not reported individually in the statistics, but rather as part of a derived variable. Here, during editing, analysis should be performed of the effects of suspected errors in the derived variable, although suspicion is often calculated for the original variables.

2.2.5 Output

A survey may have anything from a few clearly defined users and limited output to a general (public) use and extensive statistical reporting. It can be natural to focus the editing process on impacts within the principal reporting.

2.2.6 Registers

We distinguish between:

1. Observation registers for statistical surveys
2. Administrative registers

When administrative data are used, it is the maintainer of the register who primarily performs the editing; and the remaining editing performed by Statistics Sweden is principally output editing.

In SCB (2002), the CBM-handbook “Guide to granskning” (pp. 30 ff.), it is stipulated that the purpose of such editing is to identify errors, sources of error, and to provide a basis for quality assurance of the register for statistical use. There it is also emphasized that in a register with many variables, it is not reasonable to check all the variables. Only those of importance for official statistics and for key areas of use need to be checked.

As well as output editing, *reconciliations* must be performed against other statistics at aggregated level, in order to uncover errors in the data, and also to describe differences in sets of statistics to the user.

2.2.7 Data

Data from previous rounds of statistical production are needed to set checks with effective threshold values to trace major deviations in observed values in the current study. A precondition for being able to introduce and also adjust already established methods and parameters for effective editing is that unedited data are available from previous survey rounds. Also, editing codes that show which checking rules have generated error signals from unedited data are used in such analysis.

We may have:

- data on units collected so far in the current survey round,
- data for the same units on the latest occasion,
- data for the same units on many previous occasions,
- data for another sample of observations,
- some registry data (usually), e.g. from a sampling frame.

We can choose whether or not to utilize imputed values.

Data can be used in a:

- cross-sectional analysis
- time-series analysis

Time-series analysis means that more information can be extracted from the data, since time also becomes a variable. Trends, seasonal patterns etc. can be estimated and used in forecasts for the period in question.

3 Selective data editing – basic concepts

3.1 Errors and edits

The traditional approach to data editing, still in practice in many surveys at Statistics Sweden and elsewhere, takes an “accountant’s view” of editing. According to this approach, one should strive to detect and correct every possible error, preferably by re-contacts or, if necessary, by imputation. This often leads to very tedious and expensive editing efforts with no clear strategy. It is quite likely that a large number of errors with little impact on final estimates are followed-up and changed while a few large errors may still slip through the system and ruin the quality of the statistical estimates. Traditional editing often leads to a heavy workload for the editing staff and for the respondents.

Modern approaches to data editing emphasize the search for “significant” errors, accepting that final data sets do contain a number of errors with no noticeable effect on the statistical estimates. Automatic imputation rules can be added to smooth the data and eliminate item non-response and inconsistencies of no statistical importance. The term significance editing has been used to describe this type of editing, see e.g. Latouche and Berthelot (1992), Lawrence and McDavitt (1994), Lawrence and McKenzie (2000) and Farwell and Raine (2000). A more commonly used term is **selective editing**.

Data errors are often classified into two main categories. Errors such as inconsistent responses, invalid entries and item non-response, are characterized as **fatal errors,** also called **non-statistical errors**[10].

Suspected errors, also called **statistical errors**[11] are caused by data values that deviate strongly from other records, from information supplied on previous occasions or from other relevant a priori information.

The search for errors are traditionally done by edit checks, **fatal edits (hard edits)** for fatal errors and **query edits (soft edits)** for suspected errors.

[10] *Uppenbara fel* in Swedish

[11] *Misstänkta fel* in Swedish

Definition 3.1:

A **query edit** is the combination of:

- a test variable,
- a grouping of data into homogenous groups,
- acceptance regions for each group to which the observed value of the test variable is compared.

Henceforth we will assume that there are *n* **primary sampling units,** whether selected by probability sampling or on a take all basis, and within each primary sampling unit k there may be n_k **secondary observed units.** Under one-stage sampling the primary and the secondary unit coincide and $n_k = 1$ for all *k*.

Before proceeding we need some notation.

Definition 3.2:

Let for primary sampling unit *k* (k = 1, … ,n) , secondary / observed unit *l*, (l = 1, … , n_k) and the *j*:th variable:

$y_{j,k,l}$ = unedited (raw) value for variable Y_j (j = 1, … , J)

$t_{r,k,l}$ = derived **test variable** as a function of unedited values $\left(y_{1,k,l}, y_{2,k,l} \ldots . y_{J,k,l}\right)$ and other variables, for example register data or data from earlier production rounds, (r = 1,2,….,R)

Hence a test variable is an arithmetic expression, based on the collected variables and usually also data from earlier production rounds for comparison. A test variable may be a simple difference between two variables or a complex expression involving a comparison of collected values with forecasted values from estimated time-series models of historical data.

The test variables are used to form the fatal and the query edits. The relation between the test variables and the unedited data variables is sometimes one-to-one in practice, but in general terms the relation is many-to-many. Hence as well as a test variable may be built on several y-variables, as seen from its definition, an unedited data item $y_{j,k,l}$ may be involved in several test variables.

A fatal edit searches for item non-response, invalid entries (e.g. invalid codes for occupation, type of industry etc.) or inconsistent data.

The following balance edit is an example of a fatal edit for inconsistencies.

If the sum of the variables *a*, *b* and *c* should equal *d* by definition then the balance edit can be based on the test variable:

$t_{k,l} = d_{k,l} - a_{k,l} - b_{k,l} - c_{k,l}$ and the fatal edit in this case would be $t_{r,k,l} = 0$ for any observation (*k*,*l*).

If this edit fails, i.e. if $t_{k,l} \neq 0$, then all data items involved (*a, b, c* and *d*) will be suspected.

A typical query edit takes the form $\tilde{t}^{(L)}_{r.k,l} < t_{r,k,l} < \tilde{t}^{(U)}_{r,k,l}$ where the lower and upper limits, provided by the user, define the acceptance region. The choice of the groups and limits can have a critical effect on the efficiency of the editing procedure. We will return to this issue later.

3.2 Suspicion

The limits of the query edit $\tilde{t}^{(L)}_{r.k,l} < t_{r,k,l} < \tilde{t}^{(U)}_{r,k,l}$ define an acceptance interval $(\tilde{t}^{(L)}_{r,k,l}, \tilde{t}^{(U)}_{r,k,l})$. If the test variable falls outside this interval then the data items $y_{j,k,l}$ that are involved in $t_{r,k,l}$ are all flagged for suspicion. If the test variable falls inside the acceptance interval then this particular query does not cause any suspicion.

One can argue that the use of a dichotomized measure of suspicion (0/1) is a way to destroy information. If $t_{r,k,l}$ for case (*k*,*l*) deviates more from the acceptance limits than $t_{r,k',l'}$ for case (*k´*,*l´*), then there would be reason to assign a higher degree of suspicion to case (*k*,*l*) than to case (*k´*,*l´*). Later in this report we will introduce a continuous measure of suspicion. But for now we limit the discussion to the special case of a dichotomized measure $Susp_{j,k,l}$.

Definition 3.3:

A) If a fatal edit fails then all data items $y_{j,k,l}$ involved in the test variable $t_{r,k,l}$ are assigned a measure of suspicion $Susp_{j,k,l}$ = 1.

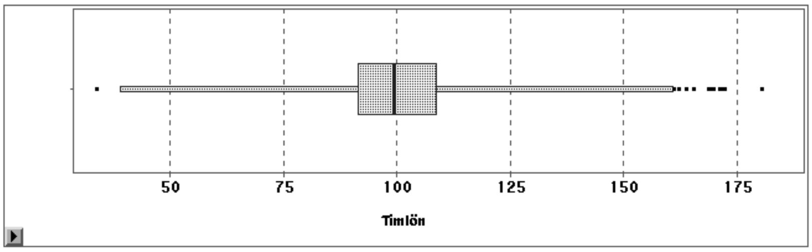
B) If a test variable $t_{r,k,l}$ of a query edit falls outside its acceptance interval then all data items $y_{j,k,l}$ involved in $t_{r,k,l}$ are assigned a measure of suspicion $Susp_{j,k,l}$ = 1.

C) Data items that are not involved in any failed fatal or query edit are all assigned a measure of suspicion $Susp_{j,k,l}$ = 0.

Example: Finding acceptance limits for the wage statistics[12]

One of the variables to be followed-up in a wage survey is the hourly wage. The acceptance limits are most easily determined using a box plot. Here, the length of the "whiskers" are 3 times that of the interquartile range.[13] A feasible proposal is that the acceptance limits are set at 40 – 160 SEK per hour.

Figure 3.1. Hourly wage in the entire population.



If excessively wide acceptance boundaries are applied across the entire data set, there is a risk that errors of importance to the statistical estimates are not identified. In order to identify suspected data values more efficiently, it is essential that acceptance limits take account of large differences between

[12] CBM 2002:1

[13] The whiskers are the narrow boxes on each side of the box, which are delimited by the quartiles. The whiskers stretch to the final observation that lies within k times the interquartile range from the nearest quartile. In the example, *k* = 3.

homogeneous groups, such as various sectors of industry and groups of varying size.

Example: Finding acceptance limits for the wage statistics (cont.)

For most sectors of employment, the acceptance limits 40 – 160 are probably reasonable initial values, although this may not apply in the case of sector 7 as seen in figure 3.2. Further breakdown of the sectors, e.g. at two-digit level, may provide for better adapted acceptance limits. One should make use of subject matter knowledge in combination with the rules given in section 4.2.1 of SCB (2002).

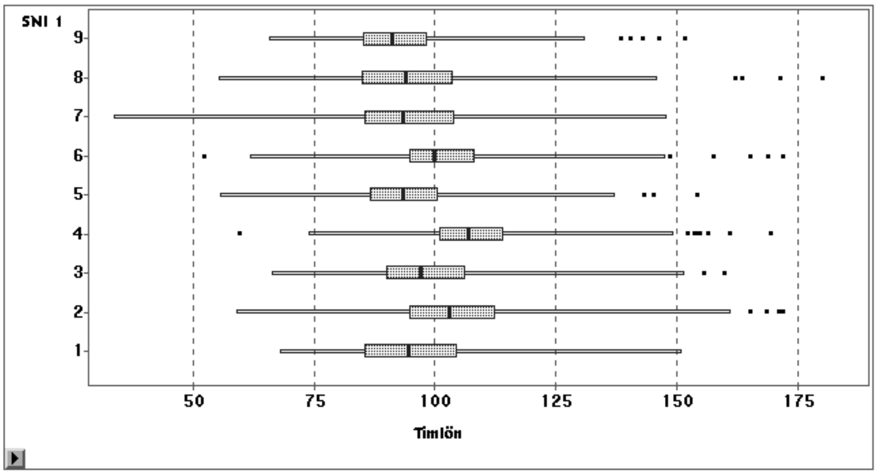
The following edits may be reasonable[14]:

if (Industry in (’1’ ,’3’, ‘4’, ‘5’, ‘6’, ‘8’, ‘9’) and (Wage < 65 or Wage > 140) then Flag = ’1’

if Industry = ’2’ and (Wage < 40 or Wage > 160) then Flag = ’2’

if Industry = ’7’ and (Wage < 35 or Wage > 150) then Flag = ’3’

Figure 3.2. Hourly wage distributed by SNI code at one-digit level[15].



3.3 Impact

In the traditional approach to data editing, all suspected data items –whether from fatal or query edits - call for manual intervention, perhaps by a re-contact with the respondent in order to check all suspected data items. However, a suspected error is ignored in selective editing if its **impact** on the statistical estimates is judged to be unimportant.

For brevity and ease of notation we will in the rest of this chapter assume that the data are collected under a one-stage sampling design. Hence the secondary level l can be ignored.

[14] Flag = ’1’, ’2’ etc. are proposed for error codes in this example

[15] The SNI code equals the EU industrial classification NACE rev.1.1

Definition 3.4.

A): Let $y_{j,k}$ be the original (unedited) value of variable y_j for observational unit k.

B): Let $y_{j,k}^{(e)}$ be the edited value of variable y_j for observational unit *k*, for data that are not followed-up, let $y_{j,k,l}^{(e)} = y_{j,k,l}$

C): Let w_k be the sampling weight for observational unit *k*.

D): Let $\hat{T}_j^{(e)} = \sum_k w_k y_{j,k}^{(e)}$ be an estimator from the sample, based on final, edited data $\{y_{j,k}^{(e)}\}$ and let $SE(\hat{T}_j^{(e)})$ be the standard error of $\hat{T}_j^{(e)}$.

E): Let $\hat{T}_j = \sum_k w_k y_{j,k}$ be the corresponding estimator, based on the original, unedited data $\{y_{j,k}\}$.

The difference between the estimate $\hat{T}_j = \sum_k w_k y_{j,k}$, based on the original, unedited data and the estimate $\hat{T}_j^{(e)} = \sum_k w_k y_{j,k}^{(e)}$, based on the edited data can be quite large in many surveys , indicating that at least some degree of editing is necessary.

Definition 3.5:

$Im\,p_{j,k} = w_k \cdot (y_{j,k} - y_{j,k}^{(e)})$ measures the **impact** made on the estimate $\hat{T}_j^{(e)} = \sum_k w_k y_{j,k}^{(e)}$ if the data item $y_{j,k}$ is not edited.

In the following section we will introduce two important concepts, **score functions** and **pseudo bias** which both play a vital role in selective editing methodology.

3.4 Tools for planning & evaluation of systems for selective editing

The impact $Im\,p_{j,k} = w_k \cdot (y_{j,k} - y_{j,k}^{(e)})$ of $y_{j,k}$ can be evaluated only when $y_{j,k}^{(e)}$ is known, i.e. after being followed-up. In order to form the basis of a workable editing system the edited data $\{y_{j,k}^{(e)}\}$, being unknown, must be replaced by **predicted values** $\{\tilde{y}_{j,k}\}$.

Several ways to establish the predicted values $\tilde{y}_{j,k}$ have been suggested, such as using the value from the previous production round, a predicted value from time series analysis with seasonal adjustment and a prediction by regression analysis using various sets of explanatory variables. This issue will be discussed in more detail in chapter 4.

As a proxy for impact we introduce the **potential impact**, which can be computed for each item before a possible follow-up**:**

Definition 3.6:

The **potential impact** $Potimp_{j,k} = w_k \cdot (y_{j,k} - \tilde{y}_{j,k})$ is a proxy for $Im\,p_{j,k} = w_k \cdot (y_{j,k} - y^{(e)}_{j,k})$ in measuring the impact made by the unedited data value y_{jk} on the estimate $\hat{T}_j^{(e)} = \sum_k w_k y_{j,k}{}^{(e)}$.

We need a few more concepts:

Definition 3.7:

The **anticipated impact** is the product of suspicion and potential impact: $Antimp_{j,k} = Susp_{j,k} \cdot Potimp_{j,k}$.

Suppose that a survey has been exposed to extensive and rigorous editing of collected data. Consider the following function:

$$LScore_{j,k} = \frac{Susp_{j,k} \cdot w_k \cdot \left| y_{j,k} - \tilde{y}_{j,k} \right|}{SE(\hat{T}_j^{(e)})}.$$

The quantity $LScore_{jk}$ is called the **local score** (or item score) for data item $y_{j,k}$. It follows from definitions 3.6 and 3.7 that the local score can be written in the following form:

$$LScore_{j,k} = \frac{\left| Antimp_{j,k} \right|}{SE(\hat{T}_j^{(e)})}.$$

The local score is anticipated impact related to some aggregate such as the estimated total or the standard error. In this particular case we have chosen to use the standard error. A more general definition of the local score function is presented in chapter 4.

Furthermore, set $GScore_k = \max_j \left\{ LScore_{j,k} \right\}$. $GScore_k$ is called the **global score**[16] for observational unit *k.*

Let all units with $GScore_k > \beta$ for some predetermined threshold β be edited while the original data are kept for all other units.
The following function

$$RPB_j(\beta) = \frac{\sum_{k \in \{GScore_k \le \beta\}} w_k (y_{j,k} - y_{j,k}{}^{(e)})}{SE(\hat{T}_j^{(e)})}$$

measures the error imposed on $\hat{T}_j^{(e)}$ if units with $GScore_k \le \beta$ are left unedited. $RPB_j(\beta)$ is called **relative pseudo bias,** see Latouche and Berthelot (1992), Lawrence and McDavitt (1994).

The previous discussion can be summarized in the following procedure, which is similar to a large degree with the approach by Latouche and

[16] There are several different functions suggested in the literature. Another such function is $GScore_k = \sum_j LScore_{j,k}$. We will return to the choice of global score in chapter 4.

Berthelot (1992). It outlines how selective editing may be carried out. We will propose a more generalised procedure in chapter 4.

Outline of a procedure for selective editing

1) Carry out extensive and rigorous editing of collected data during some production round. Compute $Susp_{j,k}$ for all *j* and *k*. Save both original and edited data , $\{y_{j,k}\}$ and $\{y_{j,k}^{(e)}\}$. Choose predicted values $\{\tilde{y}_{j,k}\}$ for $\{y_{j,k}^{(e)}\}$ by some prescribed method.
2) Compute for all *j* and *k:* $LScore_{j,k} = \dfrac{Susp_{j,k} \cdot w_k \cdot \left| y_{j,k} - \tilde{y}_{j,k} \right|}{SE(\hat{T}_j^{(e)})}$

   and $GScore_k = \max\limits_j \{LScore_{j,k}\}$.

Select a threshold value β for *GScore* such that

3) $RPB_j(\beta) = \dfrac{\sum\limits_{k \in \{GScore_k \le \beta\}} w_k \cdot (y_{j,k} - y_{j,k}^{(e)})}{SE(\hat{T}_j^{(e)})}$ are sufficiently small for all *j*.
4) For the current production round, use the selected threshold β , compute $GScore_k = \max\limits_j \{LScore_{j,k}\}$ and select only the observational units with $GScore_k > \beta$ for follow-up.

In traditional editing systems the concept of suspicion plays a major role in the forming of edits and decision rules with no visible role for the concept of impact. In selective editing, impact plays a major role, and a question is whether selective editing can do without the suspicion concept. In other words, would the above procedure work without $Susp_{j,k}$, i.e. with $Susp_{j,k}$ = 1 for all pairs of *j* and *k*? Practical experience e.g. from several of the case studies mentioned in chapter 2 indicates that procedures based on impact solely tend to generate a large proportion of “false alarm” i.e. a large proportion of units above the threshold with little or no impact on the statistical estimates.

Hence a main conclusion from practical experience is that both concepts, suspicion and impact are necessary in order to form an efficient editing system.

Example of the simultaneous use of suspicion and impact:
In the Intrastat “price editing” suspicion is based on the price per kilo ratio for reported transactions. Suspicion is solely an attribute of the ratio in comparison with the dispersion of such ratios historically. It is a continuous measure – see chapter 4 ahead for a general definition. The potential impact is measured in term of effect on trade in Swedish kronor. The overall impact is determined for domains of study according to several classifications. Impacts on output depends on the value of the transaction and the sizes of domains of study where it has influence. Thus, there can be a weak correlation between suspicion and potential impact. In logarithmic scale we have

$$\log\left|Antimp_{jk}\right| = \log Susp_{jk} + \log\left|Potimp_{jk}\right| .$$

The decision rule is to flag units with $\log\left|Antimp_{jk}\right|$ greater than a threshold. These are the units shown in the upper right corner of the diagram.

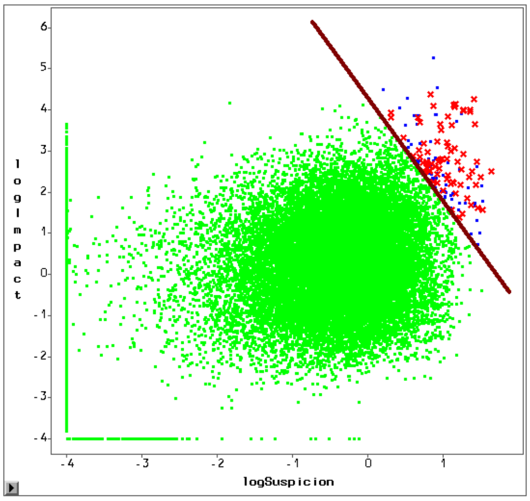


Figure 3.3. Logarithmic potential impact on the vertical axis and logarithmic suspicion on the horizontal axis in the “price editing” of Intratstat.

4 Generalized concepts in selective editing

4.1 Introductory remarks

An editing system must be able to treat many variables, several secondary observational units within each primary unit and statistical estimates for many different domains of study simultaneously. One data item $y_{j,k,l}$ may contribute to several domain estimates. These are some aspects that have been deliberately disregarded so far in order to make the text more easily accessible.

In the previous chapter we introduced some important concepts for selective editing, notably suspicion, impact and score functions. These all need to be given in more general forms to cover all situations that an edit system will face.

4.2 Several variables

With some exceptions, such as for Structural business statistics [SBS] a survey includes a fixed number of variables, the same number for each respondent. Some surveys are restricted to a single main variable, such as price.

The variables may be measured on different scales, e.g. deliveries in SEK, kilograms, employees’ hours worked or wage totals in SEK. Here, we relate the impacts of potential errors to estimates of the totals or standard errors in, for example, preceding surveys, thus giving us an dimensionless indicator.

We introduce a coefficient (parameter) for the relative importance of each variable *j*. The person in charge of the statistics should know which variables are most important for the users.

Another possible use for such a coefficient is also to adjust the score functions to variations in how well the predicted values function for the variables.

Which variables should be included in the score function? The statistical output should be decisive in designing the editing. Suspicion can be computed for all collected variables. Potential impact is computed only for variables presented in the statistical output. For variables that are not explicitly presented in the output, but contribute for example to a sum that is an output variable, the local score is computed for the sum variable using the max suspicion value from all these variables.

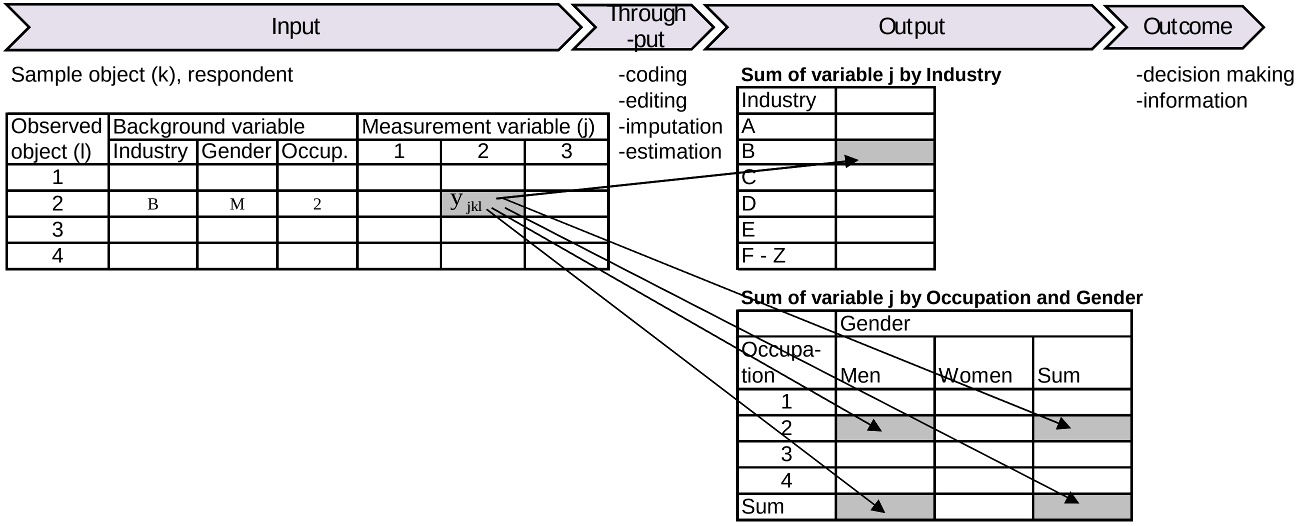
Some variables included in the score function may be strongly correlated. In the Short-term wages and salaries in the private sector survey, this is a matter of a derived variable that is the sum of its components. What effect this has on the result of prioritising units depends, inter alia, on how global scores are constructed. Summing of local scores may have greater importance than using maximum local scores. *Further analysis is needed into correlated variables.*

4.3 Several classifications in the reporting

When there are several classifications in the grouping of the statistical output into domains of study, we need to be able to “globalise” the anticipated impacts for data values obtained on the output as a whole. The surveys Wage and salary structures in the private sector and Rents for dwellings as presented in chapter 2 are examples of statistics with several classifications. Rents are reported by category of owner, type of housing unit, and region. On the opposite, some short-period statistical surveys, such as PPI, have inclusion in the overall gross domestic product (GDP) as the most important final use. In such cases, there is no need to evaluate variable impacts on the statistics in terms of classifications.

The person in charge of the statistics should provide a description of what is most important, evaluated in relative numbers. This information should be used to adjust parameter settings etc. in order to use follow-up for those output table cells and margins where it is most needed.

Figure 4.1. Impact on statistics. One data item has impact on several output table cells.


Input
Through -put
Output
Outcome
Sample object (k), respondent
-coding
-editing
-imputation
-estimation
Sum of variable j by Industry
-decision making
-information
Sum of variable j by Occupation and Gender
y jkl

Observed object (l)	Background variable			Measurement variable (j)		
	Industry	Gender	Occup.	1	2	3
1						
2	B	M	2		y_{jkl}	
3						
4						

Industry	
A	
B	▨
C	
D	
E	
F - Z	

	Gender		
Occupa-tion	Men	Women	Sum
1			
2	▨		▨
3			
4			
Sum	▨		▨

Example 4.3.1: Producer price index (PPI)

Price indices at producer and import stages have a very clear primary use. The statistics are used for the fixed-price computation of the value of production and consumption. The final use of these is GDP. Thus, for each individual item (price data), we can approximately calculate the relative impacts of potential errors in unedited input data.

Example 4.3.2: Different classifications in the SLP survey

A central classificatory variable is *personnel category,* with the categories *workers* and *white-collar employees*. For workers, *average hourly wage* is the most important parameter, for graduate employees *average monthly salary*. Other classificatory variables are *gender, occupation* (three- and four-digit SSYK[17]), *industry* (two-digit SNI), *age, education* and *region* (NUTS2[18]). However, the following two combinations are the most important:

o *personnel category*gender*industry* – for the parties of the labour market (employers and employees),

o *personnel category*gender*occupation* (four-digit) – for the public.

Example 4.3.3: Different classifications in the Intrastat survey

Foreign trade of goods is reported in values in SEK, broken down by inflow and outflow in total, with groupings of goods according to 2- and 3-digit SITC[19] and 2-, 4- and 6-digit CN[20].

A single erroneous value will thus give rise to errors in six output tables. For this survey, the potential error is related to historical levels for each domain of study. The producer of the statistics in cooperation with the methodologist has determined coefficients that specify how important each classification is relative to the others. When the anticipated impact of a suspected data item has been estimated for each of the six classifications, the maximum of the six is chosen.

4.4 Several observations (records) per respondent

As exemplified by the surveys Wage and salary structures in the private sector, Rents for dwellings, Intrastat and PPI, it is common for respondents to have a highly variable number of observations. Thus, editing leads to respondents obtaining a variable number of observations (records) for follow-up.

It would be an attractive property of an editing system if the selective editing could focus on respondents, i.e. maximise the effect on statistics and minimising workload for respondents. For practical reasons it might be necessary to focus on primary sampling units. In the case of Rents for dwellings this is especially a drawback. Here the primary sampling unit is a flat and it turns out that one respondent can get many primary sampling units to respond for. There will be no effective selective editing unless respondents are prioritised for their total response.

Conceptually, the difference between variables and records is not always self-evident. In the SBS-survey variables are stored by line. In practice, this leads to accounting items being envisaged as records. By analogy, product items in Intrastat might be envisaged as variables, but then there would be around 20 000 variables, virtually all of which would be empty in the case of any single respondent. There is also a certain dissimilarity between the SBS and Intrastat in that for the former, the number of variables (accounting items) varies between enterprises even within the same sector, but even so they are not as numerous as the number of product items in Intrastat.

[17] Standard för Svensk Yrkesklassificering, SSYK 96, Swedish Standard Classification of Occupations 1996, is an adaptation of the International Standard Classification of Occupations, ISCO-88.

[18] The Nomenclature of Territorial Units for Statistics

[19] Standard International Trade Classification

[20] Combined Nomenclature

Example 4.4.1 of global scores in relation to observations: Wage and salary structures in the private sector

At first global scores were computed by summation of local scores for all observations. It became necessary to avoid situations where large enterprises obtain a high local score solely because they have many individuals, albeit with low item scores. As an alternative, global scores (at enterprise level, LeU) can be computed as the sum of the local scores that exceed a certain threshold value.

Example 4.4.2 of global scores in relation to observations: Intrastat

Every product item in Intrastat receives an item score based on suspicion and potential impact. This is certainly “global” with respect to the six different classifications in the statistical output. Currently, 1 230 product items are flagged for follow-up per month, regardless of respondent, i.e. only on the basis of scores for product items. Around 740 respondents are responsible for the 1 230 observations, most providing just one observation, but a few as many as 50.

It might be appropriate to have a global score for respondent (identifier: Orgnr/TillNr). First, set a threshold (*T1*) for the local (product item) score. Sum the truncated item scores greater than the threshold *TI* for each respondent. The respondents whose sum of truncated item scores is greater than threshold *T2* need to be re-contacted and asked about the product items already extracted. A cost function is needed to find a cost-effective design. Here, the following function has been employed:

$$Total\ cost = \left[No.of\ \ respondents\right] \cdot C1 + \left[No.of\ \ product\ items\right] \cdot C2\ .$$

Let *C1* = *C2* = 20 SEK. With the same total cost as at present we have shown empirically that the largest sum of local item scores is attained with around 620 respondents and 1 350 product items instead of 740 respondents and 1 230 product items. This gives an efficiency gain of 5 percent.

4.5 General form of a suspicion measure for query edits

As noted in chapter 3, a typical query edit takes the form $\tilde{t}^{(L)}_{r,k,l} < t_{r,k,l} < \tilde{t}^{(U)}_{r,k,l}$ where the lower and upper limits are provided by the user.

The test variables $t_{r,k,l}$, *r=1...R,* are derived from the data $y_{j,k,l}$ and from other variables, and provide a basis for a measure of suspicion assigned to each data item $y_{j,k,l}$. We focused on a dichotomized suspicion measure in chapter 3 but indicated that a continuous measure might be preferable. In this section such a continuous measure on the interval [0,1] will be proposed.

Although the relation between *t* and *y* is generally many-to-many it may be the case that $t_{r,k,l} = y_{j,k,l}$ or $t_{r,k,l} = \log y_{j,k,l}$ where $y_{j,k,l}$ for example is a price ratio in the PPI survey. Sometimes a test variable is the difference or ratio between a received value and a corresponding value from a register or the previous survey.

Notice that the suspicion measure is primarily defined for each edit, i.e. suspicion is assigned for the test variable $t_{r,k,l}$. We denote this measure $Susp^{(t_r)}_{k,l}$. A suspicion measure is then assigned for each data value $y_{j,k,l}$ as the maximum over *r* of all $Susp_{k,l}^{(t_r)}$ for edits where data item $y_{j,k,l}$ has been involved. We denote this measure $Susp_{k,l}^{(y_j)}$ to separate it from $Susp_{k,l}^{(t_r)}$.

It should also be emphasized that every failed fatal edit always implies $Susp_{k,l}^{(y_j)} = 1$ for every data item $y_{j,k,l}$ involved.

The test variable $t_{r,k,l}$ is a function of one or several collected variables $y_{j,k,l}$. Let $\tilde{t}_{r,k,l}$ be the predicted test variable value for $t_{r,k,l}$. Furthermore, let as above $\tilde{t}_{r,k,l}^{(L)}$ and $\tilde{t}_{r,k,l}^{(U)}$ be the lower and upper acceptance limits of the edit.

Next we will propose a continuous measure of suspicion that always lies between 0 and 1.

Definition 4.1:

A) Let $Ratio_{r,k,l} = 0$

if $\tilde{t}_{r,k,l} - KAPPA \cdot (\tilde{t}_{r,k,l} - \tilde{t}_{r,k,l}^{(L)}) < t_{r,k,l} < \tilde{t}_{r,k,l} + KAPPA \cdot (\tilde{t}_{r,k,l}^{(U)} - \tilde{t}_{r,k,l})$

where *KAPPA* $\geq$ 0 is a parameter to be discussed below.

B) Let $Ratio_{r,k,l}$ =

$$= \left(\tilde{t}_{r,k,l} - KAPPA \cdot \left(\tilde{t}_{r,k,l} - \tilde{t}_{r,k,l}^{(L)}\right) - t_{r,k,l}\right) / \, max\left\{\left(\tilde{t}_{r,k,l}^{(U)} - \tilde{t}_{r,k,l}^{(L)}\right), SUSP\_ALFA \cdot \tilde{t}_{r,k,l}\right\}$$

if $t_{r,k,l} < \tilde{t}_{r,k,l} - KAPPA \cdot \left(\tilde{t}_{r,k,l} - \tilde{t}_{r,k,l}^{(L)}\right)$,

where *SUSP_ALFA* > 0 is a parameter to be discussed below.

C) Let $Ratio_{r,k,l}$ =

$$= \left(t_{r,k,l} - \tilde{t}_{r,k,l} - KAPPA \cdot \left(\tilde{t}_{r,k,l}^{(U)} - \tilde{t}_{r,k,l}\right)\right) / \, max\left\{\left(\tilde{t}_{r,k,l}^{(U)} - \tilde{t}_{r,k,l}^{(L)}\right), SUSP\_ALFA \cdot \tilde{t}_{r,k,l}\right\}$$

if $t_{r,k,l} > \tilde{t}_{r,k,l} + KAPPA \cdot \left(\tilde{t}_{r,k,l}^{(U)} - \tilde{t}_{r,k,l}\right)$

D) Let $Susp_{k,l}^{(t_r)} = Ratio_{r,k,l} / (TAU + Ratio_{r,k,l})$

where *TAU* > 0 is a parameter set by the user , for example *TAU* = 1.

Parameter names written as a full name, such as TAU instead of $\boldsymbol{\tau}$*, refer to parameters which can be set by the user to adjust the score calculations in the generalized tool for editing. See next chapter.*

Notice that the dichotomized suspicion measure as defined in chapter 3, may be obtained (with arbitrarily good approximation) by setting *TAU* to a small number, e.g. *TAU* = 10^{-6} and *KAPPA* = 1.

The parameter *KAPPA* takes a value equal to or larger than 0, most often 0 or 1 and almost never larger than 3. This parameter is a means to decide how much of the sample that will have zero suspicion. More focus on potential impact, even when data looks reasonable, is achieved when *KAPPA* = 0. Focus on a minor part of data that is really suspected is achieved when *KAPPA* > 2.

The choice of *KAPPA* is however dependent on available data from previous survey rounds when analysing and deciding parameters in THE LAB. Within earlier acceptance regions there are no edited data. Small values of *KAPPA* are not useful except for the situation where a sample is used below the global threshold or when simulated data are used.

The parameter *SUSP_ALFA* is necessary to make the ratio computable when the dispersion range is zero. We have little empirical experience, it is suggested to be 0,05.

Examples of suspicion as a function of KAPPA and TAU

We are modelling suspicion with two parameters KAPPA and TAU and a homogenous group of cold deck data. Here we exemplify by setting predicted value as the median and the measure of dispersion as the quartiles in the cold deck data.

KAPPA =0 gives suspicions >0 when the observed unedited value differ from the median of the cold deck data., i.e. for practically all data.

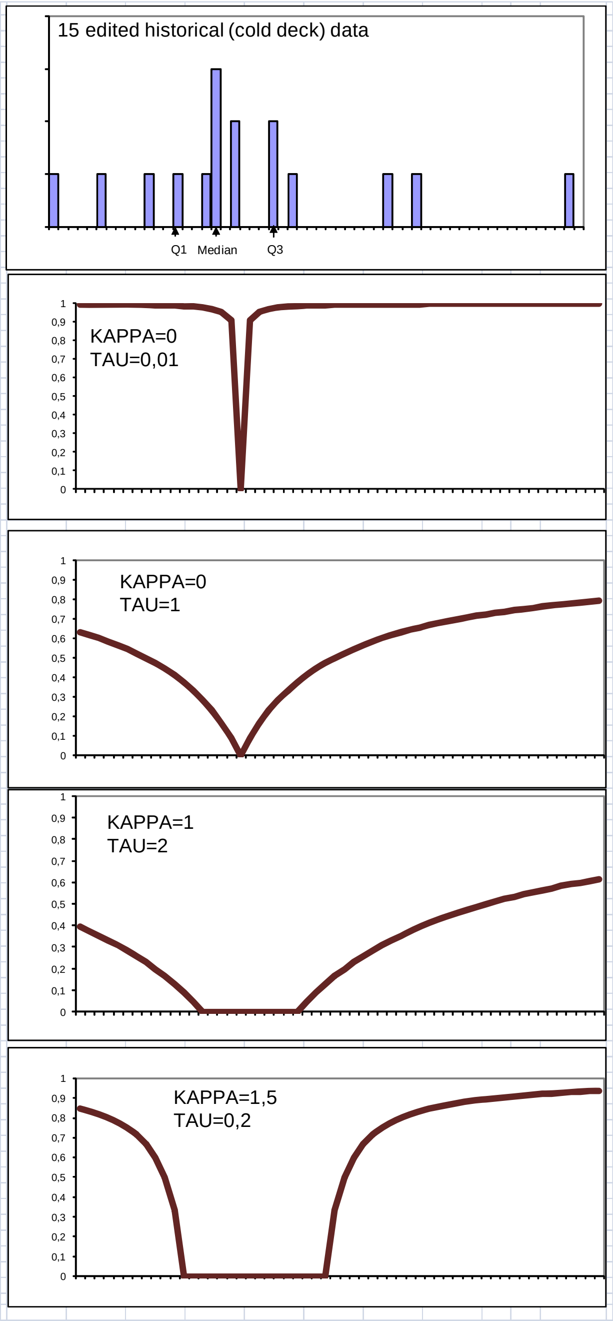
KAPPA =1 gives suspicion=0 for unedited data between the lower and the upper quartiles of the cold deck data.

Larger KAPPA´s broaden the range where suspicion is set to zero.

TAU is decisive for the shape of the curve. For small TAU-values suspicion is close to 1,0 when an observed unedited value lies outside the interval decided by KAPPA.

A large TAU makes the suspicion almost proportional to the distance away from the centre of the distribution in the cold deck data.

Figure 4.2 Modelling suspicion on cold deck data



4.6 Impact

4.6.1 General definitions

For each cell *d,j* we estimate the totals $T_{d,j}(t)$ from a census or a sample for the reference period *t*. In most surveys this estimate is a linear combination of observed unedited and edited data from period *t*. If so, the actual impact on the estimated sum for any cell *d,j* is the weighted difference between the unedited and the edited data item:

Definition 4.2: Impact

A) $IMP_{d,j,k,l} = w_k \cdot w_{k,l} \cdot (y_{j,k,l} \text{ - } y^{(e)}_{j,k,l})$ if $(k,l) \in d$ [21]

B) In case of non-linear estimators, such as ratios, a linear Taylor's expansion for impact often suffices. We then have:

$$w_k \cdot w_{k,l} \cdot \left(y_{j1,k,l} - y^{(e)}_{j1,k,l} - \theta_{d,j1/j2} \cdot \left(y_{j2,k,l} - y^{(e)}_{j2,k,l}\right)\right) \cdot \frac{1}{T_{d,j2}},$$

where $\theta_{d,j1/j2} = \dfrac{T_{d,j1}}{T_{d,j2}}$ is the ratio of totals for variables Y_{j1} and Y_{j2} in the domain of study *d*.

The actual impact is unknown until after the follow-up. If the edited values $y^{(e)}_{j,k,l}$ are replaced by the predicted values $\tilde{y}_{j,k,l}$ we can introduce the potential impact of $y_{j,k,l}$ when it is included in a domain *d*:

Definition 4.3: Potential impact

A) $Potimp_{d,j,k,l} = w_k \cdot w_{k,l} \cdot (y_{j,k,l} - \tilde{y}_{j,k,l})$ if $(k,l) \in d$

B) In case of non-linear estimators, such as ratios, a linear Taylor's expansion for impact often suffices. We then have:

$$w_k \cdot w_{k,l} \cdot \left(y_{j1,k,l} - \tilde{y}_{j1,k,l} - \hat{\theta}_{d,j1/j2}(t-1) \cdot \left(y_{j2,k,l} - \tilde{y}_{j2,k,l}\right)\right) \cdot \frac{1}{\hat{T}_{d,j2}(t-1)},$$

where $\hat{\theta}_{d,j1/j2}(t-1) = \dfrac{\hat{T}_{d,j1}(t-1)}{\hat{T}_{d,j2}(t-1)}$ is the ratio of totals for variables Y_{j1} and Y_{j2} in the domain of study *d* for some historical survey round (*t-1*).

Definition 4.4: Anticipated impact

A) The anticipated impact is suspicion multiplied with potential impact:

$$Antimp_{d,j,k,l} = Susp^{(y_j)}_{k,l} \cdot w_k \cdot w_{k,l} \cdot \left(y_{j,k,l} - \tilde{y}_{j,k,l}\right) \text{ if } (k,l) \in d\,.$$

B) And analogously for ratios:

$$Antimp_{d,j1/j2,k,l} =$$

$$= Susp^{(y_{j1,j2})}_{k,l} \cdot w_k \cdot w_{k,l} \cdot \left(y_{j1,k,l} - \tilde{y}_{j1,k,l} - \hat{\theta}_{d,j1/j2}(t-1) \cdot \left(y_{j2,k,l} - \tilde{y}_{j2,k,l}\right)\right) \cdot \frac{1}{\hat{T}_{d,j2}(t-1)}$$

4.6.2 Impact from fatal error

A fatal error in $y_{j,k,l}$ will always imply $Susp^{(y_j)}_{k,l}$ = 1. If the variable y_j is subject to item non-response, i.e. $y_{j,k,l}$ is missing, then the expression for anticipated impact cannot be evaluated.

[21] A data value which contributes to several domains will have one impact value for every such domain.

To assess the gravity of the missing $y_{j,k,l}$ we can replace it with $\tilde{y}^{(L)}_{j,k,l}$ which is the lower limit for some measure of "confidence" of $\tilde{y}_{j,k,l}$ in analogy with $\tilde{t}^{(L)}_{k,l,r}$ as defined earlier.

Hence, the expression $w_k \cdot w_{k,l} \cdot \left| y_{j,k,l} - \tilde{y}_{j,k,l} \right|$ for potential impact would then be replaced by $w_k \cdot w_{k,l} \cdot \left| \tilde{y}^{(L)}_{j,k,l} - \tilde{y}_{j,k,l} \right|$. However, the user may in some cases prefer not to let item non-response for a certain variable y_j contribute to the score function. To account for this situation we introduce a parameter *NONRESP_IMPACT* which can be set by the user.

The whole relation for anticipated impact is then replaced by:

$$Antimp_{d,j,k,l} = w_k \cdot w_{k,l} \cdot \left| \tilde{y}^{(L)}_{j,k,l} - \tilde{y}_{j,k,l} \right| \cdot 1 \cdot NONRESP\_IMPACT\ . \qquad (4.6.1)$$

In the case of a ratio we have for example

$$Antimp_{d,j1/j2,k,l} =$$

$$Susp^{(y_{j1,j2})}_{k,l} \cdot w_k \cdot w_{k,l} \cdot \left(\tilde{y}^{(L)}_{j1,k,l} - \tilde{y}_{j1,k,l} - \hat{\theta}_{d,j1/j2}(t-1) \cdot \left(\tilde{y}^{(L)}_{j2,k,l} - \tilde{y}_{j2,k,l} \right) \right) \cdot \frac{NONRESP\_IMPACT}{\hat{T}_{d,j2}(t-1)}$$

If $NONRESP\_IMPACT = 0$ then the missing y_j does not contribute to the score function. The default value for *NONRESP_IMPACT* equals 1 while $Susp^{(y_j)}_{k,l}$ always equals 1 for fatal errors. The notation $Antimp_{d,j,k,l}$ ("Anticipated Impact") is not quite appropriate here since $\tilde{y}^{(L)}_{j,k,l}$ is only an arbitrary replacement for the missing value. But we use this notation nevertheless to make the appearance of (4.6.1) and the definition (4.3) of anticipated impact correspond.

When a value $y_{j,k,l}$ exists but is considered as a fatal error then the expression amounts to $Antimp_{d,j,k,l} = w_k \cdot w_{k,l} \cdot \left| y_{j,k,l} - \tilde{y}_{j,k,l} \right| \cdot 1 \cdot 1$, where $Susp^{(y_j)}_{k,l}$ and *NONRESP_IMPACT* have both implicitly been set equal to 1.

4.7 Importance parameters

The design of the selective editing can be adjusted with so called importance parameters in order to capture survey-specific needs. To set these parameters judgementally good knowledge about the survey is needed. Thus it is recommended to involve the person(s) in charge of the survey at this stage. Analysis of empirical data from previous survey rounds are valuable when setting the parameters. The default value for all importance parameters is set to 1, which will do for many surveys.

Importance of variables; parameters *VIOLIN$_j$* (j = 1, …, J)

The parameter *VIOLIN$_J$* can be varied between variables that are included in the selective editing process. Variables for which the value of *VIOLIN$_j$* is high will tend to be prioritized in favour of variables with lower values.

Importance of classifications; parameters *CLARINET$_{c(d)}$* c (c = 1, …, C)

The parameter *CLARINET$_{c(d)}$* is the relative importance of the classification c (c = 1, …, C) in reporting of the statistics. A domain of study d is constructed by the use of a classification c. Classifications for which the value of *CLARINET$_{c(d)}$* is high will be prioritized in favour of classifications with lower values.

Importance of the size of domains of study; parameters *OBOE$_j$*

The parameters *OBOE$_j$* are introduced to give flexibility in the score function regarding absolute or relative impact on the statistics. The domains of study constructed by the classification variable *c*, can vary substantially in estimated levels $\hat{T}_{d,j,t-1}$ and standard error $SE\left(\hat{T}_{d,j,t-1}\right)$. It can be motivated to have different relative quality demands for domains of different sizes. *OBOE$_j$* can preferably be less than 1, for example 0,5, in such cases.

The three importance parameters form one part of the parameter *CELLO$_{d,j}$* which will be introduced next.

CELLO$_{d,j}$ is a unit-independent parameter specified for each output cell *d,j*. It transforms the anticipated impacts for variables on different scales and variation to comparable levels. *CELLO$_{d,j}$* is also a tool to prioritize specific tables, variables or cells that are more important than others. Besides the parameters the formula for *CELLO$_{d,j}$* contains the estimated total from a previous period, $\hat{T}_{d,j,t-1}$, its corresponding estimated standard error, $SE\left(\hat{T}_{d,j,t-1}\right)$, and parameter *CELLO_ALFA*. For the linear estimator case A:

$$CELLO_{d,j} = \frac{VIOLIN_j \cdot CLARINET_{c(d)}}{\left(maximum\left\{CELLO\_ALFA \cdot \hat{T}_{d,j,t-1}, SE\left(\hat{T}_{d,j,t-1}\right)\right\}\right)^{OBOE_j}} \quad (4.6.2.A)$$

CELLO_ALFA is the means to choose whether anticipated impact is related to $\hat{T}_{d,j,t-1}$ or $SE\left(\hat{T}_{d,j,t-1}\right)$. This is a choice made by the product manager. Because of the risk that some domain of study can have the estimated standard error = 0, it is wise to at least set a small *CELLO_ALFA*, for example *CELLO_ALFA* = 0,01 – 0,05. This parameter is universal for all variables.

> **Special case 1:** Setting *VIOLIN$_j$* = 0 means that variable *j* will be excluded from the calculations of the scores.
>
> **Special case 2:** Setting *OBOE$_j$* = 0 means that only the nominator will contribute to the value of *CELLO$_{d,j}$*. The contribution of the nominator will decrease with increasing values of *OBOE$_j$*.
>
> **Special case 3:** Setting a large *CELLO_ALFA$_{j,}$* for example 10, means that the bias is relative to the estimated sum of *T$_j$*.

In case of non-linear estimators, such as ratios, we have used the linear Taylor's expansion for impact. We have:

$$CELLO_{d,j1,j2} =$$

$$= \frac{VIOLIN_{j1,j2} \cdot CLARINET_{c(d)}}{\left(maximum\left\{ALFA_{j1,j2} \cdot X_{d,j1,j2,t-1}, SE\left(\hat{\theta}_{d,j1,j2,t-1}\right)\right\}\right)^{OBOE_{j1,j2}}} \quad (4.6.2.B)$$

X can be some positive measure of size which could be preferred rather than the standard error or to rescue the computation in case of zero standard error. In some surveys the *X* can be the estimated ratio. In price indices *X* can be the inverse of product weight.

The general tool must provide the means for manual adjustment of *CELLO*-values through direct access to a table including all *CELLO$_{d,j}$*-values.

Stability of $CELLO_{d,j}$

The edit-parameters $VIOLIN_j$, $CLARINET_{c(d)}$ and $OBOE_j$ of the parameter $CELLO_{d,j}$ will in general be kept constant between survey rounds. One reason for changing the settings of the edit-parameters might be that major changes have been made in the survey.

If $\hat{T}_{d,j,t-1}$ and $SE\left(\hat{T}_{d,j,t-1}\right)$ in the $CELLO_{d,j}$ formula are relatively stable over time, there is little gain in updating this information frequently. For short term statistical surveys it might be enough with updates of $\hat{T}_{d,j,t-1}$ and $SE\left(\hat{T}_{d,j,t-1}\right)$ every third, sixth or twelfth survey round. Thus the value of $CELLO_{d,j}$ will often be kept constant over different time periods. This is a good feature since the computation made in a single survey round should be as easy and fast as possible.

> **Example:** The following example is taken from Jäder and Norberg (2006). Foreign trade with goods involves the participation of all enterprises (LeU) with a turnover above a certain threshold. Today, around 15 000 enterprises are included. Large enterprises have several respondents, with *TillNr* as identifying variable. Each respondent supplies information on no, one or several goods items per month. The accounting and measuring instruments entail that some respondents can report on several items in the same flow, product code – country, i.e. items that it would have been possible to sum in advance.
>
> The suspicion $Susp_{j,k,l}$ is computed through comparison of price per quantity with medians and quartiles for similar goods items for all enterprises over the last 24 months (cross-sectional analysis).
>
> As predicted value $\tilde{y}_{j,k,l}$ is computed as recorded quantity for the product item in question multiplied by the historical median of unit price for as homogeneous a product group as possible. $\hat{T}^{t-1}_{d,j}$ is an average over 24 months.

Table 4.1. Classification in the reporting of foreign trade and the relative importance of classes, $CLARINET_c$, c = 1-6. The single size parameter *OBOE* = 0.4.

Report classification (*c*)	*c*	"Importance" of the classification $CLARINET_c$	Number of items flagged	Number of items not flagged
Other CN8	6	0.00001	0	0
Important CN8	5	0.067	13	4 871
CN6	4	0.1	600	145 761
SITC3	3	0.33	488	137 340
SITC2	2	0.5	390	90 605
Total in and outflow	1	1	0	0

The potential impact $Potimp_{d,j,k,l} = \dfrac{1 \cdot 1 \cdot \left| y_{j,k,l} - \tilde{y}_{j,k,l} \right| \cdot CLARINET_{c(d)} \cdot 1}{\left(\hat{T}^{t-1}_{d,j} \right)^{OBOE}}$

is computed without weighting, since the survey Intrastat is a total summation, and without the coefficient $VIOLIN_j$ for the relative significance of variable *j*, since the editing focuses on invoiced value. In reality, it is difficult for the editing process to show whether invoiced value or quantity is suspicious, so all impacts are expressed in value. The maximum of the six local scores for classifications is the global score for the observation.

Global scores have not yet been computed for respondents. A case study that followed up the method, proposes that the global score for respondent across observations is calculated using the sum of local scores with a deduction for a certain threshold value.

When the method was introduced in 2004 the number of edited records could be reduced by about 50 percent.

4.8 Local and global scores

The score function can be divided into three parts; the suspicion (*Susp*), the potential impact (*PotImp*) and $CELLO_{d,j}$. Notice that the $CELLO_{d,j}$-parameters can be regarded as weights applied to the potential impacts in order to make them comparable. The product of the three parts is the score at the most detailed level;

$$SCORE5_{d,j,k,l} = Susp_{k,l}^{(y_j)} \times Potimp_{d,j,k,l} \times CELLO_{d,j} \qquad (4.8.1)$$

The scores are globalised step by step, over domains of study (*d*), variables (*j*) and observations (*l*) to primary sampling or respondent unit (*k*) as:

$$SCORE4_{j,k,l} = \left(\sum_d \left(max\{0, SCORE5_{d,j,k,l} - THRESHOLD5\} \right)^{LAMBDA5} \right)^{1/LAMBDA5}$$

(4.8.2)

$$SCORE3_{k,l} = \left(\sum_j \left(max\{0, SCORE4_{j,k,l} - THRESHOLD4\} \right)^{LAMBDA4} \right)^{1/LAMBDA4} \qquad (4.8.3)$$

$$SCORE2_{k} = \left(\sum_l \left(max\{0, SCORE3_{k,l} - THRESHOLD3\} \right)^{LAMBDA3} \right)^{1/LAMBDA3} \qquad (4.8.4)$$

where *THRESHOLD5- THRESHOLD3* are local threshold values. The threshold for the global score is *THRESHOLD2*. The idea behind this very general formula is found partly in Hedlin (draft 2009).

The purpose of using local thresholds is to eliminate the effect that many small values of local scores for variables and respondent units all together contribute too heavy to the global score. This could be the case when using the sum parameter for LAMBDA parameters (LAMBDAx=1).

When using the continuous measure of suspicion (*Susp*), flags for variables are not set in AUTO-SELEKT to point out those variables that need to be followed up. In this case it is preferable also to use the local thresholds as the limit for follow-up.
We recommend as default value *THRESHOLD5* = 0. As a rule of thumb, the local threshold values *THRESHOLD4* and *THRESHOLD3* can be a tenth and a fifth respectively of the global threshold *THRESHOLD2*, if an elaboration has not been performed to find the best values.

Special case 1: If local threshold = 0 and *LAMBDA* = 1, we get the pure sum function.

Special case 2: If local threshold > 0 and *LAMBDA* = 1, we get sum of local scores with a deduction for a certain threshold value.

Special case 3: If local threshold = 0 and *LAMBDA* = 2, we get the sum of squares function.

Special case 4: If local threshold = 0 and *LAMBDA* = 10, we get a function very close to maximum.

4.9 Probability sampling of suspected units

4.9.1 Basis for evaluation

Nothing is known about the units that have not been followed-up. Hopefully, selective editing is efficient enough to decrease the total number of flagged units, compared to traditional editing. On the other hand this may be a disadvantage when it comes to evaluation. One possibility is to intensively follow-up a larger body of material from time to time. As a basis for adjusting the threshold value prior to a new survey round, follow-ups should also be performed on items lying below the threshold value. This can be sample-based on intermittent basis.

4.9.2 Inference from a probability sample of error-flagged units

In a survey with one variable (or just a few) and with one domain of study (or just a few) it should be possible to: (1) take a pps sample of the sampled units as a sub-sample, (2) re-contact the respondents in the sub-sample and (3) estimate the error in the point estimate with unedited data, so as (4) to perform a design-based adjustment of this estimate on the basis of the sub-sample. This long-winded process can be motivated if we know that errors tend to be non-symmetrical, for example if it is likely that the respondent can forget some accounts in the total sum. Ilves and Laitila (2009) give a deeper insight to the idea.

4.9.3 Sampling method

It is appropriate to select a sample of units with a global score lower then the threshold with probabilities proportional to the global score. In this application, it is reasonable to see to it that a large proportion of the units have a positive score; in any case, there will be a few units with low scores in the sample. A feasible method could be to make a Poisson sample with probabilities:

$p = \frac{n \cdot SCORE2_{k,t}}{\sum_k SCORE2_{k,t-1}}$, where n is the targeted sample size, $SCORE2_{k,t}$ is the global score for the respondent and $\sum_k SCORE2_{k,t-1}$ is the sum of global scores under the threshold for some previous survey. A cut-off for the smallest global scores can still be relevant. Poisson sampling is a suitable method as it can be used for the first to the last respondent regardless of inflow of data.

4.10 Central and dispersion measures

There are two quite different approaches to find a predicted value $\tilde{y}_{j,k,l}$ for $y_{j,k,l}$, a predicted value $\tilde{t}_{r,k,l}$ for $t_{r,k,l}$ and a dispersion interval to the latter:

1. Time series data for the observed unit *k,l*.
2. Cross-sectional data for a homogenous group of units similar to *k,l*.

We propose the use of both time series and cross-sectional measures, setting the time series measures at priority and the cross-sectional as reserve when time series are not long enough.

4.10.1 Cold-deck data

The elementary components of the global score need good estimates of predicted values and dispersion. These estimates must be computed before the editing of survey round *t*, at least if one wants to edit data records as they arrive. Calculations are based on edited data without using raising factor, for the stated groups. All units are included in the groups for cross-sectional data no matter if they belong to an old, outgoing sampling panel.

- A decision is to be made whether to include imputed data or if they must be excluded. It seems most advisable from a theoretical point of view not to use imputed data, but it is easier not to make a difference between imputed and collected data.
- A decision must also be made whether to include data that was suspicious in the previous survey but not flagged because the potential impact was low, or not. Again it seems to be a good idea not to use very suspected data, but again it is easier not to make a difference.

For the time series data several survey rounds are needed. The set of data for time series is specified by setting start and end of period.

For cross-sectional data one survey round can do, but several can be preferable. The data should be as fresh as possible. The set of data for cross-sectional estimates is also specified by setting start and end of period. There is also a possibility to make an exception for some set of data, for example it can be wise to exclude holiday months from hot-deck if we will edit typical labour months.

4.10.2 Hot-deck data

From a methodological point of view there should not be a problem to use data from the current survey round. This can be done as a pure hot-deck. It could also be done by a successively updating of predicted values and dispersion measures based on both cold- and hot decks. If the coming IT-architecture supports this, we do not know at present.

4.10.3 Time series models

In most surveys it is not necessary to use complicated time series models. A simple average of the last three survey rounds can do. If the survey measures phenomena with a heavy seasonal pattern, such as production decline in June, steps should be taken to pay regard to this.

There is a SAS procedure for time series analysis with default modelling of an Arima-model, the UCM-procedure, see appendix. Using these models we get a predicted value and also a confidence interval to be easily used.

We introduce *TS_Min_Obs* as the parameter stating the minimum number of historical data needed for the computation.

4.10.4 Homogenous groups for cross-sectional measures

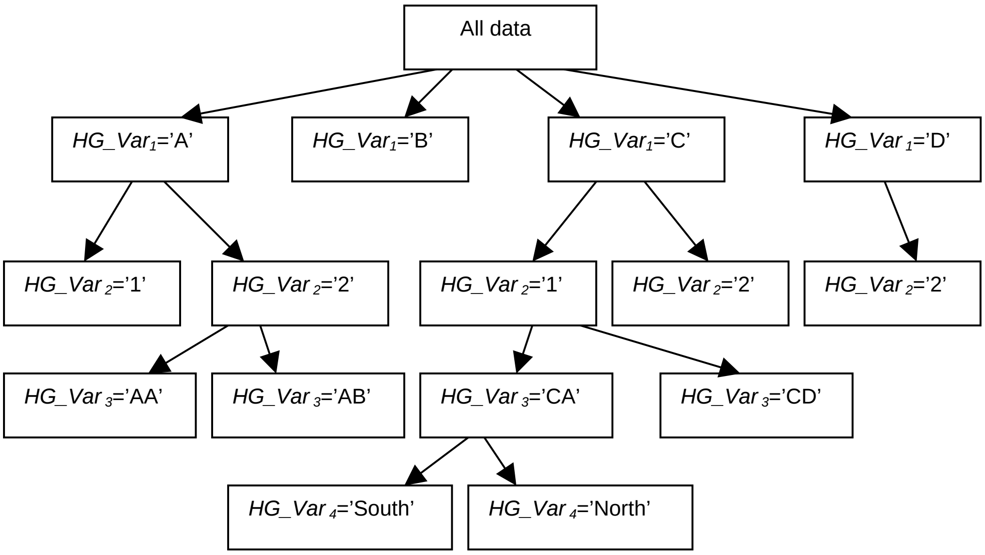
It is necessary to construct homogenous groups for which predicted values and dispersion intervals are computed. These groups may, but need not, correspond to strata or domains of study. It is our empirical view that it is more important to stress homogeneity rather than a large number of records in the calculation of predicted values for the purpose of editing.

Homogenous groups can be defined after more or less advanced analysis of cold-deck data. There are various multivariate methods to be used, without mentioning any. The result of such an exercise should be a variable that identifies all the homogenous groups.

In the SELEKT software there is a built-in functionality that constructs homogeneity groups by setting a set of classificatory variables. This set must be accompanied by the levels of each classificatory variable (number of digits). A parameter stating the minimum number of observations for the computation is also required. We introduce:

- $HG\_Var_1 – HG\_Var_H$ being H variables in the cold-deck (and current data) which are used for computation of the predicted value of Y_j and the predicted value and measures of dispersion for test variables t_{jr}.
- They are accompanied by H parameters $HG\_Dig_1 – HG\_Dig_H$ stating the number of digits to be used for each variable. The variables must be of character type.
- *CS_Min_Obs* is the parameter stating the minimum number of data needed for the computation. The homogeneity groups are constructed in an hierarchical tree-structure as long as there are at least *CS_Min_Obs* records in the cold-deck data to form the groups.

Figure 4.3. Homogeneity groups in an hierarchical tree-structure



HG_Var_1 and HG_Var_3 are the same variable in this illustration, utilizing one and two digits respectively.

There is no problem letting a homogeneity variable be the identification of an enterprise. Likewise, there is no problem having the same original variable at several levels with different digits, for example industry classification with one digit and the same classification with two digits where there are data enough. If a very complex definition of homogenous groups is desired a variable in the data must contain the definition of those groups and be used as HG_Var_1 alone

> **Example:** Intrastat has the following set of variables to get homogeneity in an hierarchical order. The more detailed levels are used as long as there are at least *CS_Min_Obs* = 5 historical data in the data base for the last 24 months.
>
> 1. Product group on 6-digit combine nomenclature
> 2. Product group on most detailed 8-digit combine nomenclature
> 3. In- or out-flow
> 4. Enterprise
> 5. Country (from which Sweden imports or to which Sweden exports)

4.10.5 Predicted value and dispersion

For cross-sectional analysis there are two immediately natural sets of measures; medians and arithmetic means, but there are more:

- Medians and lower and upper quartiles computed un-weighted across the previous survey rounds and observations *k,l* for homogeneous groups
- Arithmetic means ± standard deviation computed un-weighted across the survey rounds and observations *k,l* for homogeneous groups
- An auxiliary variable might help to predict the y-variable. If the individual ratios of the y- to the x-variable have a small variation over a homogenous group, central measure should be found for the ratio and be multiplied by the individual x-variable to yield a prediction of y.
- Prediction for observation *k,l* and variable *j* can be made on the basis of regression analysis with several explanatory variables. In SAS there is also an easy-to use procedure for robust regression, the procedure ROBUSTREG, which can be advantageous in this case. In future versions of a general box of tools this might be an option.

For time-series data there are many more possibilities:

- Edited values from the latest survey round $y_{j,k,l}^{(e)}$ and $t_{r,k,l}^{(e)}$. Here no measure of dispersion can be computed in much the same way. An interval can be defined as 0,9 and 1,1 times $y_{j,k,l}^{(e)}$ and $t_{r,k,l}^{(e)}$.
- A forecast for observations *k,l* by a time series analysis including confidence interval for the forecast can be computed in several ways, for example by the procedures UCM or FORECAST of SAS.
- Medians and lower and upper quartiles computed un-weighted across the previous survey rounds for observation *k,l*
- Arithmetic means ± standard deviation computed un-weighted across the survey rounds for observation *k,l*.

4.11 Suspicions and actions

The general concept of editing include all the following possible settings of suspicions and action taken for a suspected data value:

4. Fatal and suspected errors going thru to error list without a selective process. Flagging observations. The Flag is generated by a “traditional edit check” with test-variable, acceptance region and a homogenous group. No selective editing, the observation that fails the edit must be on the error list.
3. Flagging observations (all variables). Flag for fatal or suspected error by a “traditional edit check” as (4) above. This flag is used to set suspicion =1,00 for all survey variables for the observation. Local scores are computed and used in the forthcoming steps of selective editing. Not recommended.
2. Flagging single variables with a suspicion. The measure is generalised to allow for any positive suspicion, used from this edit check in the computation of local score for the corresponding survey variables in selective editing. Default is 1,00.
1. Suspicion is computed as a continuous measure in auto-selekt to be used in the computation of local score for the corresponding survey variables in selective editing.

4.12 Evaluation

4.12.1 Deciding the global threshold value

Achieving the goal of selective editing is dependent on finding an appropriate global threshold value. The global threshold value separates the units that will be followed-up from the ones that will pass without any manual action.[22] The global threshold value is chosen such that it will be sufficient to edit these units in order to avoid seriously biased estimates (Lawrence and McDavitt, 1994).

In case of one-stage data structure there is a one-to-one correspondence between the global threshold value, β_Q (*THRESHOLD2*), and the proportion of the responding primary units that require follow-up activities.

Let *Q* be the proportion of flagged primary units. The global threshold value is a percentile of the global scores:

$$\beta_Q = Percentile_{100-Q}(\, SCORE2_k\,) \qquad (5.1)$$

The decision of the value of the global threshold is based on an evaluation data set from at least one previous survey round where both edited and unedited data are available. The *Q*-value can be any value between 0 and 100.
A *Q*-value of 100 corresponds to a fully edited data set[23]. For every *Q*-value a corresponding so called RPB-value (se definition in next section) is computed for each output cell.

In a survey with clusters, for example enterprises and employees, we must find the best combination of thresholds for primary selected units, *THRESHOLD2* and for observed units, *THRESHOLD3*. Here, as well as in the one-stage-design, there is also the threshold for variables to be decided. To do this one must compute cost of the editing process as a function of the number of flagged primary selected units, flagged secondary selected units and flagged variables in total. Now it might be possible to find best threshold values, but the demand on data and computational resources is heavy.

4.12.2 Relative pseudobias (RPB)

RPB is a measure of the error that would be introduced in the estimates if the data had not been intensively edited. The absolute RPB-values are computed for each output cell *d, j* as follows:

$$\left|RPB_{d,j,Q}\right| = \frac{\left|\hat{T}_{d,j,Q} - \hat{T}_{d,j,Q=100}\right|}{SE\left(\hat{T}_{d,j,Q=100}\right)} \qquad (5.2)$$

where $\hat{T}_{d,j,Q}$ is an estimate of the total in an output cell, based on the evaluation data set where *Q* percent of the primary units with the highest global scores have been edited while the rest of the units have remained unedited. $SE\left(\hat{T}_{d,j,Q=100}\right)$ is the estimated standard error of $\hat{T}_{d,j,Q=100}$.

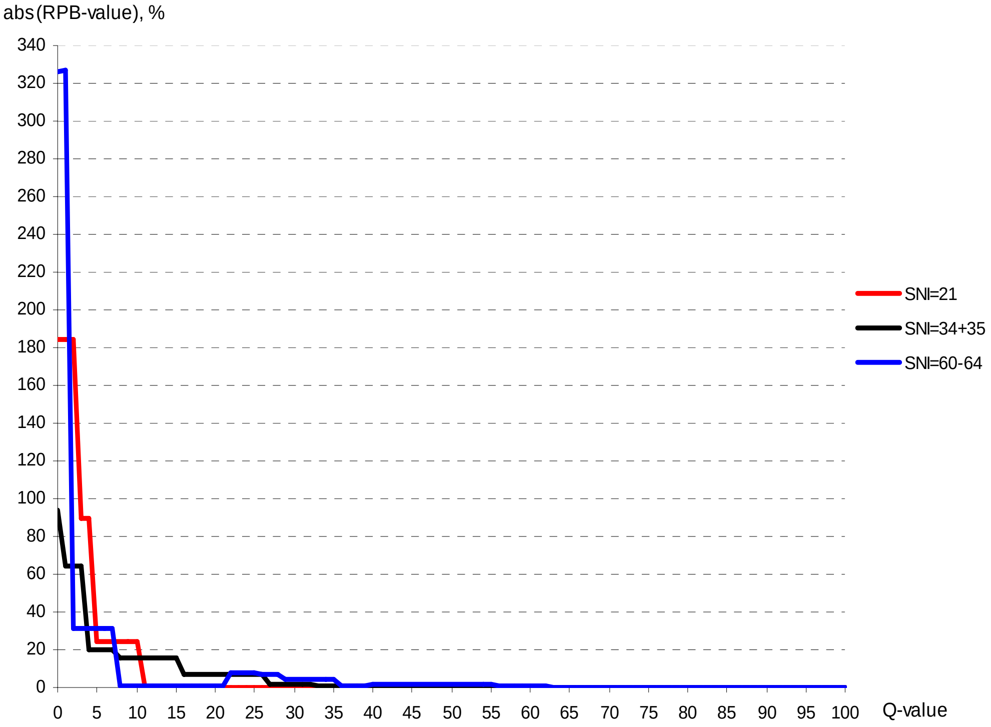
[22] Notice that non-statistical errors will always be taken care of either by imputation or by manual action whether the object is above or below the global threshold value.

[23] In practice we never have a dataset that is completely followed up by re-contacts

Notice that whenever $SE\left(\hat{T}\right)$ in the denominator is close to zero it can be replaced with some small fraction of $\hat{T}$.

Figure 5.1 shows the relationship between the RPB-values and the *Q*-values on the variable *Agreed monthly salary* for three arbitrary domains of study in *Short-term statistics, wages and salaries in the private sector*. The RPB-value often decreases rapidly when the *Q*-value is increased from zero, i.e. when the units with the largest impacts are being edited. The RPB-value will tend to zero as the *Q*-value increases, though not strictly monotonically decreasing. At the *Q*-value where all incoming units have been edited the RPB-value, of course, is zero.

Figure 4.4. Absolute RPB$_Q$-values for the variable *Agreed monthly salary* in three domains of study in *Short-term statistics, wages and salaries in the private sector*.



Each curve in figure 5.1 corresponds to a specific output cell. If we were willing to accept absolute RPB-values not greater than 20 percent it would be enough to follow-up around 10 percent of all units. This is under the condition that there are only three domains of interest. In this survey the total number of output cells that must be taken into account, when deciding the global threshold value, is about 200. The total number of output cells in a survey is given by the number of study variables multiplied by the number of (important) domains of study.

4.12.1 Analyzing simulation results

RPB-values should be produced for each combination of parameter settings and *Q*-values or costs of follow-up. An evaluation of these data, which can be substantial in amount, must be made if we want to learn about the general method and implement it in best way.

One purpose of studying the RPB-values is to search for and examine occurring extreme RPB-values. Subject-related facts and empirical knowledge should be used to determine whether an RPB-value can be accepted or not. If some RPB-values cannot be accepted one or more adjustments in the parameter settings have to be done. This calls for more simulations.

4.12.2 The resulting parameter setting

Once the best possible parameter setting including the global threshold value have been decided upon these will be used until the next evaluation. This is not done very frequently, perhaps every third year. An evaluation should always be made when the survey design is substantially changed or when the generic tool is developed with new functionality that might improve the efficiency of the selective editing.

5 The new generic editing tools

5.1 Expectations on new methods and generic tools

We think that common generic tools would be the best way to get acceptance for the change of methods. These are the expectations for the editing process:

- Standardized, common, generic tools lead to:
  - Less maintenance of IT systems, reducing a large cost for the NSI.
  - Easier planning of the manpower-demanding editing work for the total set of business surveys as individuals of the staff can work with several surveys when they are well acquainted with the tools.
  - Better work environment for the editing staff, when being familiar with an efficient tool.
  - Methodology studies are facilitated; studies of methods are possible when pre-requisites are comparable.
- Efficient editing methods (selective editing, significance editing) lead to:
  - Smaller volumes of follow-up, cheaper for Statistics Sweden and a smaller burden for respondents.
  - Better work environment for editing staff, not having to re-contact so many respondents that consider their delivered data to be correct. This is so when high hit rates of edit checks is a quality of the process.
- Structured collection and analysis of process data lead to:
  - Systematic improvement of data collection.
  - More efficient application of the editing methods and tools.
  - Better quality in statistics.
  - Information for quality declaration of statistics.

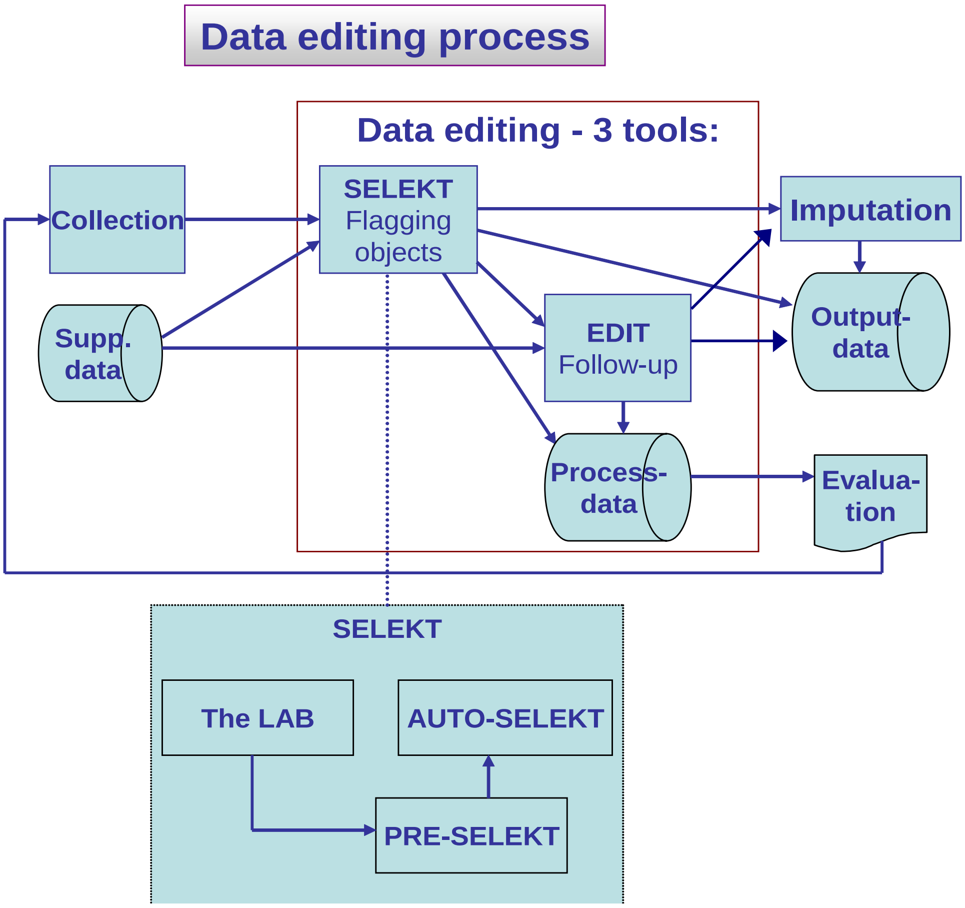
5.2 SELEKT and EDIT

The new general editing tool will be comprised of two main modules; SELEKT and EDIT. The method and the IT tools for flagging of incorrect or suspect data values through selective/ significance editing is called SELEKT. Primary selected units, observations and variables are flagged to go to manual follow-up, imputation or acceptance by SELEKT.

The specific survey conditions must be stated, such as measurement variables, domains of study and unit identifiers. These and other necessary parameters are set with the boundary intersection PRE-SELEKT for each current survey and are overhauled regularly by a process- and system-expert. The parameter values are stored in a table.

AUTO-SELEKT is a self-contained module that calculate according to the settings made in PRE-SELEKT, by reading the parameter table. AUTO-SELEKT codes all incoming units for acceptance, automatic imputation or follow-up. The units coded for follow-up with the corresponding information about the local and global scores will be presented to the editing staff in the EDIT module.

A so called laboratory environment, LAB, is a third production tool box in the SELEKT. The LAB will be used before implementation in surveys. The LAB-modules are used in order to evaluate the earlier production rounds to find best values on parameters, i.e. threshold values etc. Notice that access to unedited and edited data regarding the same time period is necessary when using the LAB. One way to make this possible is to draw and examine samples of suspected units lying below the global threshold value. The LAB is always used when implementing the general editing tool in a survey. After a successful implementation it is recommended to use the LAB to make evaluations at a regular basis. To a large extent the code for AUTO-SELEKT is used, but the LAB requires some extra functionality.



6 References

Adolfsson, C. and Gidlund, P. (2008): Conducted case studies at Statistics Sweden, *Paper presented at the UNECE Editing workshop in Vienna, April 2008*

Adolfsson, C. and Gidlund, P. (2009): Evaluation of selective editing in SLP 2008, *Paper presented at the UNECE Work session on statistical editing in Neuchâtel, Switzerland, October 2009*

Farwell, K. and Raine, M. (2000): Some Current Approaches to Editing in the ABS, *Invited paper to the International Conference on Establishment Surveys (ICES II), June 19-21, 2000 in Buffalo N.Y.*

Granquist, L. (1995). Improving the Traditional Editing Process. *In Business Survey Methods, eds. Cox et.al., Wiley*

Granquist, L. (1997) “The New View on Editing”. International Statistical Review

Granquist, L. and Kovar, J. (1997): Editing of survey data: How much is enough?, *In Survey measurement and process quality (p. 415-435) eds. Lyberg et al., Wiley*

Hartwig, P. (2009) “How to use edit staff debriefings in questionnaire design”, paper presented
at 2009 European Establishment Statistics Workshop, Stockholm.

Hedlin, D. (2009): Theory of Selective Editing. Invited paper, European Economic Survey Workshop, Stockholm, 7-9 Sept.

Ilves, K. and Laitila, T. (to be published): Probability Sampling Approach to Editing

Jäder, A. and Norberg, A. (2006): A Selective Editing Method considering both Suspicion and Potential Impact, developed and applied to the Swedish Foreign Trade Statistics, *Background facts on Economic Statistics 2006:3, Statistics Sweden*

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

Lawrence, D. and McDavitt, C. (1994): Significance Editing in the Australian Survey of Average Weekly Earnings, *Journal of Official Statistics, Vol. 10, pp. 437-447*

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

SCB (2002): Guide till granskning, *handbook in the series Current Best Methods from Statistics Sweden, CBM 2002:1. Statistics Sweden 2002*

SCB (2005): Rapport från undersökningen om granskningen i SCB:s statistikproduktion, *Working paper from Statistics Sweden 2005-03-22*

SCB (2007): Granskning Fallstudier (Nikolaus), *Project report from Statistics Sweden 2007-05-29.*

Särndal, C-E., Swensson, B. and Wretman, J. (1992): Model assisted Survey Sampling, *Springer, New York, 1992.*

Appendices

a) List of potential methodology development

i. New respondents

In the case of new units in a sample, there is a lack of access to equally satisfactory predicted values as for units that have been involved previously. This means that the computed potential impact will often be relatively high, and that more re-contacts will be made at the same threshold value. The measure of suspicion can work against this, so that a score that considers both suspicion and potential impact will be of the same magnitude as for other units. Special selections of new respondents can therefore be motivated.

ii. Variation in predictive quality

We have introduced parameters *VIOLIN*$_j$ for the relative importance of each variable. A possible use for these parameters is to adjust the score functions to variation in how well the predicted values function for the variables. We have, so far, too little empirical knowledge on how this can be used.

iii. Correlated variables

Some variables included in the score function may be strongly correlated. In the survey Short-term statistics, wages and salaries in the private sector this is a matter of a derived variable that is the sum of its component. What effect this has on the result of prioritising units depends, inter alia, on how global scores are constructed. Summing of local scores may have greater importance than using maximum local scores.

iv. The denominator in RPB

When the parameter of interest is a sum of a variable Y_j, where $y_{j,k,l} > 0$, for a domain of study *d*, it is meaningful to imagine the potential error $Potimp_{d,j,k,l} = w_k \cdot w_{k,l} \cdot (y_{j,k,l} - \tilde{y}_{j,k,l})$ relative to the estimated total. The *CELLO:s* support this choice as well as relating the potential error to the standard error of the estimated total. But when $y_{j,k,l}$ can take any real value the estimated total can be close to zero and it is not meaningful. This is also the case when the parameter is a ratio $\theta_{d,j1/j2} = \dfrac{T_{d,j1}}{T_{d,j2}}$. What is the alternative to standard error of the estimate, $\hat{T}_{d,j}$ and $\hat{\theta}_{d,j1/j2}$ in the denominator? An alternative is necessary when the standard error is zero.

v. Aggregated suspicion measures

A suspicion measure is primarily defined for each edit, i.e. suspicion $Susp_{k,l}^{(t_r)}$ is assigned for the test variable $t_{r,k,l}$. A suspicion measure is then assigned for each data value $y_{j,k,l}$ involved in $t_{r,k,l}$ as the maximum over *r* of all $Susp_{k,l}^{(t_r)}$ for edits where data item $y_{j,k,l}$ has been involved. A questions is if maximum always is the optimal function?

vi. Evaluation of old survey data

In the case of implementation of a new editing method, most often one has edited and un-edited data produced by a poor editing method. A natural question is how the search and design of a better method is affected of inadequate edited data. An alternative is to simulate errors for a file.

b) Notations and specifications of all parameters for an example

The use of all the parameters in the Selekt-methodology is demonstrated here with a fictitious example.

A survey is conducted each quarter. We have a stratified sample of enterprises. Strata are constructed by *Line of business* and *Size of enterprise* measured by number of employees in classes. The sampling fraction varies generally by size of stratum. Samples are rotated annually by 20 percent.

For each selected enterprise information on all the employees is collected. Background information on the employees are *Occupation* and *Gender*. We measure quarterly: *Hours worked*, *Salary* and *Tax*. We also have register information on Tax.

In this cut of data there are two enterprises from two different strata. The first enterprise (11001) has three employees, one male and two female persons and with two different *Occupations*. The second enterprise (12077) has six employees.

There are two fatal errors; a missing value on wage and a suspected 10 times too high wage.

Stra- tum	Pop _N	Sample _n	Enter- prise	Indu- stry	Re- gion	Employee	Profes sion	Gen- der	Year	Quar- ter	In- Hours	InSalary	InTax	Register- Tax	Hours	Wage	Tax
A3	185	35	11001	B	2	2000001	2	M	2008	2	806	159 042	94 528	65 658	622	290807	86923
A3	185	35	11001	B	2	2000002	2	K	2008	2	425		70 788	70 788	425	183684	70788
A3	185	35	11001	B	2	2000003	3	K	2008	2	484	272 493	96 126	96 126	484	375898	96126
A4	136	42	12077	A	6	2000001	2	K	2008	2	565	224 119	87 778	87 778	565	224119	87778
A4	136	42	12077	A	6	2000002	4	M	2008	2	585	689 122	201 365	201 365	585	689122	201365
A4	136	42	12077	A	6	2000003	3	M	2008	2	586	478 684	161 182	161 182	586	478684	161182
A4	136	42	12077	A	6	2000004	3	K	2008	2	495	451 495	93 696	68 215	495	273147	68215
A4	136	42	12077	A	6	2000005	3	K	2008	2	489	3 868 400	60 720	95 796	489	386840	95796
A4	136	42	12077	A	6	2000006	3	K	2008	2	621	412 779	115 469	115 469	476	412779	115469

The output consists of two presentation tables, in total with 5 different classifications $(c = 1 - 5)$:

Line of business	
A	c=1
B	c=1
C	c=1
D	c=1
E	c=1
F+G	c=1
Total	c=2

c=1 by Line of business
c=2 total (all employees)
c=3 by Occupation
c=4 by Gender
c=5 by Occupation and Gender

	Gender		
Occupation	F	M	Total
1	c=5	c=5	c=3
2	c=5	c=5	c=3
3	c=5	c=5	c=3
4	c=5	c=5	c=3
Total	c=4	c=4	c=2

General parameters for specification of survey design, cold-deck data etc.

ource	arameter	hort name	efault alue	xample for data bove
Design	Identification variable of respondent	Id2_var		Enterprise
Design	Identification variable of primary selected unit, PSU	Id2_var		Enterprise
Design	Id-variable of observation	Id3_var		Emp
Design	Weight variable for PSU or respondent (w_k)	Id2_wgt	1	Pop_N/ sample_n
Design	Weight variable for observation (w_{kl})	Id3_wgt	1	1
Design	Year-variable	T1_var		Year
Design	Short-period-variable	T2_var		Quarter
Manager	Name of data to be edited	EDIT_Data		
Manager	Year of data to be edited	EDIT_T1_value		2009
Manager	Value of short-period-variable for data to be edited	EDIT_T2_value		1
	old-deck-data to be used for computing predicted values and ispersion in a time-series analysis:			
Lab	Name of data	TS_Hist_Data		
Lab	Year of low limit	TS_T1_Start_Value		2005
Lab	Value of short-period-variable for low limit	TS_T2_Start_Value		1
Lab	Year of high limit	TS_T1_End_Value		2008
Lab	Value of short-period-variable for high limit	TS_T2_End_Value		4
Lab	Required number of records for computing statistics	TS_Min_obs		12
Lab	Measure of predicted value based on time-series data	TS_Expect		Forecast
Lab	Measure of dispersion value based on time-series data	TS_Spread		Forecast
	old-deck-data to be used for computing predicted values and ispersion in a cross-section analysis:			
Lab	Name of data	CS_Hist_Data		
Lab	Year of low limit	CS_T1_Start_Value		2007
Lab	Value of short-period-variable for low limit	CS_T2_Start_Value		1
Lab	Year of high limit	CS_T1_End_Value		2008
Lab	Value of short-period-variable for high limit	CS_T2_End_Value		4
Lab	Required number of records for computing statistics	CS_Min_Obs		5
Lab	Measure of predicted value based on cross-sectional data	CS_Expect		MEDIAN
Lab	Measure of dispersion value based on cross-sectional data	CS_Spread		QUARTILES
	ny number of variables can be used to define homogeneous roups in an hierarchical order:			
Lab	First variable in the definition of homogeneous groups	HG_Var1		Last_Occupation
Lab	Number of digits in the first variable to be used	HG_Dig1		1
Lab	Second variable in the definition of homogeneous groups	HG_Var 2		Last_Gender
Lab	Number of digits in the first variable to be used	HG_Dig2		1
Lab	Third variable in the definition of homogeneous groups	HG_Var 3		Quarter
Lab	Number of digits in the first variable to be used	HG_Dig3		1
Lab	Etc.			
	ther parameters used in the computation of suspicion:			
Lab	Defining the range where suspicion is zero	KAPPA	1	1
Lab	Sets a minimum of dispersion	SUSP_ALFA	0,05	0,02
Lab	Defining the shape of the suspicion to deviation	TAU	1	1
	arameters used in the computation of impact:			
Manager	Parameter to steer the score for non-response	NONRESP_IMPACT	1	1
Lab	Parameter setting the proportion of estimated total contra estimated standard error of estimated total in denominator of score	CELLO_ALFA		
	ggregations of local scores to global scores, thresholds:			
Lab	Parameter to choose the sum, sum of squares or maximum function in the aggregation of classifications (*c*)	LAMBDA5	1 (sum)	1
Lab	Parameter to choose the sum, sum of squares or maximum function in the aggregation of variables (*j*)	LAMBDA4	1 (sum)	1
Lab	Parameter to choose the sum, sum of squares or maximum function in the aggregation of observations (*l*)	LAMBDA3	1 (sum)	1
Lab	*Parameter to choose the sum, sum of squares or maximum*	*LAMBDA2*	*1 (sum)*	*Not availabe*

	function in the aggregation of PSU (k), not available			
Lab	Threshold for local scores by classifications	THRESHOLD5	0	0
Lab	Threshold for local scores by variables	THRESHOLD4	0	0,1
Lab	Threshold for local scores by observations	THRESHOLD3	0	0,2
Lab	Global threshold for PSU	THRESHOLD2		1
Lab	*Global threshold for respondent, not available in SELEKT 1.0*	*THRESHOLD1*		*Not available*

Importance of classifications; parameters *CLARINET*$_{c(d)}$ *c* (c = 1, … ,5)

Source	**Classification**	**Classification (c)**	**Default value**	**Example for data above**
Manager/Lab	Line of business	1	1	1
Manager/Lab	Total	2	1	0,5
Manager/Lab	Gender	3	1	1
Manager/Lab	Occupation	4	1	1
Manager/Lab	Gender x Occupation	5	1	2

Survey variable[24] specific information and parameters

Source	**Survey variable number** (automatic number in SELEKT)	**Variable**	**Short name for edited variable**	**Short name for unedited variable**	**Short name for edited auxiliary variable**	**Short name for edited auxiliary variable**
Manager/ MetaPlus[25]	1	Hours worked	Last_Hours	First_Hours	Last_ATime	Last_ATime
Manager/ MetaPlus	2	Income of labour	Last_Salary	First_Salary	Last_Hours	Last_Hours
Manager/ MetaPlus	3	Tax paid by the employer	Last_Tax	First_Tax	1	1

Cont.

Source	**Survey variable number** (automatic number in SELEKT)	***VIOLIN*$_i$** Importance parameter for variables	***OBOE*$_i$** Importance parameter for the relative size of a domain of study
Manager/ MetaPlus	1	1,0	0,6
Manager/ MetaPlus	2	1,0	0,6
Manager/ MetaPlus	3	1,0	0,6

[24] Can be a derived variable, meaning variables in output

[25] Metaplus is the database/system of metadata at Statistics Sweden

Example of possible test variables

Survey variable number (automatic number in SELEKT)	**Test variable, edited**	**Test variable, un-edited**	**Type of edit check**	**Flag**	**Suspicion**
		First_Occupation	Non-valid code	T	1 for the observation, i.e. No selective procedure, the observation must be on the flag-list
1		First_Gender	Non-valid code	2	1 for all survey variables. Selective editing decides if the observation shall be on flag-list
1		First_Hours	<0 is a non-response, i.e. a fatal error	1	1 for variable 1
		First_Tax/ First_Salary	>1 are highly implausible values	0,70	Be experience, the hit rate for this edit check is 70%, so we flag accordingly
1	Last_Hours/ Pre_Hours	First_Hours/ Pre_Hours	Continuous measure of suspicion by SELEKT.	C	0-1 for variable 1

Ratios of survey variable

Source	**Survey variable number of numerator** (Number in SELEKT)	**Survey variable number of denominator** (Number in SELEKT)	***$QVIOLIN_i$*** Importance parameter for ratios	***$QOBOE_i$*** Importance parameter for the relative size of a domain of study
Manager/ MetaPlus	2	1	1,0	1,0
Manager/ MetaPlus	3	2	1,0	1,0

Register variables

Source	**Short name**	**Variable**
Manager/MetaPlus	Register_Tax	Tax paid by the employer

Traditional edits resulting in suspicion=1 for the observation (record)

Source	**Edit**
Manager/Lab	InSalary<InTax

Fatal and query edits resulting in suspicion = 1 for specified survey variables

Source	**Edit**	**To be use for survey variable;**
Manager/Lab	InHours <0	1
Manager/Lab	InHours >3000	1
Manager/Lab	InSalary<0	2
Manager/Lab	InSalary/InHours >10 000	1,2
Manager/Lab	InTax<0	3

Test variables for which continuous suspicion measures are computed

Source	Test variable number (in AUTO-SELEKT)	Definition	To be use for survey variable;
Manager/Lab	1	InHours	1
Manager/Lab	2	InSalary	2
Manager/Lab	3	InTax	3
Manager/Lab	4	InSalary/ InHours	1,2
Manager/Lab	5	InTax/InSalary	2,3
Manager/Lab	6	InTax-RegistrTax	3

The *CELLO* matrix produced by PRE-SELEKT:

				Variable number (*j*)			Ratio number	
c	Classification	Domain of study (*d*)	Condition	1	2	3	1	2
1	Industry by SNI 2002	1	Industry=’A’	$CELLO_{1,1}$	$CELLO_{1,2}$	$CELLO_{1,3}$		
1	-“-	2	Industry=’B’	$CELLO_{2,1}$				
1	-“-	3	Industry=’C’					
1	-“-	4	Industry=’D’					
1	-“-	5	Industry=’E’					
1	-“-	6	Industry=’F’ or					
2	Total	7	‘A’<=Industry<=’Q’					
3	Occupation by nomenclature	8	Occupation=’1’					
3	-“-	9	Occupation=’2’					
3	-“-	10	Occupation=’3’					
3	-“-	11	Occupation=’4’					
4	Gender (F/M)	12	Gender=’K’					
4	-“-	13	Gender=’M’		$CELLO_{24,2}$			
5	Prof by nomenclature and Gender	14	Occupation=’1’ and Gender=’K’					
5	-“-	15	Occupation=’2’ and Gender=’K’					
5	-“-	16	Occupation=’3’ and Gender=’K’					
5	-“-	17	Occupation=’4’ and Gender=’K’					
5	-“-	18	Occupation=’1’ and Gender=’M’					
5	-“-	19	Occupation=’2’ and Gender=’M’					
5	-“-	20	Occupation=’3’ and Gender=’M’					
5	-“-	21	Occupation=’4’ and Gender=’M’					

c) Description of the UCM procedure in SAS[26]

By Jahnavi Wallin, Statistics Sweden

UCM in SAS stands for Unobserved Components Model; the procedure consists of a time-series analysis that decomposes the time series into trend, cycle and season. An autoregressive term and lags for the time series are also used, and regression variables can be incorporated into the model. The method is employed for univariate time-series analysis in order, for example, to generate forecasts for the time series or to obtain a seasonally cleaned series. The components of UCM provide a description of the underlying mechanisms that govern the time series.

A UCM can be described as follows:

$$y_t = \mu_t + \gamma_t + \psi_t + \tau_t + \sum_{i=1}^{p} \phi_i y_{t-1} + \sum_{j=1}^{m} \beta_j x_{jt} + \varepsilon_t \ ,$$

where y_t describes the observed outcome for a certain variable at the different points in time t = 1,2,…,T. The term μ_t represents the trend, γ_t the season, ψ_t the cycle, and τ_t is an autoregressive component[27] The regression term, $\sum_{j=1}^{m} \beta_j x_{jt} + \varepsilon_t$, includes variables that are defined in the input data. Lags of the dependent variable are also included in the model through $\sum_{i=1}^{p} \phi_i y_{t-1}$.

The error term, ε_t, also called the irregular component, is assumed to be white noise with variance σ_ε^2. The trend refers to the long-term development of the time series and is based on structural changes to the underlying data-generating factors, while the cycle consists of more short-term transitional change. If a time series is measured over several time periods in a year, such as months or quarters, it is common for the series to display systematic seasonal variation. Exogenous effects in a time series can be modelled by including regression variables in the model. The various components are assumed to be independent of each other and independent of the irregular component.

By controlling for the presence of the various terms and specifying as accurate a model as possible, UCM can generate widely diverse types of time-series patterns. UCM can also be used after the variables have been transformed, e.g. logarithmically or via differentiation.

[26] SAS Help, proc UCM Overview, and An Introduction to Unobserved Component Models.

[27] An autoregressive term is defined as $y_t = \phi_1 y_{t-1} + ... + \phi_p y_{t-p} + a_t$, where the error terms a_t are independent and identically distributed, i.i.d. $(0, \sigma_a^2)$, and where the estimated ϕ coefficients describe the nature of the dependence on previous periods.