

Feasibility Study: A Review of Selective Editing

# Technical Report



**Under contract for Official Statistics Research and Data  
Archive Centre (OSRDAC), Research Programme Services,  
Statistics New Zealand**

**Prepared by Carl Scarrott,  
Mathematics and Statistics Department,  
University of Canterbury**

**Assisted by Vera Costa and team from  
Edit and Imputation Network, Statistics New Zealand.**

**Contract with Canterprise Ltd.**

The views presented in this report are  
those of the author and do not necessarily represent those of  
Statistics New Zealand or the University of Canterbury

**Correspondence address:  
Dr Carl Scarrott  
Mathematics and Statistics Department  
University of Canterbury  
Private Bag 4800  
Christchurch  
New Zealand**

**[carl.scarrott@canterbury.ac.nz](mailto:carl.scarrott@canterbury.ac.nz)**

**Tel. +64 (0)3 364 2987 ext 8338  
Fax +64 (0)3 364 2587**

## **Abstract**

This report forms part of a feasibility study towards the implementation of selective editing across Statistics NZ and the Official Statistics System, under an OSRDAC research programme. It provides a review of the published literature on selective editing methodologies and key aspects of current international best practice. The key features and cost benefits of the approach are detailed along with discussion of some implementation issues. As a preparatory stage for the development of an implementation strategy, this review goes on to outline a broad selective editing framework which is intended to be general enough to apply across Statistics NZ surveys. A number of recommendations have been made based on the evidence provided in the literature review, along with some suggested foundational and research questions for future consideration.

**Keywords:** selective editing, quality improvement, performance monitoring

## **Acknowledgements**

I wish to thank Richard Penny and Kimberley Cullen for involving me in this project, which has been very stimulating. The guidance and advice of Vera Costa throughout this project is gratefully acknowledged. I would also like to thank the Overseas Trade and Quarterly Employment Survey teams for their insight and expertise which helped to highlight important topics to be covered in this report. The support of the University of Canterbury Mathematics and Statistics Department and John Duncan from Canterprise are acknowledged. The feedback provided by the reviewer was also very useful in finalising this report.

## TABLE OF CONTENTS

<b>Title Page</b>	.....	<b>1</b>
<b>Abstract and Acknowledgments</b>	.....	<b>2</b>
<b>Table of Contents</b>	.....	<b>3</b>
<b>Terminology</b>	.....	<b>4</b>
<b>1. Introduction</b>	.....	<b>6</b>
1.1. Structure of Report	.....	<b>7</b>
<b>2. Background</b>	.....	<b>8</b>
2.1. Historical Perspective	.....	<b>8</b>
2.2. Motivation	.....	<b>10</b>
<b>3. Selective Editing</b>	.....	<b>12</b>
3.1. What Are Key Survey Items/Variables?	.....	<b>12</b>
3.2. Edit Rules	.....	<b>13</b>
3.3. Editing Prior to Selective Editing	.....	<b>14</b>
3.4. Focus on Influential Errors	.....	<b>15</b>
3.5. Score Functions	.....	<b>16</b>
3.6. Local Score Functions	.....	<b>18</b>
3.7. Global Score Functions	.....	<b>19</b>
3.8. Threshold Selection	.....	<b>20</b>
3.9. Performance Assessment and Quality Improvement	.....	<b>23</b>
<b>4. Some Implementation Issues</b>	.....	<b>25</b>
4.1. Discrete Variables	.....	<b>25</b>
4.2. Iteration of Selective Editing	.....	<b>26</b>
4.3. Streaming	.....	<b>26</b>
4.4. Local Score Functions	.....	<b>26</b>
4.5. Global Score Functions	.....	<b>27</b>
4.6. Threshold Selection	.....	<b>29</b>
<b>5. Key Principles and General Framework for Implementation</b>	.....	<b>31</b>
5.1. Key Principles of Selective Editing	.....	<b>31</b>
5.2. General Framework	.....	<b>32</b>
5.3. Selective Editing Framework	.....	<b>37</b>
<b>6. Recommendations for Future Consideration</b>	.....	<b>39</b>
6.1. Practical Implementation	.....	<b>39</b>
6.2. Foundational Issues	.....	<b>40</b>
6.3. Research Questions	.....	<b>40</b>
<b>References</b>	.....	<b>41</b>

## Terminology

The terminologies used in published works on this topic are often inconsistent. Therefore, the key terminology used in this review is made explicit below:

<i>Term</i>	<i>Definition</i>
<b>Survey</b>	process by which data is collected, e.g. survey questionnaire or administrative return
<b>Unit/Provider</b>	an entity that supplies data for a survey
<b>Variable</b>	output (e.g. statistic) derived from a survey
<b>Item</b>	a single entry in a survey, e.g. question or return
<b>Unit record</b>	entire survey return provided by unit
<b>Response</b>	data provided for a single item in a survey
<b>Outcome</b>	an output from a survey, e.g. survey variables or individual unit responses
<b>Cell/domain</b>	a level of aggregation across units (e.g. by region, industry, HS code)
<b>Non-response</b>	unit fails to respond to survey
<b>Partial non-response</b>	unit fails to respond to some survey items
<b>Editing</b>	process by which possible and certain errors in responses are identified and streamed for amendment/imputation
<b>Amendment</b>	procedure (usually manual) to adjust identified errors, e.g. by manual recontact or review
<b>Imputation</b>	procedure (usually automatic) for adjusting unit responses which are either missing or possibly in error
<b>Edit rule</b>	condition used to detect missing, invalid and inconsistent unit responses, sometimes called edits
<b>Edit model</b>	culmination of all edit rules, indicating acceptance region for responses
<b>Fatal edit rules</b>	indicate responses which are certainly in error, e.g. a failed range check (age must be positive value). May include (partial) non-response

<b>Query edit rules</b>	indicate responses which are possibly in error
<b>Micro-editing</b>	editing towards validity or consistency of individual unit responses, based on only current unit's response, plus maybe auxiliary information, but not usually another unit's responses
<b>Macro-editing</b>	editing towards validity or consistency of individual unit responses, based on the responses of a number of units. For example, checks on survey variables over the whole database of records or over different domains
<b>Input editing</b>	editing carried out at data collection, input and processing stages, often same as micro-editing
<b>Output editing</b>	editing carried out after data processing stage, based on analysis of many units/responses, often same as macro-editing
<b>Expected amended response</b>	estimate of item response not necessarily correct or final (amended/imputed) value
<b>Expected amended variable</b>	estimate of survey variable, not necessarily correct or final (amendment/imputation) value
<b>Local score</b>	a quantitative measure used to prioritise a unit response for editing
<b>Global score</b>	a quantitative measure used to prioritise a entire unit record for editing
<b>Hit rate</b>	proportion of responses considered for editing that were found to be in error
<b>Selective editing</b>	procedure by which the amendment/imputation of responses which are likely to be in error and have impact on survey outcomes are prioritised
<b>Significance editing</b>	selective editing where prioritization is based on the estimation of impact of correcting the error on the survey variables

## 1. Introduction

This document forms part of a feasibility study towards the implementation of selective editing across official statistics agencies, under an OSRDAC research programme. The OSRDAC Research Programme has been established to affect a fundamental shift in the way New Zealand's official statistics are collected, disseminated and used. The centre is one component of a cross-state sector initiative to strengthen official statistics to obtain the most value from Government's investment in statistical activity.

Improving the efficiency in the editing of survey, census and administrative data is one of the key statistical issues that Statistics NZ (SNZ) is currently facing. Effort needs to be focussed on errors that are obvious to users and/or have a significant impact on the survey outcomes. To this end, the organisation needs to consider implementing efficient editing techniques, such as selective editing, where these are not already in place.

The feasibility study aims to consider the key international best practice in selective editing and its applicability to SNZ, and across the Official Statistics System (OSS). The study consists of three projects running in tandem; a broad scale but focused review of the published literature on selective editing methodologies and key aspects of current international best practice (detailed in this report), and two case studies on an implementation for Overseas Trade and Quarterly Employment Survey data.

This review document focuses on the broad principles underlying the methodology, rather than specific details about particular implementations (e.g. comparisons of score functions is not undertaken). The motivation for implementation of selective editing and the key costs/benefits are outlined. The key features of the approach are detailed and some implementation issues discussed with potential resolutions where available. The review outlines a broad framework for implementation which is hoped to be sufficiently general to apply across SNZ surveys. The framework provides for selective editing to be integrated alongside alternative (usually existing) editing processes in a cohesive manner.

The two case studies provide datasets of quite different data types and structure, with very differing issues which impact implementation. They aim to provide a valuable insight into the sorts of issues involved in more widespread implementation. Both of these case studies have historical data available to them (which are reasonably temporally consistent with good sample frame overlap). Hence, they will not expose issues associated with one-off surveys or where the responses are harder to predict/impute. However, these issues are discussed and resolutions, where available, are detailed.

The time-scales available for this project limit consideration of many aspects of selective editing and it's interaction with related fields. In particular, this review does not aim to compare selective editing to alternative methodologies. At this stage it is inappropriate for the author to consider the practical/strategic aspects associated with implementation across SNZ and there is no discussion of transitional implementation issues. The issue of imputation, which is often sensibly integrated in a generalized editing and imputation framework (Granquist, 1995), is not considered. It is assumed that once the editing strategy has been defined, and if imputation is required, that one of the many off-the-shelf imputation procedures can be implemented. The availability and applicability of software/database systems for implementation have not been considered. The limited time-scale of this review has also not allowed a review of graphical editing techniques.

## **1.1. Structure of Report**

Section 2 provides a short historical perspective and motivation for selective editing. The selective editing methodology is outlined in some detail in Section 3, along with its key benefits and drawbacks. Some of the main issues to be considered when implementing the methodology across surveys are outlined in Section 4, with suggested resolutions where applicable. The main body of the report is then rounded up with an outline of the general framework for implementation across Statistics NZ surveys in Section 5. A number of specific recommendations for issues to be addresses during any future implementation are outlined in Section 6. A list of references used in the report and some supplementary material related to the selective editing are provided.

## 2. Background

The Knowledge Base on Statistical Data Editing (K-Base) define editing as the “process of detecting and handling errors in data”, which consists of 3 phrases:

- definition of a consistent system of requirements,
- their verification on given data, and
- elimination/substitution of data in contradiction of defined requirements.

The aim is usually to provide, in a timely manner, a dataset which has reached a desired quality standard, in terms of completeness and correctness, making best use of resources.

Editing can broadly speaking be broken down into two relatively distinct types. Firstly, micro-editing (often called input editing) focused on detecting and adjusting data at the unit response level. It is often desirable for micro-editing of a unit response to be independent of other units within the current survey cycle. The micro-edited data are typically then passed on to a macro-editing (often called output editing) process, focused on errors which impact on statistics produced at some level of aggregation across a number of units. Macro-editing is normally carried out at the domain levels by which the survey outcomes are reported and will often look for errors which cannot be detected at the micro unit response level. Selective editing can bridge the gap between micro and macro-editing, as unit responses are subject to editing (suggesting micro) but a common aim is to prioritize errors which impact on the survey outcomes (suggesting macro).

Edit rules are usually defined as logical conditions to check whether a unit response is as expected, missing, in error or a suspected outlier. The combination of all the edit rules (often just called edits), is called the edit model. The edit model defines a form of acceptance region for unit responses. A unit response that fails at least one of the edit rules is said to be outside of the current edit model. It is quite common for edits to be categorised into two types, e.g. fatal edits (validity, consistency, range checks, etc.) and query edits (indicating suspicious responses). There are many alternative (finer scale) edit classifications which are discussed by Granquist (1995).

### 2.1. Historical Perspective

The traditional view of editing, where more editing should lead to improvements in data quality, has been shown by numerous studies to be incorrect (Granquist, 1995 and 1997a). Editors had long believed that more extensive and detailed edits should lead to cleaner data and therefore better quality survey outputs. However, the traditional editing process, where all edit failures are re-visited (usually by manual amendment) cannot be justified on the grounds of improved data quality for a number of reasons, including:

- Uneconomic – cost in terms of work-hours not justifiable for benefits gained
- Timeliness – delays in publication
- Overediting – data being edited to such an extent that bias and other errors are introduced, e.g. data changed to fit edit model (known as “creative editing”)
- Diminishing quality improvements – largest improvement in survey outcomes often achieved by small number of changes to highly influential units/responses, with rest of effort providing little extra improvement.

There are also many intangible costs to this overediting, for example respondent burden (in the case of recontact). Further, the amount of resources put into traditional editing often provides users of the survey data with undue confidence in its quality. These issues

are discussed in detail in the excellent reviews of editing practice by Granquist and Kovar (1997) and Granquist (1995 and 1997b), and reference therein.

In the traditional editing process, wherever possible, all units outside of the edit model are either manually recontacted or at least manually reviewed, to confirm or amend the response data. It is rather crudely assumed in this report, as is often the case, that the responses will be correct after manual recontact.

Engström and Granquist (1999) suggested that it is common for the first 10-15% of the most influential traditional edits to provide 90% of total improvement in survey variable estimates, with the first 5-10% typically bringing the estimate within 1% of final published value. Further, the hit rates for traditional edits are typically as low as 20-30%. Given this level of performance and that editing often takes up around 20-40% of total survey costs (excluding intangible costs like respondent burden, timeliness, etc), see United Nations Economic Commission for Europe (UNECE, 2000) and the UK Office of National Statistics (ONS, 2004), it is not surprising that traditional edit processes are often considered uneconomic in terms of the quality improvements that they attain.

These issues were highlighted particularly well in the World Fertility Survey (WFS), see Pullum *et al.* (1986) for a full discussion, where publication was delayed by about a year due to the extensive editing that was carried out, with a post-survey analysis showing little of the editing provided substantial gains in survey outcomes.

The report on the WFS was also critical of the lack of consistency in the level and detail of editing being carried out in several countries. The WFS experience also suggests that the edit specifications/guidelines should be pre-defined as part of the analysis plan and they should be consistent with survey objectives, with editing staff well trained, motivated and organised. Across a large organisation, involved in many different surveys, it is often desirable to define a general editing framework to ensure some level of consistency in the editing choices and procedures that are undertaken, which can lead to comparative quality assurance standards for survey outcomes. For example, both Statistics Canada (2003) and the Australian Bureau of Statistics (Williams *et al.*, 2000) have collated guidelines for their editing and imputation processes. Further, for periodic surveys where the change in survey variables over time is important, the consistency in the editing process can be just as important as concerns over accuracy.

The lack of formalized procedures for data editing can lead to ad-hoc decisions being made which can lead to bias in survey variables. For example, manual data editing can lead to data being changed to fit an edit model (so called creative editing), or can be slow to respond to actual changes in unit circumstances. Formalization is a valuable feature, but so is transparency. Transparency can allow for future re-evaluation of survey variables or unit responses and the edits applied to them. If database systems allow, there is also opportunity to rollback editing, and evaluation of the impact of any new editing strategies on the past survey variables/unit responses. End users may also benefit from a transparent editing system, so as to understand the quality level of the survey and its limitations.

## 2.2. Motivation

The “New View on Editing” proposed by Granquist (1997b) is to move away from focusing effort on cleaning data as much as possible to improve quality, to targeting editing resources on the responses which have the greatest influence on the survey outcomes. In parallel, data on the edits undertaken should be collected, stored and analysed to highlight possible quality improvements to the editing process but and other parts of the whole survey process. Thus causing a paradigm shift from “error detection and correction” to “error prevention”. Essentially editing should be moved to as early in the survey process as possible.

In particular, the editing process should provide measures by which to monitor performance and to highlight common error types, sources and distributions, which can be feedback to improve processes like survey design, data collection, and data imaging/entry. Further, the collection of data on the edit process (from now on called edit data) provides for an audit trail and to produce quantitative measures of performance (in terms of efficiency and effectiveness) and diagnostics for process improvement.

The new view on editing essentially sees selective editing as one component in a total quality management (TQM) of surveys, and that the process should provide tools for continuous monitoring, evaluation and control. Linacre and Trewin (1989) discuss the issue of optimal allocation of resources across the whole survey process to minimize survey errors. However, this report is going to focus on resource allocation within the (selective) editing process.

One of the main aims of selective editing is to focus editing resources on the most important response errors, without substantively compromising the quality of the survey outcomes. Thus ensuring that the data is of sufficient quality (in terms of conformance and design) for the (present and future) end users, but does not waste resources on unimportant errors which could better be spent on improving quality in other parts of the survey process (e.g. more macro editing). The edit data can also provide valuable information to end users on what amendments/imputations have been carried out and the reasons why, potentially allowing them the opportunity to re-evaluate those decisions.

In general, selective editing allocates every unit response, which is possibly in error, a number (score), which measures the relative importance of that possible error on the survey outcomes. The score provides a way to prioritise the editing of unit responses, possibly leading to a ranking of the most important units to be subjected to amendment/imputation. Typically, selective editing is used to define a critical stream of edits (score above some threshold value), which must be completed (manually or otherwise), before publication of survey variables. The non-critical edits (score below the threshold) can then be either automatically edited (imputed) or even left unedited, whichever is appropriate. In this way, selective editing provides a general framework for allocating editing resources to units which have most impact on survey outcomes.

Most selective editing implementations require only the current unit's responses (plus auxiliary information from past surveys) on which to base the decision as to whether they need to go into the critical stream, as the threshold is usually known beforehand. Therefore, the editing can be initiated as soon as the data is received, which in time critical applications is important. The main drawbacks of using predetermined cut-offs in selective editing are that the editing workload can vary, as can the benefit achieved (Farwell, 2004).

The potential resource savings over traditional editing approaches can be large, but will depend on the survey. For example, Granquist (2005) gives examples of surveys where only 35-50% of recontacts are required after selective editing, and that the resource saved can often be 50% or more.

Selective editing has been shown to be very effective for continuous (quantitative) variables, particularly those with non-uniform (skewed) error distributions where a small number of errors which are large in magnitude have the largest impact on the quality of survey outcomes, Latouche and Berthelot (1992). Selective editing is mainly applied to quantitative responses as there are natural metrics for defining the magnitude of the response error and therefore its impact on the survey outcomes. However, there is often no natural metric for error magnitude for discrete (categorical, ordinal, etc.) data which are common in social surveys. Hence, generating a score for ranking is rather more challenging, or even impossible. Editing for this type of data is generally carried out using alternative often more traditional techniques, e.g. Fellegi and Holt (1976) edit and imputation. In Section 4 a suggested route for providing simple score for discrete responses is discussed.

The focus of selective editing is on prioritising the editing of unit response errors which are likely to have a large impact on the survey outcomes. There is no particular focus on achieving good unit records *per se*. If the quality of the individual unit records is a critical survey outcome then other micro editing techniques should be considered. Selective editing does not necessarily provide the editor with an indication of which is likely to be the incorrect item(s), i.e. the error localization problem, which is particularly important for surveys with many key items. Error localization tries to find the minimal set of edits that is required to ensure the unit record is within the edit model, to preserve as much of original data as possible. The integration of error localization within selective editing is still an open question and is beyond the scope of this review.

The following section provides an outline of the general selective editing methodology, which is followed up by a discussion of some of the main issues encountered upon wider implementation. In common with most editing routes, selective editing performs well for many types of survey (including administrative and census data) but particular characteristics which improve likelihood of success are: periodic surveys (or those with historical data), where the survey itself and the data input/processing do not change substantially over time, where there is a good sample frame overlap and when responses/variables are reasonably predictable.

### 3. Selective Editing

The concept of “selective editing” has been around for many years, with the one of the earliest published accounts of an application to official statistics by Latouche and Berthelot (1990, 1992). Many variants of selective editing have been proposed by a number of organisations sometimes under different names, e.g. “significance editing” from the Australian Bureau of Statistics (ABS - Lawrence and McDavitt, 1994 followed by Lawrence and McKenzie, 2000), “plausibility indicators” from Statistics Netherlands (Hoogland, 2002 and 2005).

However, most variants have the common goal of prioritising the editing of unit responses based on some form of scoring function. The score provides a measure of importance in terms of the need to allocate resources to edit the associated unit response. In general, the editing workload is focused on those unit responses which have failed edit rules and have the largest impact on the survey outcomes (usually key variables). This section provides an outline of the general selective editing methodology, followed by discussion of some of the key variants.

Selective editing has been implemented by many overseas official statistical agencies, including; UK Office of National Statistics (Hedlin, 2003); Australian Bureau of Statistics (Lawrence and McDavitt, 1994), Statistics Netherlands (Hoogland, 2002), Statistics Sweden (Granquist, 1995). Some of the best practice guidelines used at these organisations, discussed in published literature, are also outlined below.

The key stages in selective editing are typically:

1. Identification of key survey variables and items and user requirements
2. Identification of in-scope responses (error analysis, edit model, prior editing)
3. Scoring (at unit response and/or unit level)
4. Prioritisation (ranking of important errors using scores)
5. Threshold selection and streaming (which are the most important errors).

The following subsections provide an outline of these stages. Following the streaming of errors comes the amendment/imputation stage, which is not within the remit of this report. However, a key feature of best practice is that any amendment and imputation should be consistent with the edit model and scoring.

#### 3.1. What are key survey items/variables?

One of the first issues to be discussed when considering implementation of selective editing is what are the key variables and items of the survey, as these will be used to judge the quality performance of the survey to ensure it is meeting the needs of its end users. Some things to consider:

- i. How important are individual unit responses?
- ii. What are key survey variables?
- iii. What are the key items the variables are based on?
- iv. What are key domains (level of aggregation)?
- v. Which variables are poorly estimated (have high variance)?

The answers to these questions are often guided by the opinion of subject matter experts, including survey users, and analysis of the survey responses/errors. They have a direct impact on the score functions to be developed and give information on what edit rules are most important.

Some careful thought should be given to question i. Is it just the survey variables that are important? Or are there other (future) uses of the unit response data? What is the impact in user confidence of remaining errors? Should the individual responses at least be (automatically) edited to ensure all serious (fatal) edits are passed? Or will this lead to overediting?

If the key survey variables can be broken down into different levels of aggregation (domains), are all levels equally important? If units/responses are scored at different levels of aggregation, how should the edits be prioritised between/within levels? What should be done with small cells, should all edit failures be subject to amendment/imputation? Are the edit resources suitably distributed across cells?

The final question encapsulates issues like which variables are difficult to estimate, low response rates, small cells, etc. If a variable has a high variance, so is difficult to estimate, then it may need to receive higher priority.

### 3.2. Edit Rules

In many implementations of selective editing the edit model plays vital role in ensuring its effectiveness. In general, selective editing allocates a score to every unit response which has failed an edit rule. However, *an edit model is not necessarily required*, a score can be assigned to all unit responses. The score is then used to prioritise the editing of those responses. Errors outside of realm of edit model will therefore not be edited. Therefore, selective editing requires a good set of edit rules to work effectively.

Consistent with the selective editing methodology, the focus of the edit model should be to highlight all errors which have a serious impact on the quality of survey outcomes. A balance need to be struck to between the meticulously design of edits to detect all possible errors of interest and not being too onerous leading to detecting unimportant errors. The idea is define a minimal set of edit rules which define an acceptance region for responses. The edit rules must be accessible and transparent (all explicit and implicit conditions noted and recorded) to ensure they can be fully understood and their limitations acknowledged, in particular by future end users.

For a new survey the edit rules should be defined using subject matter expertise, thorough error analysis, feedback from other parts of survey process (primarily editing and imputation) and an understanding of user quality requirements. For existing (periodic) surveys, an edit model is usually available but should be regularly re-evaluated in light of possible change in editing quality focus to ensure it is still being effective, due to changes in population, survey design, data collection and input, etc.

In order to provide an audit trail for the edit process it is useful to record the edit model at the time the data is processed. Further, end users may also wish to know the edit rules applied to a survey, so this can be evaluated and the impact on their study assessed.

Edit rules are often written on a survey by survey basis. Williams *et al.* (2000) surveyed a number of large statistical agencies and found most of them had no general guidelines for development of an edit model. The main exceptions being Statistics Canada who have a comprehensive set of guidelines (Statistics Canada, 2003) and the ONS who have a

standards and quality assurance team which monitor editing practices and have developed BLAISE modules which encapsulate their standards.

In some survey situations it may be possible to define edit rules to capture problematic error types which would usually only be captured at the macro editing stage, e.g. inliers (see Winkler (1997) and Mazur (1990) for discussion and examples).

### 3.3. Editing Prior to Selective Editing

Although some score functions have been designed to overcome issues like partial or full non-response (e.g. Latouche and Berthelot ,1992) or robust to serious errors which have usually failed fatal edits (typically those that have failed scope tests, validity, consistency or range checks), it is fairly commonplace for some (if not all) of these types of edit failures to be subject to initial editing prior to selective editing, either automatically or by manual amendment (including keying staff on data entry). For example, Latouche and Berthelot (1990, 1992) followed up all non-response by manual recontact and did a minimum follow-up for partial non-response to ensure that the in-scope status of the units could be assured. Hoogland (2005) suggests obvious errors like values reported in \$'s rather than \$000's can often be automatically adjusted before further editing. Thompson and Hostetter (2001) state that the US Bureau of Census (USBC) manually rectify all fatal edits (e.g. partial non-response) before selective editing of their quinquennial economic census.

An interesting example of initial editing is provided by Lawrence (1999) for the ABS. For the Survey of Employment and Earnings all reporting units were allocated to hierarchical streams (definitions slightly modified for consistency):

1. Permanently nil and defunct units
2. New units
3. Continuing units within edit model
4. Continuing units which failed a fatal edit
5. Continuing units which failed a query edits, which is further split into:
  - a. Units where resolving query edit is expected to substantially impact variables
  - b. Otherwise.

Clearly, some units may satisfy the conditions for a number of these streams, if so, they will be allocated to the highest ranked stream. Streams 1 and 3 are not subject to editing. Stream 2 is subject to close scrutiny (full traditional edits), so are out of scope of the selective editing. However, in a private communication Farwell has stated that ABS has since shown, for this survey, that excluding the new units from selective editing was not necessary. Stream 4 is subject to (manual) editing. Thus only stream 5 is subject to selective editing, giving the final critical and non-critical sub-streams.

It is certainly sensible to remove response errors which may lead to significant biases or even prevent the selective editing procedure working effectively. Granquist (1997b) suggested that if fatal edits are included in the selective editing process, then they should always be automatically corrected, even if not prioritised to be in the critical stream. One of the key reasons for this suggestion is that these types of errors are usually easily noticed and would reduce user's confidence in the data and would often lead to many users running their own error corrections routines, which is wasteful of resources.

The decision about what form and extent of any (if any) initial editing depends on a number of factors, including:

1. survey design and type
2. data collection, input and storage
3. existing edit processes,

for the obvious reasons. If there are efficient and effective routines available to deal with some of these errors then it makes sense to use them. However, if the initial editing is resource intensive then it may be sensible to investigate whether they can be superseded by the selective editing process.

Although selective editing may supersede many traditional forms of editing, it is still best practice to use whatever form of editing makes best use of available resources (Sutcliffe and Farwell, 2005). Therefore, if an efficient and effective editing process is in place which can handle the important errors then it should be used. Further, selective editing is not effective in all situations, see discussion in Section 4, so some traditional editing routes may still be required, for example to handle categorical variables.

### **3.4. Focus on Influential Errors**

Latouche and Berthelot (1992) were amongst the first authors to propose the “use of a score function prioritise and limit recontacts in editing”. However, there are forms of selective editing which do not explicitly require scoring techniques, e.g. classification into streams using logistic or classification and regression trees (Breiman *et al.*, 1984) as discussed by de Waal (2002). However, by far the most commonly used techniques use scoring so they will be expanded upon here.

Latouche and Berthelot (1992) examined a periodic business survey and defined three score functions at the unit level to provide some form of metric for how far a particular unit’s responses are away from the edit model. The edit model essentially suggested the responses usually do not change substantially between surveys. These score functions did not explicitly measure the impact of the errors on the survey variables. However, the impact on the survey variables was used to set the streaming threshold.

The term significance editing has been coined by the ABS for a key variant of selective editing, see Lawrence and McKenzie (2000) and references therein. This term is generally used to indicate the use of score functions which directly quantify the impact of the errors on the survey variables. To be more precise the score is typically an estimate of the impact of resolving the edit queries on the survey variables. However, there is some confusion over the distinction between general selective editing and significance editing within the published literature.

Latouche and Berthelot (1992) did not provide a general framework for construction of score functions (which is still an open research question). However, they did suggest four features that the unit scores should take into account (slightly modified for consistency):

1. Importance (usually size) of the responding unit
2. Magnitude and influence (impact) of suspicious responses
3. Number of suspicious responses
4. Relative importance of items or variables.

Granquist (1997b) defined a similar set of criteria that the scores should quantify:

1. Weight of record
2. Potential impact of error
3. importance of flagged item/variable.

It is clear that most these sets of criterion are equivalent, the only exception being the number of suspicious variables which is explicitly defined by Latouche and Berthelot (1992). However, this exception can be implicitly encapsulated in the potential impact of error criterion from Granquist's definition.

Before continuing discussion of score functions, it is useful to consider the difference between magnitude, influence and impact of errors. The magnitude of an error is simply a metric of its size (usually estimated). The influence is a measure of importance of response/unit in determining, say, the survey variables. The impact of the error is then a measure of the combined magnitude and influence on the survey variable. It is often useful to think about these three concepts separately. Unfortunately, their distinction is somewhat blurred in the selective editing literature. A classic reference in the regression context is Cook and Weisberg (1982).

In many survey situations the size of the responding unit (1) will be a useful proxy for the influence of that unit on the survey variables, i.e. larger units tend to have larger impacts on variables. It is also clear that often the magnitude of the response error, or more precisely, the magnitude of the effect of the error on the survey variables may also be important. The units with a high number of edit failures may also warrant prioritisation.

The final feature for prioritising edits is the relative importance of the survey items or variables themselves. This would normally be defined by subject matter experts, or end-users, but would encapsulate issues like the variability of the survey variable estimates (survey variables which are more difficult to estimate should receive higher priority, e.g. cells small in size or with low response rate).

### **3.5. Score Functions**

The score assigned to each unit response is typically a function of some, if not all, of the criterion outlined by Latouche and Berthelot (1992), either explicitly or implicitly, henceforth denoted the score function.

Hedlin (2003) categorised the score functions into two broad classes, "estimate related" and "edit related", or a combination of both. The estimate related scores provide a measure of the "predicted impact" of the unit responses error on the survey variables, which do not necessarily need to rely on the current edit model but is somewhat dependent on the survey variables and their estimator. Essentially these (significance editing type) score functions estimate the change in the survey variable due to editing the response. Whereas, the edit related scores quantify how far the unit responses are away from the edit model (e.g. number of edit failures per unit and magnitude of failure), which rely solely on the edit model.

The definition of the score function is usually guided by knowledge from subject matter experts and exploratory studies comparing properties and the effectiveness of numerous different functions, as well user requirements. There is no consistent framework for constructing score functions in the literature; this seems to still be a topic of open

research. A number of procedures, usually graphical, have been developed to assess their empirical performance which will be discussed along with threshold selection below. This section provides some discussion of key developments of score functions in the published literature.

Latouche and Berthelot (1992) pointed out that the choice of score function should consider operational aspects. For example, if data is received at a steady rate during the survey process then it might be preferable to have score function which does not rely on the availability of the data from other units on the current cycle, as the data can then be processed as they arrive.

If the survey variables are provided at different levels of aggregation then it is usually desirable for the scores to have the same distribution (or least same location and scale) across domains for comparison.

The scores given to a particular unit response are often called the local scores, based on a local score function. As already stated, a local score is typically only calculated for responses which have failed an edit rule. Local scores are only assigned to items which have been identified as important in terms of quality of survey outcomes. For example, in significance editing a local score is provided for every item which contributes to the calculation of the key survey variables.

The local scores are then combined to give a global score for the particular unit, using a global score function. Note that in some survey designs the local scores may be combined over a subset of the items (groups of items), see Section 4 for further details. It is often required for the local scores to be standardised prior to be combined, to ensure comparability across items.

The unit responses can be ranked by their local score, for streaming, or more commonly the global scores are used to rank the units in terms of their relative importance for amendment/imputation, the largest scores being the most important. Typically, the units with a global score above a certain pre-chosen threshold (discussed further below) are subject to usual subject to manual editing, with those below the threshold automatically imputed. For example, Statistics Netherlands partition records into manual and automatic editing streams (Hoogland, 2002 and 2005). There are also many examples of surveys where the critical stream is automatically imputed, and the non-critical stream are left unedited.

The literature review has not revealed a general decision theoretic framework for streaming choices (manual/automatic/no edits). Some organisations, e.g. ABS, have defined a best practice model for editing, discussed further in Section 3.9, which can allow some uniformity of decisions across the organisation. In practice, decisions are often made on a survey by survey basis, based on available resources and user requirements. For example, Latouche and Berthelot (1992) ensure all out of scope units are recontacted (to ensure they are truly out of scope), total non-response is always followed up and all units which fell outside of the edit model but with less than 10 units within their respective cell were recontacted. However, for some surveys it may be appropriate to leave the errors below the threshold unedited.

### 3.6. Local Score Functions

It is beyond the scope of this report to recommend any particular form of (local or global) score function. General criteria for defining score functions were highlighted in the previous section. In the authors view, a few interesting developments in the literature are given by:

1. Latouche and Berthelot (1992) – specific edit related
2. Lawrence and McKenzie (2000) – general estimate related
3. Hedlin (2003) – edit and estimate related discussed
4. Jäder and Norberg (2005) – suspicion and potential impact combined
5. Hoogland (2005) – plausibility indicators
6. Farwell (2002) – impact and influence.

This section uses a common score function to highlight some of the general issues associated with their calculation. The (local and global) score functions play a pivotal role in the effectiveness of selective editing, therefore a wide ranging performance evaluation must be carried out prior to its implementation to ensure it is capturing all of the errors of importance.

It is quite common for estimators of survey variables to be weighted totals (or averages) of the response for a particular item over a number of units (e.g. total employment):

$$y_p = \sum_{i=1}^{n_p} w_i x_i$$

where  $n_p$  is the number of units within the domain of interest,  $y_p$  is the survey variable estimate,  $w_i$  is the survey weight and  $x_i$  is the responses for unit  $\{i: i=1, \dots, n_p\}$  in cell  $\{p: p=1, \dots, P\}$ . The local score assigned to unit  $i$  (in a significance editing approach) estimates the impact of amendment/imputation on the survey variable given by:

$$s_i = \hat{w}_i |x_i - \hat{x}_i|$$

where  $\hat{x}_i$  is an estimate of the expected amended response (after amendment or imputation). Clearly, it would be preferred to use the actual amended/imputed value (or even the correct response) but this is typically not known beforehand, so an estimate of the amended response is required. Also notice that often the survey weight is also replaced by an estimated value,  $\hat{w}_i$ , as it is often dependent on things like the response level and therefore may not be known before the editing and imputation process is finished.

There may be a need to scale the local score to allow comparison across units, domains or items. A common scaling for estimate related scores, like that defined above, is to divide them through by an estimate of the expected value of the variable. For the score function example defined above, this would give the relative contribution of the (estimated) error to the variable estimate. Thus an estimate of expected value of the variable may also be required, which is usually based on similar types of information as for the estimate of the expected item response.

The estimate of the score (and therefore the expected amended response/variable and weights where appropriate) must be consistent with the survey estimation procedure (e.g. edit model, form of amendment/imputation and variable estimation function). The estimate is usually based on past responses for the unit in periodic surveys, with possibly

some form of growth factor applied (similarly for the weight). Alternatively it may be possible to obtain estimates from current (or past) responses from similar units (e.g. within the same domain). Hedlin (2003), Lawrence and McKenzie (2000) and many others have suggested that the performance of selective editing is not unduly influenced by the accuracy of estimation of the expected amended response/variable, as long as the variance of the estimates is smaller than for the true error.

It is preferable for the scores to be independent of the response rate and the current responses, as this allows (amongst other things) unit responses to be scored as soon as they are received. As the score is used as a relative measure of performance, the accuracy of the estimate is not so important (Hedlin, 2003 and many others), as long it is consistent across units and consistent with the survey estimation methodology.

Almost all forms of score function require some type of expected item response (not necessarily estimate of amended/imputed value as is the case with significance editing), as they are usually based on a form of metric for the difference between the actual response and what is expected according to the edit model. For example, Latouche and Berthelot (1992) defined the RATIO global score function which was based on the ratio between the current response and final recorded response from the last survey (which is the expected response in this case), as the responses are not expected to change substantially between surveys. Hedlin (2003) defined edit-related scores by “measuring the magnitude of the failure” of the edit model, which essentially means he used an expected value of the response constrained by the edit rules. Common estimators include the mean or median of (scaled) responses from units within the same domain, or from previous surveys. Section 4 discusses the situation where the unit responses are difficult to predict, which can be the case for extremely volatile items.

### **3.7. Global Score Functions**

In some implementations of selective editing the local scores are used directly for prioritisation into streams. However, commonly the local scores are combined over all the responses to give a global score for the unit, which is subsequently used for prioritisation and streaming of the entire unit record. Section 4 also discusses scenarios where the global scores combine local score over different groups of items, (where responses are related within a group, but different groups are unrelated).

A general objective framework for choosing the global score function has not been considered in the literature, pragmatic choices seem to have been made. The choice seriously impacts the performance of the selective editing and the lack of an objective framework is lacking from the literature, see discussion in Sections 4 and 6. The choice of global score function should be guided by empirical evidence, subject matter expertise and preferably an exploratory multivariate analysis of the distributions of the local scores. Such an exploratory analysis of the local scores would reveal features of the scores which could be useful, e.g. dependence (correlation) between the local scores. For example, a principal component analysis (Johnson and Wichern, 2003), could reveal the dominant features (modes of variability) in the local scores. Examination of the joint distribution, combined with an error analysis and understanding of user quality requirements, will indicate which region(s) of the joint distribution of the local scores are important in terms of prioritising resources for amendment/imputation, and therefore how to define the global score function/thresholds to capture that region effectively.

Commonly used global score functions include (see diagrams in Figure 4.1):

1. Maximum (absolute) value of local scores – units enter the critical stream if at least one of their local scores is over the threshold
2. Sum of (absolute) value of local scores – units enter the critical stream when their total score is greater than threshold
3. Euclidean measure – square root of sum of square of local scores which is sensible if the local scores are independent of each other
4. Mahalanobis measure – takes account of covariance (correlation) between local scores.

In the first implementations of selective editing the maximum global score function was used for the obvious reasons, but as problems have been identified upon implementation in other surveys increasingly more complicated functions have been considered. This list roughly follows the timeline of these developments. The adaptations seem to have essentially been motivated by a need to mitigate the problems encountered (with some validation using empirical studies), rather than rigorously defining the regions of the local score space which contain the important errors to go into the critical stream.

The global score functions often contain some form of weights to account for issues like the relative importance of items/variables, which are normally decided upon by subject matter experts. The global scores may also need to be standardised, e.g. to make them comparable across domains which is discussed in the next section. Standardisation may also be required in implementation where different units may have a different numbers of local scores contributing to the global score (e.g. not all units will have failed same set of edit rules). These issues are of minor importance and so are not considered further.

### **3.8. Threshold Selection**

The initial threshold selection is preferably carried out using a simulation approach, which is discussed in this section. For a new or one-off surveys this may not be possible. Alternative threshold selection approaches are available. For example, a theoretic derivation by Lawrence and McKenzie (2000) under broad assumptions and various online procedures, e.g. cumulative cost-benefit curves of Farwell (2004) and progress graphs of Hedlin (2003) which are discussed further in Section 4 below. However, the threshold choice needs to be periodically maintained to ensure it is still being effective, for which Thompson and Hostetter (2001) provide a nice example, particularly in the light of changes to all aspects of the survey process, e.g. questionnaire, data collection, population demographic, other editing/imputation and user requirements. For threshold maintenance these online procedures are likely to be very useful.

Depending on the data flow and timescales involved in periodic surveys, it can be preferable for the threshold used in a periodic survey to be chosen in advance. This allows responses to be edited as they are received (as long as the score function does not necessarily rely on other unit responses), which can ensure selective editing is adaptable to the data flow of any survey.

The simulation approach requires two version of the same survey dataset:

- 1) Raw dataset – which has only been subject to initial editing
- 2) Extensively edited dataset – which has been subject to more stringent editing and amendment/imputation than selective editing is expected to provide.

Usually, the extensively edited dataset is the result of the more traditional editing approach outlined above, where all edit failures are manually amended or automatically imputed. Farwell (2004) has indicated that in the case of ABS, the availability of the extensively edited data is becoming less and less justifiable for resource considerations, which is also backed up by other agencies. Hence, some of the alternative approaches (in particular the interactive approach) described in Section 4 may come to the fore.

The simulation approach is far superior as it provides a much fuller picture of the error distribution and allows inexpensive evaluation of many alternative selecting editing parameters (e.g. consideration of numerous local and global score functions and thresholds). The simulation approach also allows a more complete performance evaluation of the whole selective editing procedure.

The simulation approach can consider the complete range of thresholds. For each threshold level chosen those units (if global score is combined over units) above the threshold are assigned to the critical stream, and the raw responses are replaced with the extensively edited values to give the simulated edit dataset. The non-critical stream data are usual not changed. The improvement in the survey variable estimates due to the simulated editing is then evaluated. As a reliable estimate of the final variable value is available from the extensively edited dataset, it is possible to quantify the bias in the estimate at each threshold level.

Various measures have been defined to quantify the improvement in the survey variable estimates. The most commonly used in practice are:

- i. Absolute pseudo-bias (Latouche and Berthelot, 1992):  $\left| \frac{\hat{y} - \hat{y}^*}{\hat{y}^*} \right|$
- ii. Relative pseudo-bias (Lawrence and McDavitt, 1994):  $\left| \frac{\hat{y} - \hat{y}^*}{SE(\hat{y}^*)} \right|$ ,

where  $\hat{y}^*$  is the final (best) variable estimate from the extensively edited dataset and  $\hat{y}$  is the variable estimate using the newly simulated dataset. The absolute pseudo-bias is logically defined but is found to be somewhat temperamental if the variable values are close to zero (Lawrence and McDavitt, 1994). If no other changes of the database are made the absolute pseudo bias will tend towards zero as the threshold decreases (though not necessarily strictly monotonically decreasing).

The relative pseudo-bias relates the bias to the standard error of the estimate, which encapsulates other error sources, e.g. sampling errors (Lawrence and McKenzie, 2000). Hence, the threshold can be chosen to satisfy conditions like that the selective editing should be reduce the bias to within 10% of the estimation error (as used by Hedlin, 2003 and Lawrence, 1999). Measures of this sort could be used towards defining standards for quality assurance.

Typically, the threshold choice can be guided by exploratory plots, for example:

- i. bias statistics against various threshold levels
- ii. bias statistics against number of (simulated) amended/imputed responses
- iii. total or percentage of cumulative change in variable estimate against thresholds or number of checks, e.g. progress graphs of Hedlin (2003).
- iv. total or cumulative (cost) benefit against unit/provider ranks (Farwell, 2004)

In order to be able to reduce resources allocated to editing, the graphs must indicate that a small number of errors are causing the largest bias reduction (i.e. large drop in bias at high thresholds/small number of edits). If each amended/imputed has the same cost associated with it, then the graphs which compare the number of amendments/imputations to the change in the estimate, are essentially cost-benefit curves. Some authors (e.g. Latouche and Berthelot, 1992) have suggested that the threshold could also be chosen to achieve a desired recontact/imputation rate.

An interesting approach was trialled by Hedlin (2003) for score function and threshold selection using the simulation approach. Hedlin compared the performance of:

- i. current editing method, with no prioritisation
- ii. an edit-related score function
- iii. an estimate-related score function
- iv. ideal method, where the final values are assumed known and used directly in the estimate related score function from (iii).

The ideal method provides a valuable gold standard by which to judge the performance of other editing processes.

Many authors advocate re-evaluation of the threshold periodically. For example, Hedlin (2003) of the ONS suggested the threshold should be initially chosen using the simulation approach but then dynamically reassessed, see Section 4 for more details. Best practice across the published literature also indicates the need to periodically re-evaluate the selective editing effectiveness by sub-sampling errors from non-critical stream and subject them to manual recontact/amendment. This will help identify any important errors which are being ignored by the score functions.

A important issue that has not been discussed so far is how to set thresholds when the global scores are calculated over different domains (or over groups of common variables in a survey). Firstly, the scores may need to be standardised to be comparable between domains/aggregation levels. Then there is a question as to whether there should be a common threshold across all domains at all levels of aggregation. For example should smaller domains, in which estimation is more variable, have a lower threshold. If different aggregation levels are considered, it seems sensible that the errors that are important at the highest aggregation level (i.e. HS1 in trade transactions) should be dealt with first, then errors which impact at lower levels. However, is this always the most sensible approach? There is also a need for management procedures to be in place to ensure that if a unit is to be manual recontacted then all possible edit failures are raised at the same time, rather than them being recontacted numerous times from each domain they are within. For surveys with a large number of domains, there may be too many graphs to examine for threshold choice. Farwell (2004) has suggested treating the curves as Lorenz curves and calculating GINI indices for comparison. There is no consistent pattern in the literature on the approach for dealing with these issues; a sensible pragmatic judgement seems to be made on a survey by survey basis.

### 3.9. Performance Assessment and Quality Improvement

One of the key features of the new view on editing is to collect data on the edit process (so called edit data) to monitor performance and to measure quality improvement (Granquist, 1995). The performance monitoring allows evaluation of the efficiency and effectiveness of the selective editing process (initial edits, edit model, scoring and streaming) and should provide information to highlight improvements to other parts of the survey process. The key principle is to move away from “error detection and correction” towards “error prevention”.

Data should be collated on the common error types, their sources and distributions which can be feedback to improve survey design, data collection, capture and processing. For example, if there are particular items where the responses are often outside of the edit model (i.e. incorrect or missing) this may indicate a problem with the question itself or the process by which the data is captured/processed. The edit data will also provide an audit trail to diagnose problems in the system and possibly to allow users to re-evaluate the editing process and the decisions made.

Key considerations in specifying the edit data collection requirements for quality purposes include (Lyberg *et al.*, 1998):

1. *Identification of product characteristics and user requirements* – what are the key outcomes, available resources and quality needs in terms of design and conformance, including possible future uses of data, and any guidelines for quality standards across collections
2. *Understand process* – process/data flow charts, so as to design efficiency into the process so errors are prevented in the first place
3. *Identify key process items* – what are the process components which impact on quality, provision of cause and effect diagrams (Aitken *et al.*, 2003)
4. *Evaluate measurement capability* – what can be measured? What level of accuracy? What are the costs of measurement and recording?
5. *Reduce variability* – predictable and stable processes are more likely to lead to reliable quality
6. *System for continuous monitoring* – how is performance going to be monitored, feedback provided and how progress is to be communicated to staff.

The UNECE (2000) defined three criteria by which quality of a statistical product should be assessed; relevance, timeliness and accuracy. Statistics Canada (2003) have further defined the additional criteria of accessibility, interpretability and coherence.

Continuous process monitoring and control can be a very valuable tool to provide early warning system for problems in the editing process. Statistical process control charts (Oakland, 1986) could be used to monitor process parameters like the change in variable estimates, hit rates and streaming over the editing cycle to ensure they are within an acceptable range or if they are out of control. For example, to detect problems when serious errors are slipping through the initial editing stage and reducing the effectiveness of the score functions.

Careful consideration has to be given as to what edit data is collected in order to satisfy user requirements (where users include personnel involved in future surveys), so as not to waste resources on collecting unimportant information. Rivière (2000) discusses the general issue of quality measurement and dissemination of quality indicators to end users. He suggests consideration of two key criteria when for dissemination: brevity and

readability. The user has to be provided with only the key information they need to evaluate the quality and its impact on their study. It should also be relatively easy for them read and comprehend the quality measures that are being provided to them.

Engström (1997) provides an example where collection of data on error causes was integrated into the survey process. Editors were required to identify the error cause, record an error code and indicate whether the unit was manually recontacted or not. The study concluded that provision of error codes was useful in terms of highlighting survey problem areas, but found the error allocations burdensome on the editors. Hence, an edit data system has to be designed to facilitate the analysis and reporting of performance statistics, but also make the edit data collection efficient and not resource intensive.

Common edit data to be collated include detection rates, correction hit rates, summary statistics on impact of amendments/imputations (possibly at different levels of aggregation). Rocca *et al.* (2005) and UNECE (2000) provide a detailed overview of useful data for quality indicators. For audit trails and transparency of editing for end users, where database systems allow, consideration should be given to storing the following data (in rough order of priority):

- i. Final response values (whether they be unedited/amended/imputed)
- ii. Original responses (including imaged questionnaires)
- iii. Error codes (and key), suggested reason why wrong, source of error, how amended/imputed, etc.
- iv. All amended/imputed values (despite condition that edit model, editing and amendment/imputation should be consistent with each other, multiple revisions often still end up being made)
- v. Edit flags passed or failed (or at least indication of more than one failure, or number of fatal versus query failures)
- vi. Record of edit rules and score functions in operation at time of survey
- vii. Local and global scores.

The issue of database systems and software is an important issue to be considered upon implementation for each survey, but is beyond the scope of this review. Examination of edit data lends itself to exploratory graphical techniques. There are many examples of useful graphs presented in the literature, some of which are discussed at relevant points throughout this report, but a full review is not included.

Best practice at some international official statistics agencies (although by no means the majority) includes provision of agency wide quality guidelines and editing strategy documents. A useful survey on the provision of guidelines is provided by Williams *et al.* (2000), from which the following details have been gathered. For example, Statistics Canada (2003) have produced quality guidelines which encompass editing and imputation methods. Statistics Canada also produce edit reviews for all their household surveys, which provide information which is feedback to future survey questionnaire designers and interviewers. The edit reviews are acknowledged as resource intensive and time consuming, but considered valuable for quality assurance. The ONS have a dedicated standards and quality assurance team, the standards of which are encapsulated in the BLAISE modules (developed by Statistics Netherlands). The ONS team also conduct internal edit process reviews towards the general quality assessment of a survey. The ABS has compiled a data editing manual, which acts as an invaluable central archive of editing techniques and guidelines for implementation. Although it is acknowledged by Williams *et al.* (2000) that the manual needs updating to include modern techniques, and is often ignored in practice. ABS also suggest using a quality assurance manager to lead

on developing and reviewing editing strategies (Sutcliffe and Farwell, 2005).

## 4. Some Implementation Issues

The general principles of selective editing are fairly straightforward and have been shown to work well in a variety of survey situations. Features that enable selective editing to be successful include availability of historical data (e.g. periodic surveys), high sample frame overlap between surveys, consistent survey design and predictable survey responses and variables. Selective editing has been shown to be effective for continuous items, particularly those with skewed error distributions where a small number of errors which are large in magnitude have the largest impact on the quality of survey outcomes. However, its implementation can be hampered by issues related to survey type (periodic, one-off, administrative, etc.), survey response type or structure, data flow and database requirements, etc. In this section some of the key issues raised in the literature and from the two parallel case studies are outlined and where available possible solutions (sometimes still an open topic for further research) are discussed.

A number of these issues were concisely dealt with by the excellent review carried out by Farwell (2004), from which some of the following discussions have been assumed.

### 4.1. Discrete Variables

The usual score functions are often inappropriate for discrete (e.g. categorical) variables, as there is usually no metric for quantifying the magnitude of possible errors and their impact. Clearly, depending on your definition, it may be possible to define some form of error metric for certain ordinal variables, but not in general. An error metric certainly cannot be defined for nominal variables.

There are of course many edit and imputation procedures available for discrete variables, Fellegi and Holt (1976) and its subsequent generalisations being one of the most rigorous. Another example is nearest neighbour imputation (Bankier 1999). However, a literature review in this area has not revealed an attempt to prioritise editing of errors in categorical variables, as these procedures are usually automated and so have relatively little resource cost.

If it is desired to include discrete variables within selective editing, then it may be possible to define the score function using the feature set outlined by Latouche and Berthelot (1990):

- i. Importance of the responding unit
- ii. Magnitude and influence of suspicious responses
- iii. Number of suspicious responses
- iv. Importance of items or variables.

It is beyond the scope of this report to outline any particular form of score function, however it is clear that the score function could encapsulate every suggested feature except for (ii) due to the lack of error metric. For example, it should be possible to quantify the relative importance of the responding units (i) and items/variables (iv) and therefore provide a measure of influence to enable prioritization. If the edit rules can check the categorical responses, then the feature (iii) can also be quantified.

## 4.2. Iteration of Selective Editing

Section 3 stated that some surveys undergo initial (traditional) editing before selective editing, usually to correct serious fatal edits, e.g. field and office editing (Pullum, *et al.*, 1986). In particular, all errors which may reduce the effectiveness of the selective editing should be dealt with (Farwell, 2004). However, some of these errors may not be detected prior to instigation of selective editing. Process monitoring and control procedures must be in place to ensure these are captured as soon as possible. The selective editing may need to be re-initiated (scores recalculated) once these errors have been adjusted.

This problem raises a number of questions:

- i. Can the selective editing (scoring) be made robust to these serious errors?
- ii. If not, then should selective editing routinely be re-iterated after an initial burn-in period?
- iii. If selective editing (and amendment/imputation) has been used at one level of aggregation (e.g. regions), should the scoring be recalculated at the next level of aggregation (e.g. whole population) of interest?

## 4.3. Streaming

The decision as to what form of amendment/imputation, if any, is to be carried out to each stream is usually decided on an ad-hoc basis for each survey, based on availability of resources, error types, quality standard requirements, etc. A topic of ongoing research is considering the definition of a formalised framework to enable such decisions to be made, using resource allocation criteria. Linacre and Trewin (1989) have made some headway on the allocation of resources across the whole survey process, and Rivière (2002) has also put forward a theoretical approach based on some standard assumptions. Best practice at statistical agencies seems to point towards a movement away from manual amendment to automatic amendment as far as possible for the critical stream, with some agencies (e.g. ABS and ONS) not editing the non-critical stream. A motivation for not editing the non-critical stream is to reduce concerns about overediting.

## 4.4. Local Score Functions

Most local score functions defined in the literature require some form of estimate of the expected response and some estimate-related score functions also require an estimate of the expected variable. In periodic surveys, these expected values (response or variable) are often based on historical data. This section discusses problems associated with estimation of these values when historical data is unavailable or they are hard to predict.

### 4.4.1. One-off surveys/no historical data

If no historical data is available then alternative sources of information are sought. Firstly, expected amended responses can be based on responses from units with similar characteristics (i.e. same cell). For example, the mean or median response from all units within cell, or regression estimates (Farwell, 2004). Secondly, expected amended variables can be based on expected responses. However, Farwell (2004) has found that these can be wildly inaccurate. Although not feasible for large surveys, Farwell went on to take a pragmatic pseudo-Bayesian approach in one case study, basing “guesstimates”

of variables on past estimates from very old surveys combined with those based on the expected responses. It is certainly sensible tap all available sources for information to improve response/variable estimates. Farwell (2004) also gave an example of a score function which, when standardized, removed the need to have expected variable estimates. Other alternatives are discussed in the next sub-section.

#### 4.4.2. Difficult to predict responses/variables

If the expected responses/variables are not able to be reliably estimated (e.g. response too volatile, or not historical data), then it may be possible to try a different form of score function which is correlated with usual score. See general discussion of score functions above. Farwell (2004) gives an example of using a combination of three score functions which measure the (standardized) approximate unit's contributions to the level, movement (requires past response data) and standard error of the variable of interest. The three local scores were combined using a weighted Euclidean distance to give a global score, which is correlated with the usual measure of impact of errors, as the largest contributors often have largest impact on estimates of survey variables.

### 4.5. Global Score Functions

In general, the global score functions (if required) combine the local response errors to give a score to each unit which is used to prioritise units into streams. The functions suggested in the literature (commonly one of the maximum, sum, Euclidean, Mahalanobis of local scores) have been pragmatically adopted. This area is ripe for extensive research, initially based on an exploratory analysis of the local scores, comprehension the aim of prioritisation and any theoretical justifications.

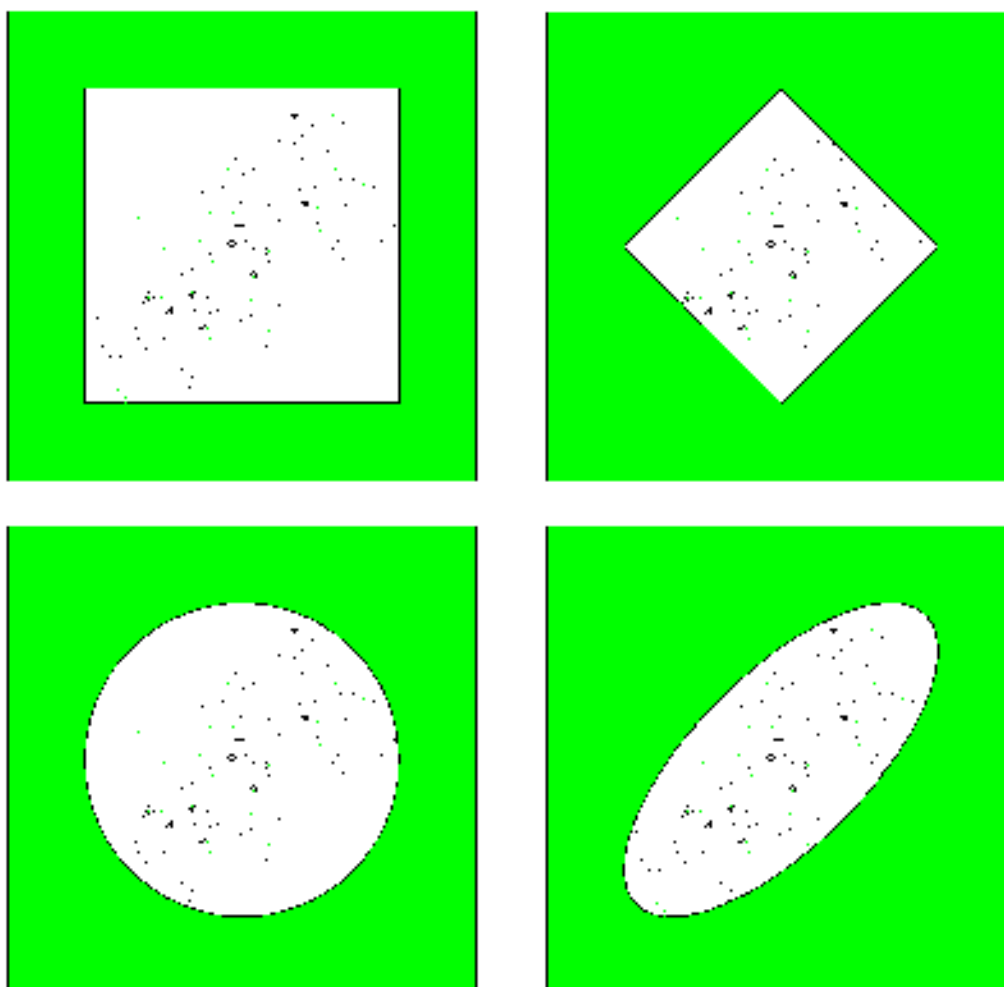
#### 4.5.1. Large number of local scores (key variables)

There has been much concern raised about this issue and ad-hoc choices of global score functions have been made to overcome it based on empirical evidence. However, a detailed multivariate analysis of the local scores will provide a better justification for the global score function choice. Firstly, it is clear that as the number of local scores increases the probability that the maximum of them is above the threshold increases, i.e. the more local scores leads the higher chance of the unit needing editing/imputation.

To motivate this result we will consider a crude example. Assume that the local scores  $X_1, X_2, \dots, X_n$  are independent identically distributed random variables then the survivor function of the maximum:

$$\begin{aligned} P\{\max(X_1, X_2, \dots, X_n) > x\} &= 1 - P\{\max(X_1, X_2, \dots, X_n) < x\} \\ &= 1 - P\{(X_1 < x) \cap (X_2 < x) \cap \dots \cap (X_n < x)\} \\ &= 1 - P\{X_1 < x\}P\{X_2 < x\} \dots P\{X_n < x\} \quad \text{as independent} \\ &= 1 - P\{X < x\}P\{X < x\} \dots P\{X < x\} \quad \text{as identical} \\ &= 1 - P\{X < x\}^n \end{aligned}$$

Therefore, as the number of local scores  $n$  increases the probability a unit's maximum local score is above the critical threshold  $x$  gets closer to one. This explains why authors have found too many units were put in the critical stream when using the maximum global score function with a large number of local scores.



**Figure 4.1.** Examples of critical threshold regions (shaded) defined by 4 global score functions combining two local scores. Clockwise from top left:  $\max(|X_1|, |X_2|) > 2$ ,  $|X_1| + |X_2| > 2$ ,  $\sqrt{X_1^2 + X_2^2} > 2$  and  $\sqrt{(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})} > 2$ . An arbitrary threshold of 2 has been used. The simulated values (representing local scores) are from a bivariate normal with common mean 0 and variance 1 and correlation 0.7.

The global score function has to ensure that the region of the space filled by the important errors that need prioritising in the critical stream can be identified. Figure 4.1 shows scatter diagrams of a set of simulated of bivariate normally distributed observations with some dependence (correlation) between them, with examples of the four commonly used global score functions. The observations are intended to crudely represent the behaviour of 2 local scores, which are related to each other. Dependence between local scores is to be expected if the items they are based on are related to each other, as if a unit answers one of the questions incorrectly they are more likely to answer other related questions incorrectly as well.

The four graphs show the critical regions defined by the four global score functions: maximum, sum, Euclidean and Mahalanobis measures. Note that the weighted Euclidean distance measure used in the literature is equivalent to turning the circular region (of non-critical errors) into an ellipse. Although the non-critical regions are not of the same size, the difference in the shape of the regions is evident. The critical (shaded) regions clearly

prioritise very different groups of errors. This shows that it is important to consider what regions contain the important errors that need to be assigned to the critical stream. Thus definition of the global score function requires a detailed multivariate analysis to understand of the distributions of the local scores.

#### **4.5.2. Provider or Group Scores**

The majority of implementations of selective editing in the literature either use the local response scores to prioritise for streaming, or they combine the all the local scores within a unit to give a global score, the global score is then used for prioritization (possibly within different domains separately). In other words either the individual unit responses or the units themselves are prioritised. However, it is not uncommon for a single provider to supply response for a group of subsidiary units (e.g. the head office of a conglomerate organisation may provide survey responses on behalf of its subsidiary businesses) or surveys where the survey can be broken down into independent components (so called “item groups” from Farwell, 2004) which are not related to each other. It is not appropriate to review the entire unit record as this is likely to be a waste of resources, only the response given in the item group which contains the important error should be reviewed. In this type of scenario, it is sensible for the global scores to combine responses within each item group. Farwell (2004) provide an example of this working in practice for the Australian Agriculture Survey. Although Farwell (2004) points out that the provider score are still a useful tool to manage manual recontacts, i.e. to make sure that if a recontact is required that all queries are dealt with in one go.

#### **4.6. Threshold Selection**

The ideal world approach to defining thresholds for selection of the critical stream is to use a simulation approach which requires original (as far as possible) unedited data and the set of intensively edited final data, which will provide a good picture of the full error distribution. In general, the simulation approach allows a full cost-benefit analysis of the edit and imputation process and the threshold chosen can be used for future surveys. The benefit of having a pre-specified threshold is that it may be possible to commence editing as soon as the data is available. Farwell (2004) also pointed out that the main drawback to a pre-specified threshold is that the number of providers selected can vary, as can the total benefit attained.

Online (graphical) threshold selection procedures have been suggested by Farwell (2004) and Hedlin (2003), which do not require assumptions about the errors and may not require the full error distribution to be known. Hedlin (2003) suggest using progress graphs of the cumulative change in a survey variable estimate against the rank of the unit scores. It is expected the largest changes will be for highest ranked units, with the variance decreasing with rank. The threshold score is chosen when the cumulative change has stayed on zero for a number of unit edits. There is no guarantee that units with scores below the threshold will have no impact on the survey outcomes, hence Hedlin (2003) suggest analysing a sample of units from the non-critical stream and subject them to amendment/imputation. Farwell (2004) suggested plotting the cumulative expected benefit against the cost of editing (same as rank if equal editing cost per unit). These graphs are easily implemented if the data comes in batch mode or the threshold is to be chosen in advance on past data, but may need to be carefully managed if data is received/edited on a more continuous basis.

### 4.6.1. No before-and-after editing data

If no before and after data are available then the thresholds can be set online (interactively), using the graphical tools mentioned in Section 3 above. The main drawback is that editor does not get to examine the full error distribution. Farwell (2004) pointed out that it is becoming increasing harder to justify resources to provide extensively edited data, so online threshold selection is becoming more common. Lawrence and McKenzie (2000) put forward a model based procedure for threshold selection under some standard assumptions. A concern with the interactive threshold selection approach is how it is managed relative to the data input flow. For surveys where the whole data base is bulk-loaded this is easier to manage, but not if data trickles in. This type of issue needs to be investigated as part a scoping study prior to implementation of selective editing for a survey.

### 4.6.2. Conservation of joint distributions

The editing and imputation procedure of Fellegi and Holt (1976) defined three guiding principles that their procedure adhere to:

1. Minimal changes to original data (preserve as many of original data as possible, the error-localisation problem)
2. Consistency between edit model, editing and imputation
3. Preservation of marginal and preferably joint distributions of responses.

Although selective editing is not an imputation procedure *per se*, it does dictate which errors are prioritised for later imputation. So there is a need to consider whether the streaming is going to impact on these sorts of guiding principles for imputation. Although imputation is outside of the remit of this report, this foundational concern warrants a short discussion.

Selective editing does explicitly uphold the second principle. But the other two are not explicitly upheld, the most serious of which is considered the preservation of marginal and joint distributions. The third principle implies that the imputation should not bias the responses towards the edit model. The selective does not implicitly ensure that the editing is uniformly spread over the population; therefore particular groups within the populations may be over-represented in the critical stream and others under-represented. If the selective editing/imputation route is not well designed, there is concern that this could cause a bias in the joint distributions of the responses.

Related concerns have been raised by Tate (2000) and the masking type effects of Hidioglou and Berthelot (1986). Tate (2002) raises a concern over units which provide data which is increasingly in error, so if score measure change between periods, then it may not pick it up. In some sense this is a systematic type error, which selective editing is not designed to detect. Macro editing would often target this type of error.

Similarly, in selective editing attention is typically focussed on the importance of errors within particular units, however, it will not necessarily detect small (more systematic) errors in individual units which can combine to impact on the survey variables. An open question seems to be whether score functions can be designed to capture these more systematic errors, or whether they have to be dealt with by macro-editing.

## 5. Key Principles and General Framework for Implementation

Previous sections have outlined the key features of the selective editing methodology and some of the issues to be considered upon implementation. Towards considering the implementation strategy across Statistics NZ surveys it is useful to outline the guiding principles of selective editing and a general framework.

### 5.1. Key Principles of Selective Editing

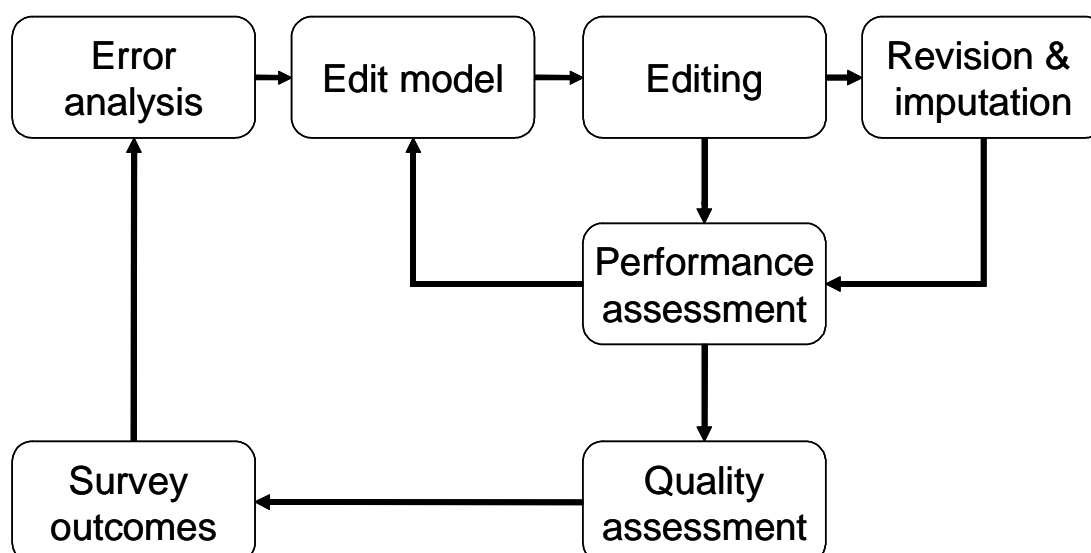
The foremost principle of selective editing is that of prioritisation. However, a range of guiding principles are outlined below, some of which are common to all edit processes.

- 1. Prioritisation.** Editing resources should be focused on errors that have the largest impact on quality of survey variables. Need to expend sufficient resources to ensure variables are at the quality standard required to be “fit for purpose”.
- 2. Efficient and Effective.** Editing needs to identify all important errors, thus requiring a well designed edit model and score functions. Further, editing needs to limit identification of false and unimportant errors, and needs to provide information for continuous performance assessment and process control.
- 3. Dynamic.** Process should not be set in stone. All components (e.g. edit rules, score functions, thresholds, record keeping) need to be periodically re-evaluated to ensure they are still efficient and effective, in light of changes to; survey design, other survey processes (pre and post editing), population and user needs.
- 4. Balanced and practical.** Balance needs of users and available resources. In order to not be too onerous, it is important the methodology can integrate with existing systems and processes and adaptable to solve real problems across surveys.
- 5. Consistent.** Editing should be consistently applied, reproducible and transparent. Move away from subjective judgment based editing, to formalized procedures which allow audit trails. Automation is likely to help achieve this goal.
- 6. Learning and feedback.** Performance systems to evaluate process and identify improvements, to provide evidence that editing is meeting needs of stakeholders and making best use of resources. Communicating feedback to whole survey process (including future surveys), e.g. common errors, sources and distributions. Performance measures can also be a useful tool towards staff motivation.
- 7. Prevention.** Paradigm shift from “error detection and correction” to “error prevention” (Granquist, 1997b). Edit data need to be recorded, analysed and reported for future error prevention, e.g. changes to survey design or data input.
- 8. Quality assurance.** To ensure quality conformance (“fitness for use”) the users requirements (past, present and future) need to be identified and provide tools to monitor edit performance. Further, users should be made aware of limitations of edit process (e.g. likely remaining errors). To ensure quality in design editing must be working efficiently and effectively, mitigating all errors which will impact later stages in process. Editing procedure must also be statistically justifiable, to ensure no systematic biases are introduced.

## 5.2. General Framework

In order to define a general framework for the implementation of selective editing it is first of all useful to understand how it could integrate into a general editing and imputation process and how it impacts on the components of that process, including feedback mechanisms. Following this an outline of a general selective editing framework will be detailed.

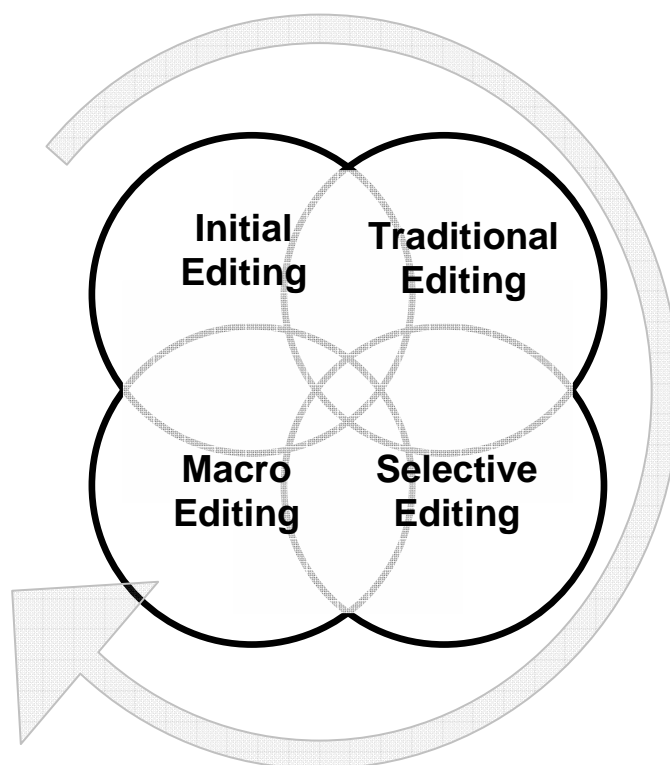
In some ways, the general framework implicitly assumes that selective editing will be combined into a pre-existing editing and imputation framework, as this is likely to be the most cost-effective implementation route in practice. However, in an ideal world the longer term approach may be to develop selective editing from the ground-up, fully integrating with other editing and imputation processes. For example, there is no need for an edit model, as the score function itself can be used to define the responses which are in-scope of selective editing. There is also potential for other efficiencies to be gained. For example, the error localisation problem may be aided with information from all stages of the editing and imputation process.



**Figure 5.1.** Edit and imputation process flow chart.

- 1. Survey outcomes and uses.** It is important to have a clear definition of the aims and objectives of the survey, including user requirements, so that procedures and processes can be put in place to ensure they are fulfilled in an efficient and effective manner. For example, understanding of data context, timescales, quality requirements and available resources. For selective editing to be effective it needs to make sure edits are prioritised to minimize the remaining bias in the survey variables and ensure effective use of resources. In order to do this effectively the key survey outputs (and therefore key variables/items) and domains of interest need to be identified. All components of editing (edit rules, score functions and imputation) must be in accordance with the primary outcomes of the survey.

2. **Error Analysis.** In order to improve quality of outcomes, it is important to know the error types, their source and distribution. This will often be based on exploratory (often graphical) analysis of the survey data, at various levels of aggregation. Comprehension of the expected errors should lead to effective definition of edit rules and a way to measure the edit performance and quality assessment. In particular, how to define edit model and score functions to capture all errors which seriously impact survey variables. Depending on the data collection flow it may be possible to carry out an error analysis online, which could provide early warning of problems in survey design or data collection/input, thus leading to early quality improvements.
  
3. **Edit Model.** Selective editing does not require an edit model, as such. However, most implementations are based on a previous micro-editing system, so often only those items which have failed edit rules will be in-scope of selective editing. Williams *et al.* (2000) suggest that edit rules should be defined considering four available sources of information: knowledge of key items/variables, understanding of conditions likely to influence responses, knowledge of type and impact of errors on quality and performance of the edits. Effective and efficient editing and imputation requires well thought out and meticulously defined edit rules. Firstly, errors which are outside of the edit model are likely to slip through early edit stages and are, at best, likely to be detected at the macro-editing stage. Efficient editing requires sensibly defined edit rules so as not to make editing too onerous, e.g. loose range checks lead to many possible errors to be considered whereas tight range checks may lead important errors being missed. The edit model needs to be transparent (all explicit and implicit conditions identified) to users and editors, to ensure it is known what errors are identified (or not).



**Figure 5.2.** A common sequence of different forms of editing.

- 4. Editing.** The editing stage should ensure the bias on survey variable estimates is minimized, without overediting and making efficient use of resources. There are four commonly defined stages to editing which have a natural ordering, see Figure 5.2. The reason for the overlap between editing stages, is to represent the possibility that within a particular survey the different forms of error can often be captured during different edit processes. Further, the optimal balance of the routes depends on the survey and the editor's philosophical viewpoint. Key criteria for editing defined by Williams *et al.* (2000) is that it must be appropriate, effective, efficient and defensible.
- a. **Initial edits** – in general these cover serious errors, e.g. incorrect units, validity checks, non-response, imaging errors. These are usually obvious, mainly fatal errors which are commonly easy to identify. Keying staff often used to correct these types of error on data entry, but with survey imaging becoming more prevalent these errors often require intervention. Although often a traditional edit process, it is preferable to automatically amend these errors before they impact on later stages of the process.
  - b. **Traditional edits** – incorporates a small number of important errors, detected by edit rules, which cannot be incorporated in selective editing, e.g. certain categorical variables, or may prevent selective editing working effectively, e.g. extremely large errors which impact estimation of scores. Sutcliffe and Farwell (2005) also point out, although the long term goal is to replace traditional edits with well thought out selective editing, in the short term some traditional edits may be needed in the transition period.
  - c. **Selective edits** – prioritising errors which have impact on survey outcomes. Indication of editing priority, to allow streaming of edits, e.g. manual recontact/intervention, automatic imputation, no editing required. Extensive testing of score functions is required to ensure they can be calculated and are effective in the range of likely error scenarios. Significance editing can also provide measure of expected impact of error amendment on survey variables.
  - d. **Macro edits** – provides final check that data to be published are as expected, usually given subject matter knowledge. Taking a look at responses as a whole and the information they provide and check they make physical sense, and that there are no remaining errors which have a substantial impact on the survey outcomes. For example, detection of systematic errors (like under-reporting of profits) which is difficult to identify at micro level. Should check multivariate relationships between items/variables. Information on consistent patterns of errors from this phase should also be passed back to earlier stages, to improve in particular the edit model and score functions in selective editing.

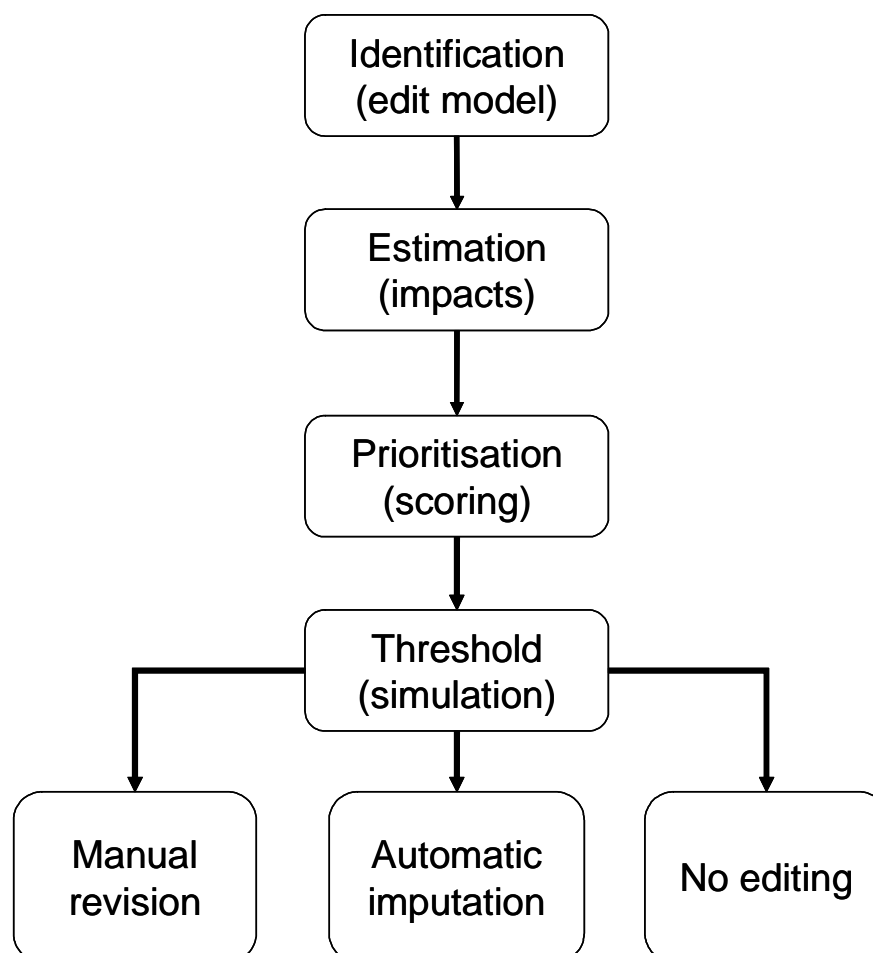
- 5. Revision and Imputation** – The editing stage should indicate what form of revision or imputation is required. A key principle of imputation is that it should be consistent with the edit model. Fellegi-Holt (1976) showed it is desirable to make minimal changes to the responses to satisfy the edit model, and that any imputation should conserve the marginal and joint distributions of the responses. In general, it is considered preferable to move towards automatic imputation rather than manual intervention, for resource efficiency and consistency reasons. However, it is to be acknowledged that an automatic imputation process needs to be well designed and regularly reviewed to ensure good performance.
- e. **Manual** – Traditional approach to data revision. Usually relies on recontact or decisions made by editors using available information without recontact, e.g. past responses, responses from similar units within cell. It is generally the most costly form of response revision, however, it may be the only option if the responses are difficult to predict. In the case of recontact or follow-up, which is often considered the best way to lead to improvements in data quality, it can increase respondent burden which can impact on quality of response in future surveys. Manual revision without recontact also has many potential drawbacks, in terms of lack of consistency (often based on subjective decisions, which differ between editors), transparency, reliability and reproducibility, shown to lead to creative editing, revision to satisfy edit model. It is also difficult to quantify the benefit of editor expertise in assessing quality.
  - f. **Automatic** – Often the preferred route. It will generally provide results which are consistent and reproducible (in a statistical sense). Further, the quantification of quality is somewhat easier. There are many off-the-shelf imputation routines (and software packages) which can handle most survey designs, response data types, etc.
  - g. **Not Edited** – Depending on the survey outcomes and the user requirements, it may be possible to ignore non-serious (non-fatal) errors which have little or no impact on survey outcomes. It is often that case that non-systematic random errors will cancel each other out in the estimation of survey variables. Reduces possibility of bias introduced due to overediting. However, the decision for this must be justified with considerations of available resources and quality assurance requirements.
- 6. Performance Assessment** – A key component of best practice for editing and imputation is the need for regular editing and imputation reviews. To diagnose problems and to assess performance, in particular in light of changes to survey design, other survey processes (pre and post editing), population and user needs. In order to measure performance there is a need to record meta-data at all stages of the editing and imputation process. The sort of data that could be useful for performance assessment are; records of what and how many edits were undertaken, any changes made to responses, information (coding) on possible sources of errors. In particular, for selective editing records of changes to the estimates of the survey variables should be recorded to ensure the thresholds are set at an appropriate level. Statistical process control ideas could be incorporated to continuously ensure the whole process is within control, i.e. meeting

performance targets, thus providing an early warning system for problems. In particular, the selective editing needs to have periodic reviews to ensure there is no long term biases due to its continued use. For example, many authors have suggested periodically sub-sampling errors below the threshold and place them in the critical stream to evaluate performance. The ability to assess performance can also provide useful information for motivating the editing staff to achieve performance targets.

7. **Quality assurance** – The measures and information provided to judge performance of editing can also be used for quality assurance purposes. To quantify the impact of editing and imputation on the survey variables. Further, feedback of common errors will lead to better understanding of survey outcomes, and improvement of error analysis, edit model, etc., thus improving quality of whole survey process. The move towards more formalized automated (less subjective) approach to editing and imputation could enable more transparency in the edit process. Editing and imputation reviews could also allow users to re-evaluate the editing that was carried out and outline limitations of the process.

### 5.3. Selective Editing Framework

The rest of this section defines a general framework for the implementation of the selective editing process. For a detailed discussion of the issues involved in implementation the reader is referred to Sections 3 and 4.



**Figure 5.3.** General framework and flow for selective editing.

1. **Identification** – The first thing to decide is which items and errors are in scope of selective editing. In general, most implementations of selective editing define the responses which are outside of the edit model as in-scope. If this is the case, then the edit rules need to be meticulously designed to account for all possible errors of potential interest. Consideration also has to be given as to which items and errors may need to be dealt with using initial, traditional and macro-editing processes given available resources.
2. **Estimation** – In order to prioritise units for editing, each unit requires some form of score allocated to it. The score will generally (implicitly or explicitly) encapsulate (due to Latouche and Berthelot, 1992) the importance (size) of the responding unit, the magnitude and influence of the error, number of suspicious responses per unit and relative importance of the items or variables. Generally,

the score will measure distance from the edit model (edit-related score – due to Hedlin, 2003) or the impact on survey variables (estimate-related). In fact the significance editing variant outlined by ABS, scores based on an estimate of the expected impact on the survey variables of correcting the error. Consideration has to be given to how to estimate expected response/variables, e.g. using historical data or from similar units within cell.

- 3. Prioritisation** – In most selective editing implementations it is necessary to combine unit response score to give a global unit score, in order to prioritise units for streaming in review/imputation stage. Although, it is possible that in certain situations it is desirable to prioritise responses based on local scores, or groups of responses. A desirable feature of some score functions is that they have the same distribution (or at least same location and scale) across different cells. For surveys with different levels of aggregation a decision needs to be made as to how to prioritise units across different aggregations.
- 4. Threshold selection** – The final phase of selective editing is deciding how to stream priority units for review/imputation, see item 5 in section 5.2 for details. The decision for whether the units are subject to manual (recontact or review), automatic imputation or no selective edits is made on a survey by survey basis, and should consider issues like; available resources, performance of automatic imputation techniques, end user tolerance of remaining errors. The selection of threshold is usually carried out using a simulation approach (where raw data before editing and after some form of extensive editing is available), which enables analysis of full error distribution, investigation of different score functions and threshold levels. The impact of different threshold levels for the streams on the survey outcomes is usually measured by some form of (standardized) bias statistic. The threshold is chosen so as the bias is within acceptable bounds, considering the cost of review/imputation in each of the streams. However, alternative strategies for threshold selection are available using either theoretical justifications (under varying assumptions) or online threshold selection.

## 6. Recommendations for Future Consideration

Based on the evidence gathered as part of this review, a number of recommendations have been devised which the author believes should be considered in developing the strategy for implementation of selective editing within Statistics NZ. The recommendations are categorised into those related to practical implementation issues, foundational concerns and further research questions.

### 6.1. Practical Implementation

*The implementation and integration of selective editing is recommended.* Selective editing is a useful implement in the toolbox of an editor, as it allows prioritisation of editing resources based on the impact of errors on the survey outcomes. The resource savings have the potential to be quite large, depending on the survey characteristics. Another benefit of selective editing is that it is very adaptable and has shown to be easy to integrate with existing systems. However, it should be noted that selective editing is not suitable for all surveys and is not always suitable as a standalone tool. Best practice is a combination of input, traditional, selective and macro editing processes, the balance of which depends on characteristics of the survey.

*A strategic approach to implementation is advised.* This document provides a general framework which should be supplemented with the operational knowledge and experience within Statistics NZ. Statistics NZ should consider the balance it wishes to achieve between the goals of improved data quality, resource saving and timeliness. The strategy should provide for flexibility across surveys, while also enabling shared approaches, functionality and systems wherever possible.

*Selective editing should be integrated within a broader approach to quality improvement.* This will enable learning and feedback to be integrated into other parts of the survey, as this will better enable error prevention and satisfaction of user requirements. Statistics NZ should also consider using resource savings from selective editing to improve other parts of the survey process and analysis of survey data.

*The development of quality standards and editing guidelines should be considered.* This should enable staff at all levels and end users to better understand the overall objectives of editing and how they are achieved. This could also support a cultural change from the traditional “error detection and correction” mindset to modern quality focused “error prevention”. A number of national statistics agencies have adopted this approach; indeed some have quality managers/teams/networks. High level leadership, training and support will be required.

*The principle of regular review should be adopted.* This means that both the over-arching strategy and the survey specific implementations should be subject to continual (periodic) performance assessment. Of particular concern, the selective editing parameters need to adapt to changing user requirements, survey procedures and systems, technological advances and changing populations. Mechanisms for performance assessment and process monitoring and control are also desirable.

*Prior to survey specific implementation an extensive scoping study should be undertaken.* Selective editing is a powerful tool. However, ill-considered implementations risk seriously undermining data quality. The study should consider the nature of the survey,

quality requirements, training needs, available resources and thorough evaluation of components of the editing strategy (e.g. edit model, score functions, edit data).

## 6.2. Foundational Issues

The following foundational concerns warrant additional investigation and consideration by Statistics NZ.

Fellegi and Holt (1976) define three key principles of their imputation approach (slightly adapted for presentation):

1. Minimal changes to original data (error localisation)
2. Consistency between edit model, editing and imputation
3. Preservation of marginal and preferably joint distributions of responses.

Although selective editing is not an imputation procedure as such, the streaming it provides impacts upon any amendment/imputation required. In particular, there is concern that the final principle of preservation of the marginal/joint distributions is not explicitly ingrained in the selective editing framework. See Section 4 for further discussion. *Further investigation of the implication of these concerns in practice is recommended.*

Selective editing focuses attention on the important of errors within particular units. However, it is difficult for scoring functions to detect small errors in individual units which have large combined impact on the survey variables. Selective editing is certainly not explicitly designed to detect systematic errors, e.g. underreporting across a number of units. Therefore in order to capture errors which would usually have been detected in the traditional editing route, it is recommended that *the macro editing process be re-evaluated upon consideration of selective editing implementation.*

## 6.3. Research Questions

*Framework for score function derivation and threshold selection* – general positive features of score functions are well known. However, little work has gone into defining a general objective framework for their construction, based on the literature reviewed. In particular, the rather pragmatic adaptation of global score functions seems to have been somewhat ad-hoc, based on purely empirical performance evidence. Since the original submission of this report the author has been made aware of work by Farwell *et al.* (2002) who consider a technical framework which jointly considers the balance between reporting bias and cost of editing. The framework provides analytic solutions in the case of equal costs per edit, in terms of minimising reporting bias for fixed cost or minimising cost for fixed bias. However, time constraints restricted fuller evaluation of this work.

*Streaming is essentially a classification problem* – is it possible to use some modern classification statistical tools for this problem (classification and regression trees, Breiman *et al.*, 1984).

*Alternative scoring techniques* – scores often involves measurement of magnitude of error and influence on variables. Can alternative statistical tools be used effectively, for example cross-validation/case deletion diagnostics (Cook and Weisberg, 1982) or non-parametric methods? What can be learnt from the outlier testing literature (Barnett and Lewis, 1984)?

## References

A useful resource on statistical data editing is K-Base (Knowledge Base for Statistical Data Editing) at <http://amrads.jrc.cec.eu.int/k-base/new.htm>, particularly papers presented at UNECE events.

Aitken, A., Hörngren, J., Jones, N., Lewis, D. and Zilhão, M.J. (2003). Handbook on improving quality by analysis of process variables. Eurostat, European Commission.

Bankier, M. (1999). Experience with the new imputation methodology used in the 1996 Canadian census with extensions for future censuses. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing*, Rome, Italy. Working paper 6.

Berthelot, J.-M. and Latouche, M. (1993). Improving the efficiency of data collection: A generic respondent follow-up strategy for economic surveys. *J. of Bus. and Econ. Stat.* **11**(4), 417-424.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Wadsworth, Pacific Grove, CA.

Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. Chapman and Hall, London.

Cox, B.G. Binder, D.A. Chinnappa, B.N. Christianson, A. Colledge M.J. and Kott, P.S. (eds). (1995) *Business Survey Methods*. John Wiley and Sons: New York.

de Waal, T. (2002). Development of modern edit and imputation methods at Statistics Netherlands. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing*, Helsinki, Finland. Working paper 32.

Engström, P. (1997). A small study on using editing process data for evaluation of the European Structure of Earnings Survey. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing*, Prague, Czech Republic. Working paper 19.

Engström, P. and Granquist, L. (1999). Improving quality by modern editing. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing*, Rome, Italy. Working paper 23.

Farwell, K. (2004). *The general application of significance editing to economic collection*. Technical report for Australian Bureau of Statistics (available from <http://www.abs.gov.au>).

Farwell, K. and Raine, M. (2000). Some current approaches to editing in the ABS. *Int. Conf. on Establishment Surveys II*, Buffalo, USA.

Farwell, K., Poole, R. and Carlton, S. (2002). A technical framework for input significance editing. *Proc. Of Data Clean 2002*, Finland.

Fellegi, I.P. and Holt, D. (1976). A systematic approach to automatic edit and imputation. *J. Amer. Stat. Ass.* **71**, 17-35.

Granquist, L. (1995). Improving the traditional editing process. In Cox, Binder, Chinnappa, Christianson, Colledge and Kott (eds), *Business Survey Methods*, John Wiley and Sons, 385-401.

Granquist, L. (1997a). On the current best methods document: Edit efficiently. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing*. Working paper 30.

Granquist, L. (1997b). The new view on editing. *Int. Stat. Rev.* **65**(3), 381-387.

Granquist, L. and Kovar, J.G. (1997). Editing of survey data: How much is enough? In Lyberg, Biemer, Collins, de Leeuw, Dippo, Schwarz, and Trewin (eds), *Survey Measurement and Process Quality*, John Wiley and Sons, 415-435.

- Hedlin, D. (2002). Score functions to reduce business survey editing at the ONS. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing, Helsinki Finland*. Working paper 14.
- Hedlin, D. (2003). Score functions to reduce business survey editing at the U.K. Office for National Statistics. *J. of Off. Stat.* **19**(2), 177-199.
- Hidiroglou, M.A. and Berthelot, J.-M. (1986). Statistical editing and imputation for periodic business surveys. *Surv. Meth.* **12**(1), 73-83.
- Hoogland, J. (2002). Selective editing by means of plausibility indicators. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing, Helsinki, Finland*. Working paper 33.
- Hoogland, J. (2005). Selective editing using plausibility indicators and SLICE. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing, Ottawa, Canada*.
- Jäder, A. and Norberg, A. (2005). A selective editing method considering both suspicion and potential impact, developed and applied to the Swedish foreign national statistics. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing, Ottawa, Canada*. Working paper 12.
- Johnson, R.A. and Wichern, D.W. (2003) *Applied Multivariate Statistical Analysis*. Prentice Hall.
- Latouche, M. and Berthelot, J.-M. (1990). Use of a score function for error correction in business surveys at Statistics Canada. *Proc. of the International Conference on Measurement Errors in Surveys*.
- Latouche, M. and Berthelot, J.-M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *J. of Off. Stat.* **8**(3), 389-400.
- Lawrence, D. and McDavitt, C. (1994). Significance editing in the Australian Survey of Average Weekly Earnings. *J. of Off. Stat.* **10**(4), 437-447.
- Lawrence, D. and McKenzie, R. (2000). The general application of significance editing. *J. of Off. Stat.* **16**(3), 243-253.
- Lawrence, G. (1999). Significance editing in the survey of employment and earnings. (available from <http://www.oecd.org>).
- Linacre, S.J. and Trewin, D.J. (1989). Evaluation of Errors and Appropriate Resource Allocation in Economic Collections. *Proc. of the 1989 Annual Research Conference, U.S. Bureau of the Census, 197-209*.
- Lyberg, L.E., Biemer, P.P., Collins, M., de Leeuw, E. Dippo, C., Schwarz, N. and Trewin, D.J. (eds). (1997). *Survey Measurement and Process Quality*. John Wiley and Sons: New York.
- Lyberg, L.E. and Biemer, P. (1998). Quality improvement in surveys – A process perspective. *Proc of ASA Survey Res. Meth. Sect.* (available from <http://www.amstat.org/sections/srms/Proceedings>).
- Manzari, A. (2004). Combining editing and imputation methods: An experimental application on population census data. *J. Roy. Statist. Soc.* **167**(2), 295-307.
- Mazur, C. (1990). A statistical edit for livestock slaughter data. *US National Agricultural Statistics Service, SRB Research Report No. SRB-90-01, Washington D.C., US.*
- Oakland, J.S. (1986). *Statistical Process Control – A Practical Guide*. Heinemann Ltd, London.
- ONS (2004). Methodology for selective editing. *National Statistics Methodology Advisory Committee paper NS MAC (04) 1*.
- Pullum, T.W., Harpham, T. and Ozsever, N. (1986). The machine editing of large sample surveys: The experience of the World Fertility Survey. *Int. Stat. Rev.* **54**, 311-316.

- Rivière, P. (2000). A general framework for the implementation of quality indicators in a business register. *Proc. of 14<sup>th</sup> International Roundtable on Business Survey Frames*, Auckland, New Zealand. Session 6 paper 2a.
- Rivière, P. (2002). General principles for data editing in business surveys and how to optimize it. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing*, Helsinki, Finland. Working paper 16.
- Rocca, G.D., Luzi, O., Signore, M. and Simeoni, G. (2005). Quality indicators for evaluating and documenting editing and imputation. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing*, Ottawa, Canada. Working paper CRP3.
- Statistics Canada. (2003). *Statistics Canada Quality Guidelines*. Fourth Ed. Catalogue No 12-539-XIE.
- Sutcliffe, P. and Farwell, K. (2005). A general approach to editing. *Proc. ISI Annual Conference 2005*, Sydney Australia.
- Tate, P. (2002). Developing selective editing methodology for surveys with varying characteristics. *Proc. Of Data Clean 2002*, Finland. (available from <http://www.oecd.org>).
- Thompson, K.J. and Hostetter, S.L. (2001). Investigation of selective editing procedures using quinquennial data. *Proc. Of Federal Commission on Statistical Methodology Conference 2001*. (available from <http://www.fcs.gov/01papers/>)
- Underwood, C. (2001). Implementing selective editing in a monthly business survey. *Econ Trends* 577, 41-45.
- UN Department of Economic and Social Affairs. (2001). *Handbook on Population and Housing Census Editing*. Series F number 82.
- UNECE. (1997). *Statistical Data Editing. Volume 2: Methods and Techniques*. Conference of European Statisticians, UNECE Work Session on Statistical Data Editing. Statistical Standards and Studies 48.
- UNECE. (2000). *Evaluating Efficiency of Statistical Data Editing. General Framework*. Conference of European Statisticians, UNECE Work Session on Statistical Data Editing.
- Weir, P. (1997). Data editing and performance measures. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing*, Prague, Czech Republic. Working paper 38.
- Williams, H., Kennedy, E. and Tam, S. (2000). Statistical data editing design principles for household surveys, with application to a computer assisted interviewing (CAI) environment. Technical paper present at Methodology Advisory Committee, Australian Bureau of Statistics (available from <http://www.abs.gov.au>).
- Winkler, W.E. (1997). Problems with inliers. *Proc. of Conference of European Statisticians, UNECE Work Session on Statistical Data Editing*, Prague, Czech Republic. Working paper 22.