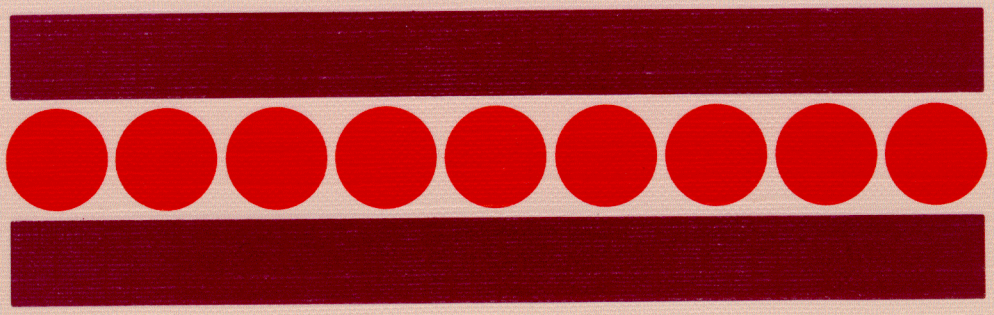
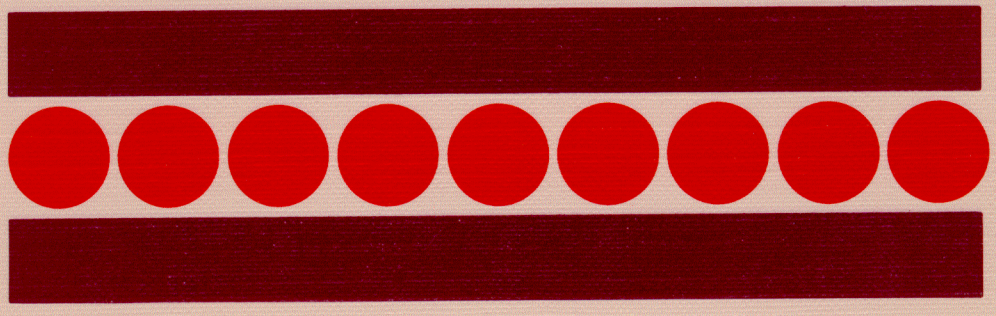
NAZIOARTEKO ESTATISTIKA MINTEGIA EUSKADIN 1991 SEMINARIO INTERNACIONAL DE ESTADISTICA EN EUSKADI



MACRO-EDITING

METHODS FOR RATIONALIZING THE EDITING OF QUANTITATIVE DATA

LEOPOLD GRANQUIST



EUSKAL ESTATISTIKA - ERAKUNDEA
INSTITUTO VASCO DE ESTADISTICA

EUSTAT

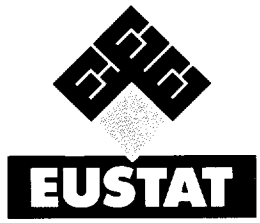
NAZIOARTEKO ESTATISTIKA MINTEGIA EUSKADIN 1991 SEMINARIO INTERNACIONAL DE ESTADISTICA EN EUSKADI

MACRO-EDITING

METHODS FOR RATIONALIZING THE EDITING OF QUANTITATIVE DATA

LEOPOLD GRANQUIST

KOADERNOA 24 CUADERNO



***Lanketa** / Elaboración:*
Euskal Estatistika-Erakundea /
Instituto Vasco de Estadística

***Argitalpean** / Edición:*
Euskal Estatistika-Erakundea /
Instituto Vasco de Estadística
C / Dato 14-16 01005 Vitoria-Gasteiz

***Botaldia** / Tirada*
500 ejemplares
XII-1991

Inprimaketa eta koadernaketa /
Impresión y encuadernación:
ITXAROPENA, S.A.
Araba kalea, 45 - Zarautz (Gipuzkoa)

***Lege-gordailua** / Depósito legal:* S.S. 1.258/91

ISBN: 84-7542-127-10 (o. c.)
ISBN: 84-7749-107-0 (t. 24)

BIOGRAPHICAL SKETCH

Employed at Statistics Sweden since August 1964 at Statistical Methods Unit of the Research and Development Department. During 1983-1989 Head of the Statistical Methods Group of the Department of Economic Statistics. Since August 1989, senior statistician at Statistical Methods Unit of the Research and Development Department and project leader of the Data Editing Project. Since 1990 Deputy Head of the Statistical Methods Unit. Most of the research and development work has been devoted to data editing issues.

Invited speaker/teacher to: UN Workshop on the Statistical Computing Project, Budapest, 1984; IBGE Workshop on Data Editing for Latin America, Rio de Janeiro, February 1990; Statistics Canada Symposium 90, Measurement and Improvement of Data Quality.

CONTENTS

1 EDITING PRACTICES TO-DAY

1.1 CONTENTS

This introductory chapter presents the need and the environment for the particular macro-editing methods which is the object of this report. Thus, the editing process, editing practices and the cost of editing are treated in a very brief and schematic way. The presentation is mainly based on Ferguson (1989), Mayda et al (1990), DEFS (1990): Subcommittee on Data Editing in Federal Statistical Agencies, Granquist (1991)1991.

A more detailed treatment of data editing options and considerations are described in Pierzchala (1990). Another recommended reference (in Spanish) on editing issues and editing procedures is Villan & Bravo (1990).

1.2 EDITING SUBPROCESSES

In Ferguson (1989) data editing is defined as the process for the review and adjustment of collected data. In that paper the process is sub-divided into some major sub-process areas. These areas are:

a) Survey Management (completeness checking, quality control, audit trails, and collection of cost data).
b) Data Entry ("Heads-down" and "Heads-up").
c) Data Review (consists of both error detection and data analysis)
d) Data Adjustment (data editing and imputation)

Survey management

Survey management functions are not data editing functions per se but, many of the functions require accounting and auditing information to be captured during the editing process. Survey management must thus be integrated in the design of data editing systems.

Granquist (1984 b, 4.6.4) claims that it is absolutely necessary for an editing program to include facilities for producing statistical reports on the impact of the editing process. However, frequently the survey management functions of generalized editing programs do not produce statistical reports that are easy to analyze and give a basis for improvement of the total survey design (including the production process).

This fact is one of two very important reasons why one chapter of this report is devoted to measuring the impact of editing on the estimates. The other reason is that results from studies on the impact of editing may serve as a basis for decisions whether to implement one or some of the macro-editing methods which form the subject of this report.

Data entry

Data entry may occur in two modes. "Heads-down" data entry refers to the key-entry of data without any error detection occurring at the time of the key- entry.

"Heads-up" data entry refers to the key-entry of data concurrently with a review of the entered data taking place.

Other terms for Heads-up data entry are "Data entry editing" or "Data capture editing".

Data review

Data review consists of error detection and data analysis.

"Manual" data review may occur prior to the data entry, i.e. the data are reviewed and prepared/corrected prior to key-entry. This procedure is also termed "input-editing" or "input data review" and is typically used when followed by "heads-down" data entry.

Generally, too much resources are allocated to this input editing procedure. This holds true even in those editing systems where the procedure is followed by intensive computer editing. Linacre and Trewin have found that this editing procedure may be counter-productive and sometimes involves "creative editing". Creative editing arises when clerks try to avoid that a record will fail edits in the successive computer editing programme and therefore manipulate data without having a defendable basis for doing so. Besides from being an inefficient procedure, creative editing may give the survey officers a completely wrong idea of the reporting capacity among the respondents.

"Machine" data review involves an immediate review of the questionnaire after adjustments are made. "Interactive editing" is another term used in the literature.

"Computer-Assisted Interviewing" combines interactive data review with interactive data editing while the respondent is an available source for data adjustment. Additionally, data capture (key-entry) might occur at the interviewing time.

"Batch" data review occurs after data entry and consists of a review of many questionnaires in one batch. This procedure is a fundamental part of most of the editing systems to-day.

Data adjustment (Data editing and imputation)

"Manual" data adjustment is the term used when the selection of a more reasonable value is done by a person. It may take place in an input editing process, in data entry editing, in a batch process or in interactive data editing.

"Automated" data adjustments occur as a result of computer actions.

1.3 EDITING PRACTICES

Types of edits

Editing can be carried out at all stages between data collection and imputation, and different kinds of edits may be applied at each stage. There are three principal types of edits: validity, consistency, and statistical or distributional edits.

The role of subject matter specialists

A major aspect of editing today concerns the role of subject matter specialists, who are heavily relied upon for their knowledge of the data. In DEFS (1990) is reported that 60 % of the surveys reviewed refer all the data that fail edits to subject matter specialists. This is a time-consuming and expensive process, involving a great number of skilled professionals.

Integrated editing

The approach adopted by Statistics Canada according to Mayda et al is to integrate editing in the data capture, follow-up and imputation stages. A consistent strategy must be developed so that the edits at different stages are not contradictory. A key to this approach is the use of selective micro-editing, which concentrates on records that have a significant impact on the

estimates. The large or unusual cases identified through this process are referred for follow-up with the respondent.

The cyclical feature of the editing process

Data review and adjustment is a cyclical process (see Bethlehem et al (1989), p.5). The adjustments require further review in order to catch any new inconsistencies introduced by the adjustments.

Minimizing the number of iterations is the subject of much discussion. Of particular note are the ideas of Fellegi & Holt (1976) and the advent of interactive data editing.

The Fellegi-Holt Methodology

In Ferguson (1989, p.8) the Fellegi-Holt Methodology is described as follows: The logic that was used for data review is analyzed by the machine and used to generate machine logic that identifies the minimum set of items on a questionnaire that have to be corrected to resolve an inconsistency. Correction of these items ensure that no new inconsistencies are introduced during the data adjustment.

Mayda et al (1990) describe that in 1976 Fellegi and Holt advocated the following guidelines for editing in surveys: 1) all data on every record should satisfy all edits, 2) as much of the original data as possible should be preserved, 3) the imputation rules should be derived automatically from the edit rules, and 4) the marginal and joint frequency distributions of the data should be preserved.

Some examples of systems based on the Fellegi-Holt Methodology are: DIA from INE, Spain, SPEER from U.S. Bureau of the Census, and GEIS from Statistics Canada. A detailed description of the methodology and these applications is given in Villan and Bravo (1990).

Interactive (on-line) editing

In an interactive data editing environment, the questionnaires are presented and resolved as they are keyed or in one adjustment session only. An example of an interactive editing system is BLAISE by the Netherlands Central Bureau of Statistics (see Bethlehem et al (1989)).

Automated Edit Checking, All Error Correction Done by Analysts or Clerks

This is a short generalized description or definition of the most common editing process. It holds true for 60 percent of the surveys in federal statistical agencies according to DEFS (1990), and certainly for most of the surveys in other statistical agencies.

1.4 COST OF EDITING

Editing has always been expensive. The advent of computers has not decreased the cost of editing essentially, in spite of all efforts to rationalize the editing processes.

According to Pritzker et al (1965) "rough estimates of editing and correction or imputation costs as percent of total costs" for the 1963 Manufactures Census, 1963 Business Census, and the Export Statistics are 18, 18, and 16 respectively.

In 1975, i.e. ten years later, a study of editing costs at Statistics Sweden, is the basis for the following sentence in Granquist (1984 b, 4.3.2). Although micro-editing to a large extent is supported by computer programs, the cost of editing can amount to 40 % or more of the total survey cost, at least in business surveys.

The Subcommittee on Data Editing reports in DEFS (1990) that in their study of editing practices they had found that the median editing cost as a percentage of the total survey cost was 35 %. Editing costs as a percentage of the total survey cost varied greatly by the type of survey. The median for surveys of individuals and households was 20 percent compared with 40 % for economic surveys.

DEFS (1990) states that a key question is: "What is the cost/benefit relationship of this extensive manual review?" The subcommittee recommends the following activities to be undertaken by all survey officers:

* Evaluate and examine the cost efficiency, timeliness, productivity, repeatability, statistical defensibility, and accuracy of the current editing practices versus alternative systems.

* Evaluate both the role and the effectiveness of editing in reducing nonsampling errors for the surveys.

* Evaluate the cost/benefit relationship of extensive manual review on resulting estimates.

1.5 RATIONALIZING THE EDITING PROCESS

Automated Edit Checking, All Error Correction Done by Analysts or Clerks

This is a generalized description of the editing process today at national statistical agencies (see 1.3).

Efforts to rationalize the typical editing procedure are carried out in different ways for example by:

* Removing the cyclical feature by some Generalized Editing System based on the Fellegi-Holt Methodology (example DIA in Spain) or by interactive editing in the adjustment phase or in the data capture phase or in the data collecting phase (example BLAISE in the Netherlands).

* Developing generalized programs which permit subject matter departments to set up their error detecting program themselves (example CONCOR of the U.S. Bureau of the Census).

* Developing tools for the EDP departments permitting them to meet the demands from the subject matter statisticians to get application programs for their editing task in a short time, mainly by specifying parameters (example GODAR from Croatia).

* Integrating the data editing in different phases, for example data capture editing, data editing with follow-ups, automatic imputations, and out-put editing (see Mayda et al (1990)).

* Performing automatic imputations (automated data adjustments), thus reducing the need for manual adjustments.

* Using statistical methods for deducing efficient bounds for the checks, which is the aim of the macro-editing methods presented in this paper, or for selecting the records which require follow-ups (e.g. score functions, see Latouche & Berthelot (1990)).

2 MEASURING THE IMPACT OF EDITING

2.1 BACKGROUND

It has long been apparent at Statistics Sweden, Australian Bureau of Statistics and more recently at Statistics Canada that it is very difficult to get survey managers to change editing procedures, because the editing tradition is very strong at statistical agencies. There is a strong need for different types of documented evidence and arguments to convince the survey staff of the benefit even to question their editing practices.

It is believed that a good way of accomplishing real changes in the editing culture is to provide the survey officers with data on the editing process of the survey for which they are responsible. For the time being, the only way to get such data is to conduct studies on the editing, because the editing process itself does not furnish data of that kind.

2.2 AIM AND CONTENTS OF THE CHAPTER

The object of this chapter is to present a number of reasons why editing procedures should be evaluated or studied and to give hints, ideas or guidelines on how studies might be conducted.

Actually the chapter constitutes a plea that editing procedures in general (i.e. irrespective of the type of procedure applied) should be periodically evaluated as a regular survey operation.

Besides, it has been found that data on the current editing give an excellent basis for assessing whether the type of macro-editing methods discussed here will be a successful weapon against over-editing for the particular survey under study.

2.3 IMPETUS FOR EVALUATIONS OF TRADITIONAL EDITING AND IMPUTATION

The traditional micro-editing practice

The target for the macro-editing methods presented in this report is to rationalize the traditional editing procedure, which in 1.2 above is

characterized as follows:

"Automated Edit Checking on Micro Level, All Error Correction Done by Analysts or Clerks"

This micro-editing procedure is applied in at least 60 percent of the surveys in federal statistical agencies according to DEFS (1990), and certainly in most of the surveys in other statistical agencies. It means that most of editing and imputation systems are covered by this report.

Cost of Editing

The micro-editing procedure is very expensive because all data that fail edits are referred to subject matter specialists. The process is time-consuming and expensive, involving a great number of skilled professionals.

Editing and imputation usually account for 20 - 40 percent of a survey budget (see 1.4). That cost does not include the development and maintenance of software, which in most agencies is made in-house.

The heavy cost of editing is an obvious and quite sufficient reason for continuous evaluations of editing procedures. However, that survey officers do need more arguments is an experience of for example Statistics Sweden and the Australian Bureau of Statistics.

Editing and Quality

The resources spent on editing are always justified by quality arguments. Editing is considered a guarantee that the survey will meet some specific quality requirements.

However, it is hard to find any evaluation of an editing process, which definitely shows that the quality was improved or that the editing was worth the cost. See e.g. Granquist (1991), which is a review of different kinds of evaluations of editing processes, covering about 30 surveys carried out by different statistical agencies all over the world.

In fact, all published evaluation studies of editing processes indicate that editing is counter-productive or has not had any impact on the estimates. In some cases where the estimates were notably changed, a very low percentage of the errors was found to account for almost all of the total change.

If the editing operation has the important role for the data quality claimed by almost all survey officers, then it should be studied whether the aims of the editing really are accomplished in a cost efficient way. Nevertheless, the number of evaluations of editing processes undertaken by statistical agencies is very low. Most of the evaluations reviewed in Granquist (1991) are from the last few years. They were carried out because the ends and means of editing were questioned by some methodologists (not survey managers) of a few statistical agencies.

However, at Statistics Sweden a first step has already been taken to reach the goal of making studies of the editing a regular element of survey processing. All the four subject matter departments have to conduct studies on the editing of one or two important surveys. Guidelines for such studies have been formulated by a Data Editing Project Team at Statistics Sweden. These guidelines are based on the method presented in Greenberg and Petkunas (1987), which will be discussed below. They contain examples from the studies carried out on the Annual Survey of Financial Accounts in Sweden, which is also discussed in a section of this chapter.

Editing pays off

The lack of evaluations in spite of the heavy costs may be due to the belief that editing pays off to such a great extent that it is not necessary to carry out evaluations. This is understandable because it is rather easy to arrive at the conclusion that micro-editing always pays off.

In every survey there are apparent errors in the data to be micro-edited. These errors are easily detected and a great many of them can be corrected. Evidently the number of errors will decrease by the micro-editing operation. Even if the error rate of every variable is low, many questionnaires or records are affected (see Granquist (1982)). Sometimes errors of great magnitude are detected and corrected. Then the quality is somewhat improved. When serious errors are detected after the editing phase or after the publishing of the results, i.e. when the quality of the edited data is found to be poor, this is considered to indicate that more resources should be spent on the editing of that survey.

All this leads to the opinion that micro-editing is essential, and that it is quite unnecessary to investigate if the resources spent on micro-editing really pay off or if the resources could be more rationally allocated.

Over-editing

Every traditional editing procedure contains a vast number of checks with too narrow bounds. They produce a great number of error messages, many of which imply expensive contacts with the respondents. Instead of focusing the checks on possible important errors, the experts invent all possible kinds of checks, viz all that can be checked is checked. This particular problem, nowadays termed over-editing, is faced by a growing number of statistical agencies.

A very striking example of over-editing is the machine editing of the World Fertility Survey. The conclusion given in Granquist (1988) in his review of the report by Pullum et al is that:

(i) The machine editing had no impact at all on the analysis
(ii) The machine editing caused a delay for every national survey of one whole year.

Editing should improve the knowledge of the survey data

Another serious objection against the application of a traditional editing procedure is that nothing is learnt about errors, error structures, and the problem areas of the survey design. According to Granquist (1984) this is the essential role of editing.

Editing should be a source of information for rationalization

Production of statistics, as all production of goods or services, has to be as rational as possible. Statistical agencies have the responsibility to produce the best possible statistics within the limits set by available resources.

A great part of the total survey costs is due to errors both in the collecting phase and in the processing phase. Accordingly one important way of rationalizing a survey is to prevent errors. The editing process is probably the best source of information about the errors of a particular survey, because one aim of editing is to detect errors. Thus, if the detected errors are analyzed to find appropriate preventive measures and an appropriate editing tool is found to fight the remaining important errors, the editing process might have a nice effect.

Summary

A number of reasons for undertaking evaluation studies on editing processes has been indicated. This is in perfect accordance with the following recommendations given in DEFS (1990) which are based on the findings of a study of editing practices covering 117 surveys at statistical agencies in United States:

All survey officers should

* examine and evaluate the cost efficiency, timeliness, productivity, repeatability, statistical defensibility, and accuracy of the current editing practices versus alternative systems.

* evaluate both the role and the effectiveness of editing in reducing nonsampling errors for the surveys.

* evaluate the cost/benefit relationship of extensive manual reviews on resulting estimates.

2.4 MEASURING THE IMPACT OF EDITING - METHOD AND STUDIES

2.4.1 Introduction

Studies of distributions of errors and error types form a valuable basis for designing more efficient editing methods. They may also help to persuade subject-matter experts to accept macro-editing methods in their surveys.

A data error analysis which focuses on the impact on the estimate by the errors classified by size may show the potential value of implementing macro-editing methods. Such an analysis is reported in Greenberg and Petkunas (1987).

The method used by Greenberg and Petkunas is presented in 2.4.2 together with some graphs from the evaluation of the editing procedure used in the U.S. 1982 Economic Censuses. This method has also been used, albeit in modified versions, in some studies carried out at Statistics Sweden and at Statistics Canada. These are described and commented in 2.4.3 and 2.4.4 respectively. Finally, there is a brief description of a simple method for studying various types of errors detected in editing, exemplified on data

from the Swedish Survey on Financial Accounts.

2.4.2 The Greenberg-Petkunas Method

Within the evaluation study all records from six selected kind-of businesses (KB's= a sub-classification in the SIC) were analyzed after each run through the edit and imputation cycle. An immediate finding was that relatively few records accounted for a very great part of the total change.

To study the phenomenon more in detail, a change file was established. It consisted of all the records for which the keyed-in reported value was different from the tabulation field value.

X_i= reported value for case "i"

Y_i= tabulation field value for case "i"

i= 1,.....N (N= number of cases (changes))

$d_i = | x_i - y_i |$

The cases were ordered by the absolute value of the difference, i.e.:

if $i \geq j$ then $d_i \leq d_j$

$$D = \sum_{i=1}^{N} d_i$$

Then d_i / D = the proportion of the total change contributed by the i^{th} case.

q_i and p_i were defined:

$$q_i = \left(\sum_{j=1}^{i} d_j / D \right) 100\ \%$$

$$p_i = (i / N)\ 100\%$$

for i= 1,...N.

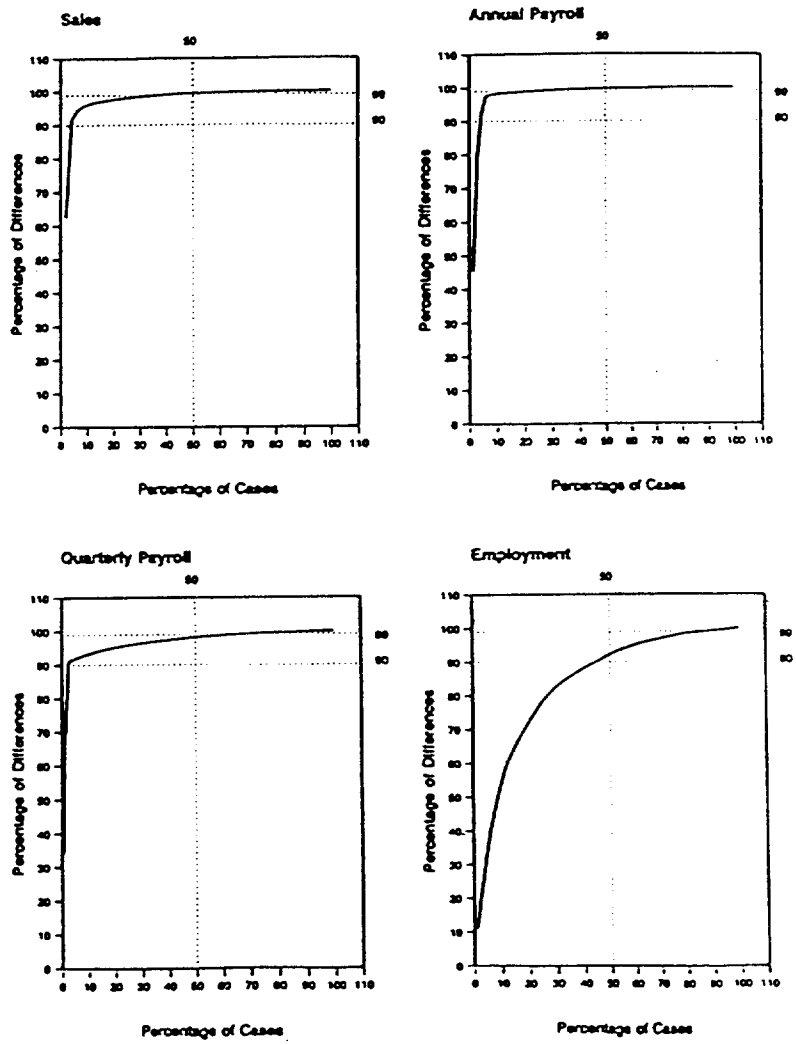
q_i represents the percentage of total change contributed by cases for which the change was equal to or greater than the change for case i.

p_i represents the percentage of cases in the change file for which the change was greater than or equal to the change for case i.

Some results are shown in the graphs below (see Figure 2-1).

For all items studied and all the six KB's, approximately 5 per cent of the cases contributed over 90 per cent of the total change. Many of these large changes were due to reporting in units rather than in thousands.

FIGURE 2-1 shows graphs of the percentage of total change by the percentage of cases for SIC 783300. (Reprint from Greenberg and Petkunas (1987))



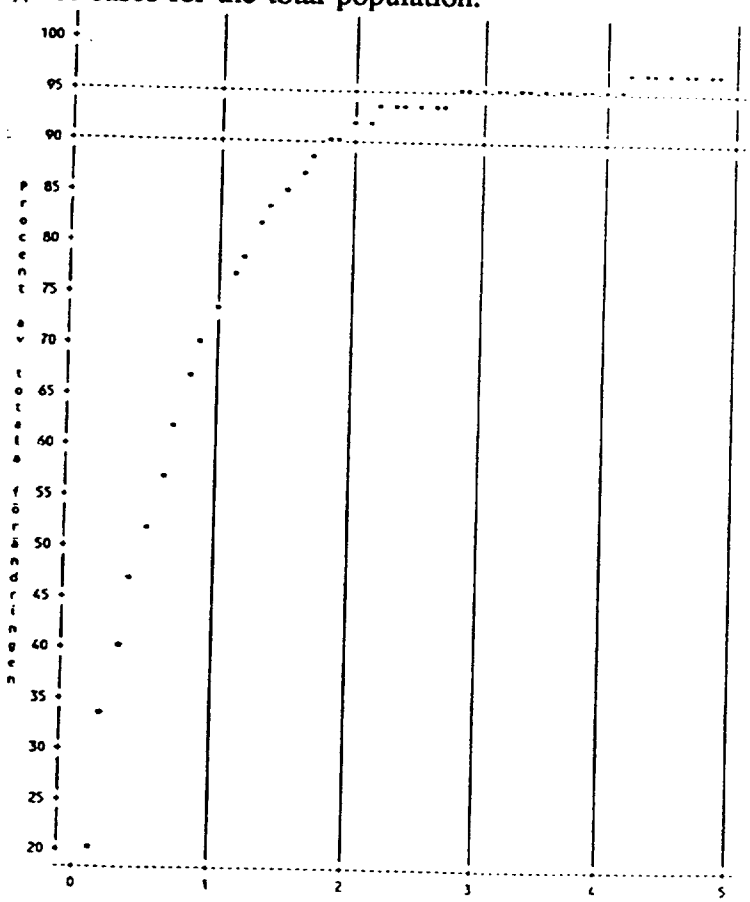
2.4.3 The Study on the Swedish 1987 Survey on Financial Accounts

The study covered the editing of all enterprises whith more than 49 employees. The population size was a little more than 4 000 and the survey was edited by a traditional micro-editing procedure with ratio, validity and consistency checks. Inter-active editing was applied (see 1.2, data review).

For three selected variables all records were analyzed after the whole editing procedure was completed. For the analyses the Greenberg-Petkunas method was used.

Figure 2-2 shows a graph for the variable "Value added". The results are quite in accordance with the findings reported in Greenberg and Petkunas (1987). It should be noted that nearly one fourth of the reported values on the item "Value added" were changed by the editing procedure.

FIGURE 2-2 shows graphs of the percentage of total change by the percentage of cases for the total population.



The modified measuring method

The changes were also related to the estimated value of the variable. This was accomplished by defining p_i as above and :

N_s = the number of unchanged records (approximately 3000)

$d_i' = x_i - y_i$

$$D' = \sum_{i=1}^{N} (x_i - y_i)$$

S = the total of all (N_s) unchanged values

C(E) = the total of all changed values after editing (y_i)

C(R)= the total of all changed values as originally reported (x_i).

$$r_i = \left(\sum_{j=1}^{i} d_i' + S + C(R) \right) / \left(S + C(E) \right) 100\%$$

which means that r_i = the percentage of:

("the sum of all positive and negative changes contributed by cases for which the change was equal to or greater than the change for case i" + "the total of all cases as originally reported")

to

("the total of all values after editing")

This means that:

r_0 = the total relative change

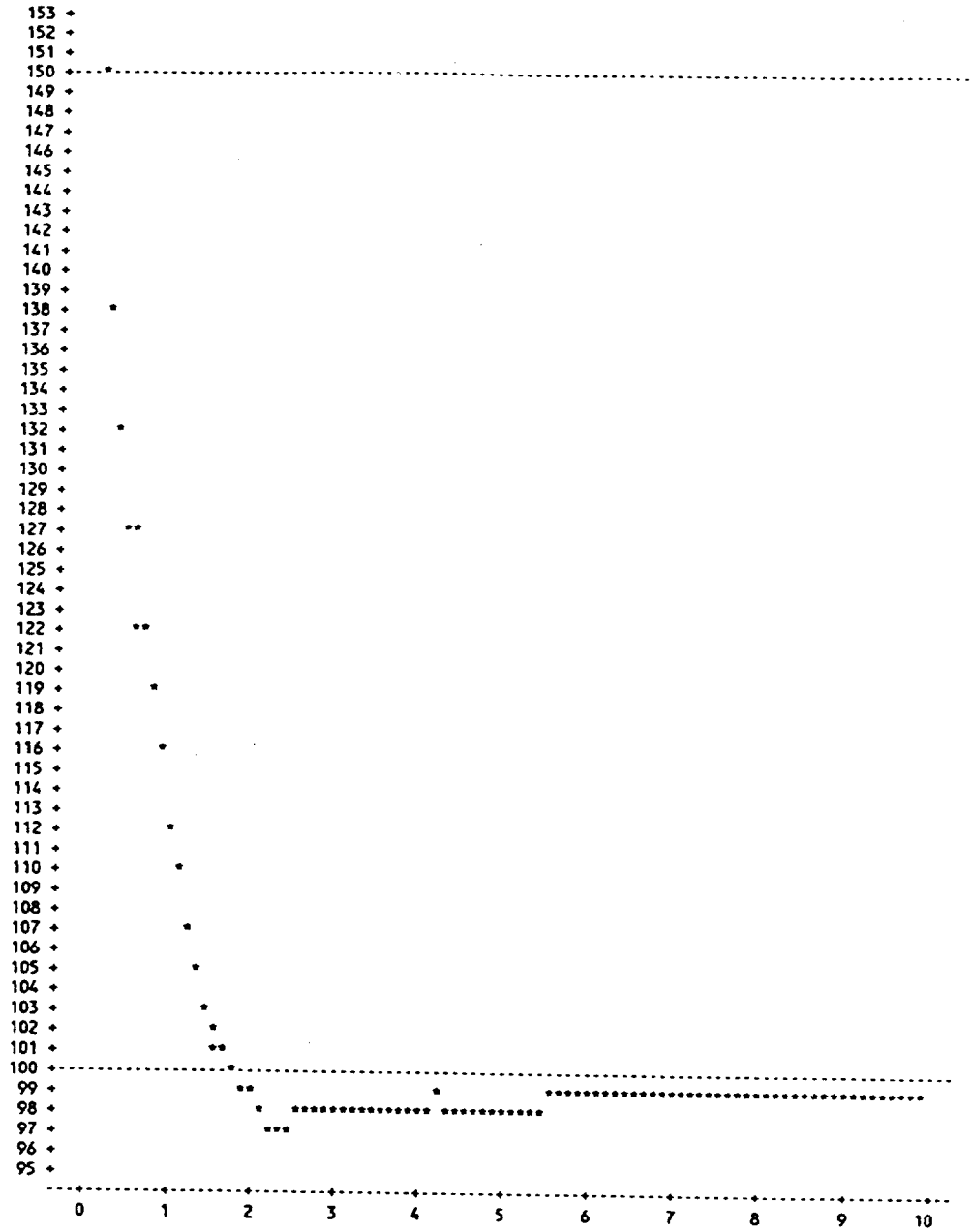
$r_{999} = 100$

(999 was the number of changes for this item of the survey)

By this modified method the same data as represented by the graph in

Figure2- 2 is displayed by the graph in Figure2-3

FIGURE 2-3



2.4.4 The Study on the Micro-Editing for the Annual Survey of Manufactures in Statistics Canada

Boucher (1991) describes the results of some studies undertaken at Statistics Canada on the impact of micro-editing in the Annual Survey of Manufactures (ASM).

2.4.4.1 The ASM Editing System

The ASM is an establishment based survey covering all operations carried out under one ownership at a single physical location. These establishments are classified according to the Canadian Standard Industrial Classification (CSIC).

The study was limited to the about 11000 establishments which are required to respond to a long questionnaire. Each record is processed and followed-up on an individual basis and subject to all editing steps without discrimination. Individual follow-up takes place in cases of non-response and as reaction to edit failures. The edit process is linear and iterative. In some cases a record may be manually handled as many as 20 times.

The first stage of the editing is the manual input data review. Clerks make every effort to ensure that the data will pass the machine editing and to avoid rejects. Then the individual records are submitted to the Questionnaire Information Processing System (QUIPS), which includes data entry and machine editing. Follow-ups usually take place during the QUIPS stage. There is also an out-put editing termed pre-analysis, which can result in telephone follow-ups to confirm or correct the data.

2.4.4.2 The Study

Most ASM records fail various edits and require extensive cleaning. Before the study the impact of such processes had never been evaluated. The main object of the study was therefore to quantify the impact.

Because of the input data editing, the original data provided by the respondents had to be reconstructed to make it possible to measure the impact of the editing and processing. As a first step, the study focused on observing, measuring and analyzing selected "Principal Statistics". A sample of establishments from the 1988 ASM was selected.

Boucher (1991) presents a comparison of data at two points in the

processing:

Raw data, which essentially represent the reconstructed information as reported by respondents prior to any editing, coding or edit follow-up.

Final data, which are the actual data after coding, editing and follow-up by the operations staff.

The study found a relatively high incidence of reporting in dollars when in fact responses are requested in thousands of dollars. An adjustment was made remove the effect of this reporting bias prior to calculating and analyzing the impact of editing.

It should be noted that this adjustment is one of the two main differences to the Swedish study (see 2.4.2). (The other difference is that the records where item-nonresponse occurred were not excluded from the raw and the final data files.)

2.4.4.3 Findings

* When variables are taken individually, there appears to be scope for rationalizing the editing with minimal impact on the quality of the estimates.

* The distribution and impact of the editing efforts vary significantly among industries and establishments.

* Three out of four edit changes result in increasing the value of the reported data.

* On the average, three mechanical edit iterations are necessary to validate individual units, even after extensive manual editing and follow-up.

* Questionnaire design problems are responsible for a significant number of respondent errors.

The editing impact on the selected items on the total level is shown in Table2-1.

TABLE 2-1 shows the editing impact on each selected field for all the establishments in the study. The "impact" represents the aggregated net difference between the "final" and the "raw" data expressed in terms of the final.

VARIABLE	$ RAW	$ FINAL	% IMPACT
FUEL & ELECTRICITY	2.992	2.956	-1.2
RAW MATERIALS	64.931	66.315	2.1
SHIPMENTS	103.140	105.222	2.0
WAGES	9.340	9.622	2.9
EMPLOYEES	267	271	1.7

(Values in millions of Canadian dollars)

In order to give a basis for assessing whether and to which extent there was any unnecessary editing in the current micro-editing process the cumulative effect of correcting errors was analyzed, from the largest to the smallest edit change, in absolute terms.

For each CSIC industry, the study presents the number of establishments in the sample, the industry level impact and the number of establishments with changes contributing to this impact. Then are calculated, **"the largest changes needed"**, representing the level of changes required (percentage and count of changes) to bring the estimate within 0.25%, 0.50% and 1.0% of the actual "FINAL DATA", had the changes been made to the corresponding largest contributors.

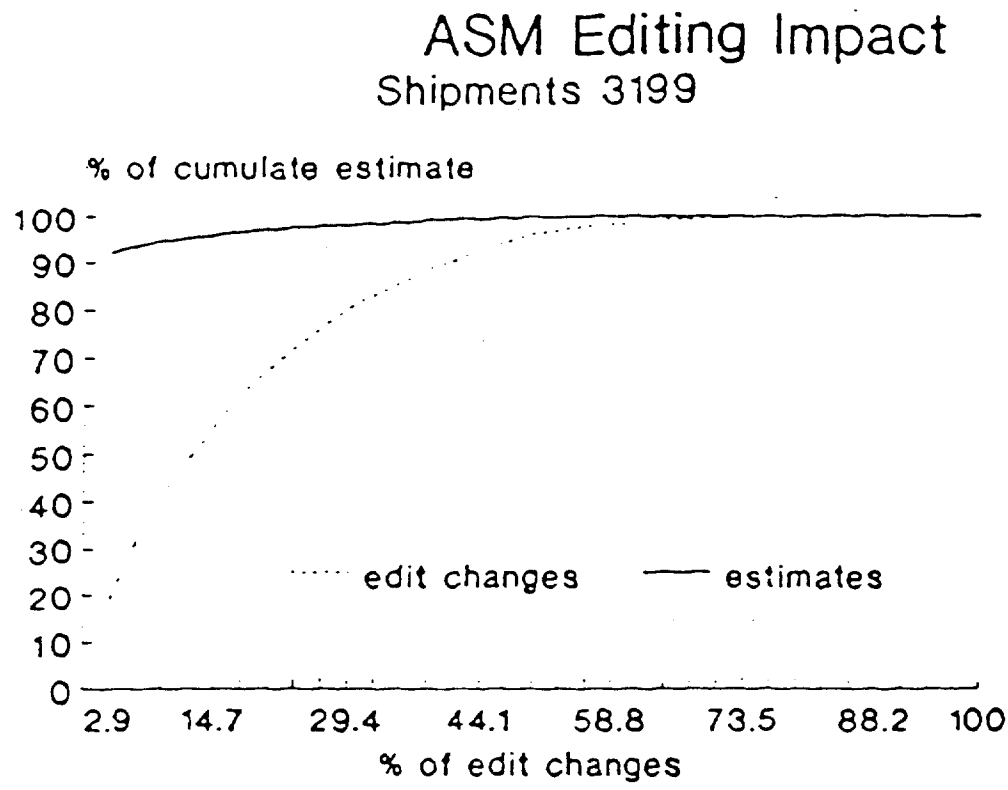
TABLE 2-2 shows the largest changes needed as a measure of editing contributions for shipments.

Industry	# of cases with changes	+/- 0.25% of final		+/- 0.50% of final		+/- 1.0% of final	
		%	#	%	#	%	#
Machinery	5	80.0	4	80.0	4	60.0	3
Sawmills	54	44.4	24	37.0	20	24.1	13
Petroleum	6	33.3	2	16.7	1	16.7	1

A series of plots were produced to indicate the relative contribution of the changes made during the editing for each variable and industry. Figure 2-4

is a graphical representation of table 8 for Machinery (CSIC 3199).

FIGURE 2-4 shows the estimate curve (the top curve) and the "edit-changes" curve (the bottom curve) on total level (see the explanations below).



Explanations

Following the Greenberg-Petkunas method (see 2.4.2) the cumulative changes, in absolute terms, constitute the *"edit changes-curve"* shown as the bottom curve. Its first point represents the largest individual edit change, at the establishment level. The last point represents 100% of the total absolute changes, i.e. all the changes done.

The top curve (*estimate-curve*) shows how close the estimate of a given data field would have been to the final data, in relative terms, had the changes been made only to the corresponding largest contributors. Here, the modified Greenberg-Petkunas method is used (see 2.4.3).

2.4.4.4 Conclusions

The results obtained at the national/industry level clearly indicate that it is sufficient to deal with the largest establishments with edit failures to ensure acceptable data quality. This permits substantial resources savings.

An editing alternative for ASM is outlined based on "selective edit follow-up". The authors believe that they can develop selection criteria, a minimum set of edits, and a macro-editing package that will be operational and lead to resource savings and improved timeliness in the ASM program.

2.4.5 A Feasibility Study on Measuring the Impact of Editing Changes

2.4.5.1 Background

A feasibility study on estimating the impact of types of editing changes has been carried out on the Swedish 1987 Survey on Financial Accounts. The aim of the study was to elaborate guidelines for estimating the impact of errors found in editing. The underlying reasons are expressed by the following two sentences formulated and discussed in 2.3:

Editing should improve the knowledge of the survey data

Editing should be a source of information for rationalization

A great part of the total survey costs is due to errors both in the collecting phase and in the processing phase. Accordingly, one important way of rationalizing a survey is to prevent errors. If the detected errors are analyzed to find appropriate preventive measures and an appropriate editing tool is found to fight the remaining important errors, then the survey will be conducted in a more efficient way.

Guidelines have now been elaborated and are supposed to serve as a help for the subject matter departments at Statistics Sweden to undertake studies on the impact of editing (see 2.3, editing and quality). The report on the

feasibility study (Forsman (1991)) is annexed to the guidelines as an example.

2.4.5.2 The Study

The study was made on the same data as were used in the study reported in 2.4.3. It covered all enterprises whith more than 49 employees. The population size was a little above 4 000. The item Value Added was selected. A Poisson sample was taken. The records for which the value had been changed by the editing process were identified. The reasons for the changes were classified by 5 different types by the most experienced statisticians. It was a difficult task because rather a long time had elapsed since the data were edited. However, the study succeeded. Then the number and total for each change reason were estimated.

2.4.5.3 The Method as a Current Editing Operation

The clue of the method is the classification of error types (change reasons). This should be done as a part of the editing. Items should be selected and known error types listed. Then the most serious error types should be selected and when changes are made during the editing all changes should be classified according to the selected types. It then becomes very easy to estimate the impact of every single selected error type. A procedure like this has been suggested by the Data Editing Project Team to become a compulsory operation in survey processing at Statistics Sweden.

3 WHAT IS MACRO-EDITING

3.1 CONTENTS

A discussion of errors, according to Granquist (1984 b), will here serve as a background to the concept of macro-editing as it was introduced in Granquist (1984 a). Then a working definition of macro-editing adopted by Statistics Canada is presented together with a few explanatory notes according to Chinnappa et al (1990). The two definitions of macro-editing are compared and a definition of the type of macro-editing methods described in this paper is presented as a conclusion of findings from the first attempts at Statistics Sweden to elaborate a macro-editing procedure on live data.

3.2 ERRORS

3.2.1 Contents

The concepts of negligence errors and misunderstanding errors are introduced in order to facilitate a discussion of the micro-editing and macro-editing concepts. The contents of this section constitute a summary of the discussion of errors in Granquist (1984 b).

3.2.2 Negligence errors - Some Definitions

Discernible errors

Errors are classified as "discernible", when the values obviously are incorrect from a formal or theoretical point of view.

Identified errors

A discernible error is denoted as an "identified" error, if the erroneous observation can be replaced by the correct value (be corrected) without going back to the source of information (e.g. the respondent).

Suspicious errors

An observation is "suspicious", if it for some reason is conspicuous. The errors which are discovered by inspection of "suspicious" observations are

designated as "suspicious"errors.

Negligence errors

The discernible and suspicious errors as "defined"above, will here be designated as "negligence" errors. A "negligence" error is the result of carelessness either by the respondent or by the survey process up to the editing phase. In a repetition of the survey the same error should probably not be found for the same item of the same record (questionnaire).

3.2.3 Characteristics of Misunderstanding Errors

Misunderstanding errors are errors that arise due to ignorance or misapprehension of questions, concepts or definitions, but they also comprise tactical errors, e.g. when respondents deliberately give inaccurate answers for some reason (to mislead competitors, gain benefits, protect their privacy). In contrast to negligence errors, which arise randomly, a specific misunderstanding error will probably affect the same item of the same record if the survey is repeated under the same conditions. These errors will usually be repeated in periodic surveys, if measures are not taken to avoid them.

A particular misunderstanding error usually affects specific groups of respondents, i.e. there are groups of respondents which have a rather high probability to commit (or to be affected by) the error.

3.2.4 Classification of Errors

Thus, errors are here classified either as negligence errors or as misunderstanding errors on the basis of the two dimensions:

number of cases, and

reason for the error.

Misunderstanding errors are those that affect a number of records and are consequences of ignorance, poor training or misunderstanding, or are committed deliberately for some reason or other.

All other errors are classified as negligence errors.

3.3 MACRO-EDITING AS A WEAPON AGAINST MISUNDERSTANDING ERRORS

3.3.1 Introduction

The difficult editing problem are the unknown misunderstanding errors. The problem of finding unknown misunderstanding errors (though usually not expressed in these terms) is commonly tackled by using a great number of edits (a mass checks approach) with as tight bounds as possible. This editing "strategy", which is applied in numerous surveys by almost all the statistical agencies of the world, may be considered as arisen from a more or less unconscious adoption of the "safety first" principle. This principle is based on the assumption that the more and tighter the checks, the better the resulting quality. Anyway, the "strategy" will certainly lead to extensive over-editing.

This mass checks approach is not at all an efficient method for detecting negligence errors and is far from being an acceptable method for tackling misunderstanding errors. In respect to misunderstanding errors, it is an indirect method because it is not focused on the specific problems of the actual survey.

What is worse is that this approach very easily conveys an illusory confidence in the quality of the data to the survey staff. They may then not consider it worthwhile to review carefully and critically both the data and the survey operations.

3.3.2 The Macro-Editing Approach

The problem of unknown misunderstanding errors should be tackled by "macro-editing" as the concept is defined by Granquist (1984 a).

In short, macro-editing in that sense means

(i) to identify problem areas

(ii) to estimate and document the impact of the problem areas

(iii) to take measures to counteract or adjust for the misunderstanding errors in the current survey, if possible, and otherwise store the knowledge for future surveys.

The prerequisite of macro-editing is a solid knowledge of both the subject matter and the production system. Consequently macro-editing, aimed at finding misunderstanding errors, should be carried out by survey statisticians.

3.4 THE MACRO-EDITING CONCEPT AT STATISTICS CANADA

3.4.1 Definition

Macro-editing is defined as the detection of errors in data through the analysis of aggregated data.

In Statistics Canada, macro-editing is known by several names including certification, validation, analysis, reconciliation, integration, and consistency checking.

3.4.2 Discussion of the Macro-Editing and Micro-Editing Concepts

According to Statistics Canada, macro-editing implies a comparison of aggregated data or estimates at aggregated levels (e.g. publication cells)

(a) within a survey (or data source) or

(b) with information from other sources in order to detect inconsistencies, anomalies or errors in data.

Micro-editing, on the other hand, is the detection of errors in data through checks of individual records.

Micro-editing can contribute to data analysis and to improving survey processes, because it helps in identifying reasons for variations and uncovering sources of error. However, at Statistics Sweden we consider this opinion of Statistics Canada, valid only in theory (cf. 3.3.1). In fact, we have not yet found any example of manual review work that has resulted in a feed-back to the survey officers for making changes in the editing systems.

In Chinnappa et al (1990) it can be read: "Whereas, traditionally micro-editing has been a labour intensive and time-consuming operation where each individual record or questionnaire is cleaned up through verification

and checks, there is much room for efficiency by concentrating the micro-editing efforts on only those records and variables which have a large impact on major estimates". It should be noted that this sentence constitutes in a very concentrated form the background and the aim of the type of macro-editing methods considered in this report.

Macro-editing in Chinnappa et al (1990) is described as follows. "In the survey process, after micro-editing, correction and imputation, preliminary aggregated estimates are prepared by the data gathering unit. Analysts subject these preliminary survey estimates to macro-editing usually based on time series of estimates from the same survey or known models that these estimates are expected to satisfy in order to detect any major errors or inconsistencies. Such analyses lead to further examination of the estimates at lower levels in an effort to identify the causes of the suspected errors. The focus here is on making adjustments at the micro level. At this stage, records that contribute the most to the estimates or that show highest changes from the previous occasion may be examined to isolate outliers and errors that were missed by the micro edits. Macro-editing could thus result in the detection of outliers (which may be adjusted) and errors at the micro-level (which are corrected)."

As can be seen in the following this particular role of macro-editing is exactly the aim of the type of macro-editing methods which are the subject of this paper. The difference is that at Statistics Sweden we have found that macro-editing with the described aim can replace traditional micro-editing methods and be carried out during the same phase of the data processing as micro-edits.

According to Chinnappa et al (1990) the methods of macro-editing at Statistics Canada are divided into five general categories, as follows:

(i) Comparison with like information from other sources

(ii) Time series analyses

(iii) Structural analyses of the data

(This means comparisons of similar and dissimilar subgroups of the data. For similar subgroups similar characteristics should be expected and for dissimilar subgroups, different characteristics should be expected. For example, an unusually high fertility ratio in the younger female population would be a reason to investigate possible sources of error.

(iv) Comparison with related information

(v) Comparison of aggregates calculated using accounting identities.

At Statistics Canada they state that "micro-editing is one of the many ways of checking for errors. However, checks of this type, although effective in controlling errors on data we collect, tell us little about the data we should, but do not, collect." This and the following quote from the paper of Chinnappa is completely in agreement with our points of view concerning the other and much more important role of macro-editing, namely to detect misunderstanding errors.

"Macro-editing plays a fundamental role in the statistical process in as much as it is in most cases, the only independent means of verifying the quality of our figures. Although direct comparisons can rarely be made, they can be extremely useful in signalling whether coverage is drifting, definitions are no longer relevant, etc. Macro-editing is therefore not only considered desirable but an essential component in keeping Statistics Canada products relevant and maintaining high quality standards."

3.5 MACRO-EDITING AS A WEAPON AGAINST OVER-EDITING

3.5.1 Background

In Granquist (1984 a) macro-editing is introduced as a means for tackling unknown misunderstanding errors.

At Statistics Sweden we started experimenting with macro-editing methods by performing edits on aggregates. When a suspect aggregate was found, the first step was to find out whether there were one or two individual observations which caused the aggregate to be classified as "suspect". When working on means for identifying such data, we found that the macro-editing methods were more efficient than traditional micro-editing methods in detecting suspect data. Then we decided to work on this first step of the macro-editing concept only, until we managed to get such methods extensively implemented at Statistics Sweden.

From our experiences, we have found it appropriate to define macro-editing (this type of macro-editing methods) as "statistical edits" applied to expanded data. The methods which are reported in this paper may be characterized as methods to provide micro-editing methods with more

efficient acceptance bounds. They bring a priority thinking to the verifying work.

3.5.2 Definition on Macro-Edits for Detecting Errors at Micro-Level

This paper deals with checks on quantitative data which flag "suspicious" data for manual review. This type of checks may be considered the opposite of validating checks which indicate data that are erroneous. Ferguson (1989) calls the first type "Statistical Edits". They use the distributions of current data from many or all questionnaires, or historic data of the statistical unit, to generate feasible limits for the current survey data.

In this paper macro-editing imply a procedure for pointing out suspicious data by applying statistical checks/edits based on the weighted keyed-in data. The upper and lower limits of a macro-editing check (macro edit) should be based

only on the data to be edited and

on the importance of the data on the total level.

4 MACRO-EDITING METHODS

4.1 INTRODUCTION

4.1.1 Objective and Contents

The objective is to present macro-editing methods as a weapon against over-editing. That is why emphasis has been given to the rational aspects of macro-editing as compared to micro-editing. The presented experimental studies in data processing environments and in production show that these method are superior to micro-editing methods in editing quantitative data. Savings of up to 70-80 percent of the manual review work are quite feasible.

This chapter is mainly devoted to descriptions, studies and results of *"The Aggregate Method"*, *"The Hidiroglou-Berthelot Method" ("Statistical Edits")* and *"The Top-Down Method"*. These methods are described in different ways, each one serving a special purpose. First there is a brief, schematic description aimed at presenting the basic ideas or principles in an easy way. Then there is a detailed description of the method in connection with the presentation of some studies or applications. The purpose is to provide sufficient information to serve as guidelines for an implementation. Finally the method is described in such a way as to serve as a basis for a programmer to implement it into an application.

"The Box-Plot Method" and *"The Box-Method"* are also reviewed, but as modifications or developments of the Aggregate Method and the Top-Down Method respectively. Because studies on these methods have not yet been undertaken by Statistics Sweden, they are described only in the schematic form.

"The Cascade of Tables Method" developed by Juan Pons Ordinas (see Ordinas (1988)) for the editing of the Survey of Manufactures in Spain is only given as a reference in the reference list. The reason is that it is somewhat different from the type of macro-editing methods treated here. We classify it as an output editing method.

A common evaluation method has been utilized in all the studies reported below. Another common feature is that the simulations and the applications have been carried out on data from only two surveys. The evaluation

technique and these two surveys are consequently presented in this introduction to the chapter.

4.1.2 The Evaluation Technique

The studies on the methods reported below have been simulation studies on real survey data. The results have been compared with the results of the micro-editing methods applied when the survey was processed. The changes made as a result of the micro-editing process were entered on a change file and the study consisted in investigating (by calculating a few measures) which data in the change file were flagged by the macro-editing method and which were not. The rationalizing effect was measured as the reduction in the number of flagged data, and the "quality loss" as the impact of the remaining errors (the errors found by the micro-editing of the survey, but not by the macro-editing method under study)

4.1.3 The Experimental Data

The studies were carried out at Statistics Sweden and used data from the Survey on Employment and Wages in Mining, Quarrying and Manufacturing (SEW) and the Survey of Delivery and Orderbook Situation (DOS).

4.1.3.1 The Survey on Employment and Wages in Mining, Quarrying and Manufacturing (SEW)

The survey is a monthly sample survey on employment and wages in the Swedish industry. The number of reporting units (establishments) is about 3000. The main variables are: *number of workers, number of working hours, the payroll and wages/hour.*

A traditional editing procedure is applied, but with inter-active up-dating, which includes checking against the edits. This editing procedure was recently revised by extending the acceptance limits of the ratio checks, co-ordinating all the checks and tuning the checks. The verifying work was then reduced by 50 % without any drop in quality. It should be noted that the rationalizing effect of the macro-editing methods studied here is compared to the revised procedure.

4.1.3.2 The Survey of Delivery and Orderbook Situation (DOS)

The survey is a monthly sample survey of enterprises. There are about 2000

reporting units (kind-of-activity units) in a sample drawn once a year from the Swedish Register of Enterprises and Establishments. DOS estimates changes in the deliveries and the orderbook situation (changes and stocks) for both the domestic and the foreign market (six variables) for the total Swedish manufacturing industry and for 38 divisions (classified according to the Swedish Standard Industrial Classification of all Economic Activities).

The questionnaire is computer printed. The entry of the questionnaire data is carried out in three batches every production cycle by professional data-entry staff. The Top-Down macro-editing method is applied.

4.2 THE AGGREGATE METHOD

4.2.1 Contents

A somewhat generalized description of the aggregate method is first presented (an edited version of 6.1.2 in Granquist (1991)). The description is to serve as an overview of the method and contains some generalized conclusions from the studies.

Then, from Granquist (1988 b) a report is given of the studies of the original method carried out by a main-frame prototype applied to the Survey on Employment and Wages (SEW). There is a short description of the SEW editing system and a revision of it made in 1985.

A micro-computer application of a somewhat modified version of the Aggregate Method is then reported in such a way that it may serve as a detailed guide for implementing the Aggregate Method. All principal steps are carefully reported together with some facts and figures from a study on SEW data. (It is an edited version of Lindström (1990 a)).

Finally there is description of the method as an EDP function (a cut from Lindström (1990 b)).

4.2.2 A Generalized Description of the Aggregate Method

Editing on aggregate level followed by editing on micro-level of the flagged aggregates

The basic idea is to carry out checks first on aggregates and then on the individual records of the suspicious (flagged) aggregates. All records belonging to a flagged aggregate of any of the variables form the error file. The checks on the individual level are then carried out on that error file.

Another realization (which is studied) of the idea is to give the records of a flagged aggregate a specific signal telling which of the variables that does not pass the edit. The check on the micro level is then applied only to those questionnaires which belong to the aggregates which fail the check for that variable.

The acceptance bounds are set manually on the basis of the distributions of the check functions

The most essential feature of the aggregate method is that the acceptance limits are set manually by reviewing lists of sorted observations of the check functions. Only the "s" largest observations and the "m" smallest observations of the check functions are printed on the lists.

Both the check function A on the aggregate level and the check function F on the individual level have to be functions of the weighted (according to the sample design) value(s) of the keyed-in data of the variable to be edited. By using the weighted values in function A, the checks on the aggregate level can be calculated in the same way as is done in traditional micro-editing.The macro-editing process can then be run as smoothly as a micro-editing process.

The lists of the sorted observations can be used either directly as a basis for reviewing observations manually (if identifiers are printed out together with the observations) or indirectly as a basis for setting acceptance limits for an error detecting program, which then can produce error messages of suspected data for manual reviewing. The advantage of using the error-detecting program is that to the reviewer the process can be made to look identical to the old one. To implement the Aggregate Method, only programs for printing the lists of the sorted observations are needed. Such programs can easily be added to the old system.

Improvements by providing the lists with statistics or graphs

Obvious improvements are to provide the lists of the distribution tails with such statistics as the median, the quartiles, the range, interquartile range or with graphs. Here, Box-plots (see Tukey (1977)) are recommended.

Some results and findings

This macro-editing concept is a realistic alternative or complement to micro-editing methods, which reduces the amount of manual review work to a considerable extent (by 34, 66 and 80 per cent in below reported studies), without losses in quality or timeliness.

The acceptance intervals of the checks should be wider and should not be symmetric around 1 for ratios and around zero for differences, when ratios and differences are used as check functions.

The best strategy is to set the limits as close as possible to the first outlier on both sides. The definition of "extreme out-liers" by Tukey (1977) may here serve as guidelines for efficient limits according to our findings, which is confirmed in Anderson (1989 b).

From the studies of the modified version of the Aggregate Method (see A4) it is found that the aggregate checks can be skipped, when there are no problems with either the storage capacity or the computer cost. Then it is recommended to provide the tails of the distribution of the check function with Box-Plots (see Tukey (1977)).

4.2.3 The Main-Frame Application on the SEW

4.2.3.1 A Short Presentation of the Survey

The SEW is a sample survey on employment, absences, discharges and wages in mining, quarrying and manufacturing. Data are collected every month from about 3000 establishments belonging to about 2500 enterprises. They report the number of workers on a normal day during the end of the month, the number of working hours for a chosen period, the payroll for those working hours, the number of working days in the month and data about absences, newly recruited, discharges and overtime. The Swedish standard industrial classification of all economic activities (SNI) is applied. It is identical with the 1968 ISIC up to and including the four digit-level and

has in addition a two-digit national subclassification.

4.2.3.2 Present Editing Procedures

To-day a traditional micro-editing procedure is used, which means that a computer programme is run in batch with checks prepared in advance pointing out data as "suspicious" or not consistent with other data. The reconciliation is completely manual, but as data entry is carried out interactively, changed data are shown directly on the screen and checked against the programme.

This computer assisted editing is preceded by a manual pre-edit process, aimed at finding out if the questionnaire is possible to run and to facilitate the data entry. However, this pre-editing work is much more comprehensive than necessary.

The data entry is carried out in another department, without any editing at all. The idea of integrating the pre-edit and the data entry operations has not yet been accepted.

The questionnaires are processed in lots on arrival at the office. The last questionnaires are processed interactively. There are four or more editing runs in batch every month. The records which have been up-dated are processed together with new questionnaires from later runs.

Within the production system there is also a macro-editing process. However, this process, the cell-list editing, is only applied when the time schedule permits it and as a final editing of the micro-edited data.

In this cell-list editing the estimate, the ratio and the difference to the estimate of the preceding month for every domain of study are printed out on a so called cell-list, one for every main variable. The lists are then scrutinized manually. If there is a "suspicious" estimate a search is made for "deviating" records to find out whether there are errors in those records.

4.2.3.3 Revision of the Micro-Editing Process

In our data editing project we have recently made revisions of both the processes within the original systems. Concerning the micro-editing process, we have reduced the number of flagged data by about 50 %, without any drop in quality.

For this revision we used our interactive editing programme GRUS, described in GRUS. We studied every check separately and how it interacted with any other check in the system.

We found that some checks were unnecessary in the sense that they did not discover any errors. It was noted that they were not redundant in the "Fellegi-Holt" meaning (see Fellegi & Holt (76)).

Another finding was that almost every check had too narrow bounds. They had been set according to the safety first principle. In this case only deviations of up to ±10 % from the values of the previous month were accepted.

By studying the impact of every single check and all combinations of single checks on SEW data, we constructed a new rather well-tuned system. The main difference to the old system was much wider acceptance limits in the checks. In fact they could have had still wider limits, but at that time the subject matter specialists were not willing to accept any greater changes.

The immediate result of using the new system was a slow raise in quality, as measured by the fact that more errors were found, due to twice the time to review every flagged data. We can now state that the time for the manual review work has been reduced to approximately 50 %. About the quality we can only state that it is at the same level as before.

4.2.3.4 The Aggregate Method

The idea behind the aggregate method is very simple. It is to use an error-detecting system (in our case EDIT-78 described in EDIT-78 (1982)) twice. Checks are run first to aggregates and then on the records of the file with the suspicious (flagged) aggregates.

The most important feature of the procedure is that the acceptance limits are based on the distributions of the check functions of the data to be edited. This is done by a manual analysis of print-outs of the tails of the distributions of the checks variables.

These print-outs can also be considered as an alternative realization of our editing method because identifiers are printed out as well. The reason why we use the EDIT-78 programme is that it contains an error message procedure which prints out in the same message every found inconsistency and suspicious data of the same record, which facilitates the manual review

procedure. Above all we use EDIT-78 because it contains an inter-active updating procedure.

4.2.3.5 The First Experiment

The procedure was applied to the whole body of the edited and published data of the SEW for a selected month.

The methods were applied to the following variables: Number of working hours, pay-roll, wages per hour, number of workers. The checks for each variable consisted of a combined ratio and difference check against the values of the preceding month, i.e. both the ratio and the difference had to be rejected to indicate the present value of the variable as suspicious.

This experiment aimed at finding out how methods and computer programmes for testing checking methods and procedures should be constructed. Besides, we also happened to detect two serious errors which had not been found in the processing of the data of that month. This was an indication that micro-editing may not always detect even serious errors.

4.2.3.6 Evaluations by Simulation Studies

A *Simulation on processed data*

The first simulation study was carried out on unedited data and the evaluation was done against the edited data for the selected month.

The editing process of the SEW surveys cannot consist only of this macro-editing method. For two reasons there is a need for an additional cleaning of the data, namely

i) to fight those validating and consistency errors (formal errors) which cannot be detected by the macro-editing method.

ii) to make it possible to carry out the macro-edits.

In our simulation study this additional cleaning operation consisted of a manual review of errors found by an error detection programme dedicated to find certain types of errors in certain variables. Missing values and errors in the variable "number of working days" were handled by this special programme, which proceeded the macro-editing procedure.

In our study the number of totally flagged records from this cleaning and preparatory programme was 161. The number of records in the survey was 2951. The flagged values were reconciled against the edited data file.

In order to test the macro-editing procedure under realistic conditions the records of the data file were divided into lots exactly as when that SEW survey was processed. However, the records from the last runs were put together into one editing round (instead of three).

All data were inflated by the inflation factors used in the normal estimating procedure. The records were then sorted and aggregated into four-digit SNI-groups. There are 89 such groups in the SEW.

For every editing round the procedure indicated above (see 4.2.3.4) was executed. The review work consisted in checking with the edited data. Flagged values from this macro-editing procedure were considered as errors found and revised, if the micro-editing had changed the original value. By matching the macro-edited file we found out those errors in the production process which had not been detected by our procedure.

RESULTS:
Number of records: 2951
Number of flagged records: 274
Number of errors found: 76

When the round was originally processed the number of flagged records was 435 and the number of errors found was 205. Thus, we got a reduction by 161 flagged records = 34 per cent.

But, did this simulated macro-editing process detect the most serious errors? This question may be answered by the following table, which for each variable on the four-digit level shows the impact of the remaining errors was for each variable, for all the 89 groups for which SEW data are published.

TABLE 4.2-1 shows the number of aggregates by the total relative difference of the estimates.

DIFFERENCE In per cent	WORKERS	HOURS	PAY-ROLL	WAGES/HOUR
0 < - < 0.05	4	1	6	6
0.1 - 0.4	4	5	5	8
0.5 - 0.9	1	2	4	2
2.5	1	0	0	0

B *Simulation during current processing*

This simulation study was carried out almost exactly as the one described above. The only difference was that this one was run parallel with the normal processing of the survey. Thus we eliminated any possibility of being influenced by knowing the results of the micro-editing process when we set the boundaries for the checks of our macro-editing procedure.

However, the same evaluation method had to be applied. Of course we should had preferred to evaluate the macro-editing method by processing the data entirely concurrently with the normal processing. This might be done in the future if the subject-matter specialists would be interested in the procedure.

RESULTS:
Number of records: 2996
Number of flagged records: 225
Number of errors found: 50

When the round was regularly processed, the number of flagged records was 389. The number of errors found was 110. Thus, we got a reduction by 164 flagged records = 42 per cent.

The impact of the remaining errors was calculated for each variable, for all the 89 groups for which SEW data are published. However, the macro-editing procedure seemed to be less successful than in the previous simulation in detecting the most serious errors found in the normal processing (see the figures within parenthesis in Table 4.2-2).

This simulation study was carefully analyzed. First, we investigated those errors with an impact on the aggregates of more than 2 % discovered by the

usual editing but not by our simulated process. There were two errors which were accepted by the ratio checks but not by the difference checks and which were at the border of the acceptance regions of the ratio checks. A slight change upwards of the lower boundary of the ratio check should have caused these errors to be discovered by our macro-editing method.

Then the impact of the boundaries of the checks on the efficiency of the error detection was analyzed. The finding was that the acceptance intervals of the checks should be wider. The strategy had been to set the limits as close to the main body of the data as possible. A better strategy seemed to be to set the limits as close as possible to the first outlier.

The findings of the analysis can be summed up as follows. If the simulation had been carried out with a considerably higher upper bound of the acceptance intervals of the ratio and the difference checks, then we should have got the following results:

The number of flagged data of the macro-editing checks had been reduced to 134, which means a 66 % reduction of flagged values compared with the corresponding micro-editing of these data.

The quality had not been affected as shown by Table 4.2-2.

TABLE 4.2-2 shows the number of aggregates by the total relative difference in per cent of the estimates. The figures within parentheses show the outcome of the first experiment of the same study.

DIFFERENCE	WORKERS	HOURS	PAY-ROLL	WAGES/HOUR
0 < - < 0.05	4 (4)	3 (2)	13 (12)	10 (8)
0.1 - 0.4	3 (3)	6 (6)	4 (4)	8 (8)
0.5 - 0.9	1 (1)	1 (1)	6 (5)	3 (3)
1.0 - 1.9	1 (1)	2 (1)		1 (1)
2.0 - 2.9				
3.0 - 3.9	(1)	(1)		
4.0 - 4.9			(1)	

4.2.3.7 Conclusions

On the basis of the two simulations we can state that the macro-editing procedure will lead to a substantial reduction of the manual review work.

There is no notable loss in quality. It is true that all errors detected by the micro-editing process are not detected by the macro-editing procedure. However, these errors are small and many of them are not found by the ratio checks of the micro-editing system either. They were found by scrutinizing flagged values of other items and checks.

It is quite clear that it is important to learn how to tune the macro-editing procedure and to fit it properly to the micro-editing procedure to get an efficient and well co-ordinated editing system. This can only be done by gaining experiences through using the system in the processing of the SEW.

We have learnt that the acceptance intervals of the ratio checks should not necessarily be symmetric and that they should be rather wide. (The safety first principle when defining acceptance regions do not apply to macro-editing methods either).

There are no timeliness problems with this kind of macro-editing methods. They can be applied during the processing of the data under the same conditions as computer assisted micro-editing methods.

The macro-editing concept is a realistic alternative or complement to micro-editing methods, which reduces the amount of manual reconciliation work to a considerable extent, without losses in quality or timeliness.

4.2.4 A Macro-Editing Application Developed in PC-SAS

4.2.4.1 Preface

An application of the Aggregate method developed on a PC using the SAS-system for the programming is described in accordance with Lindström (1990). A comparison is made between the actual production editing and the macro-editing. It is shown that using the macro-editing instead of the production editing the number of flagged records are reduced by 80 percent.

The description is very detailed. Thus it may also serve as a guide for implementing the Aggregate Method. It should be noted that the experiences of the studies lead to the Box-Plot Method as an alternative to the Aggregate Method.

4.2.4.2 The Edits

Checks are first made on aggregates and then on records belonging to a

suspicious (flagged) aggregate. The acceptance limits for the checks are based on the distribution function of the check variables.

The application was developed on the Survey of Employment and Wages (SEW, see 4.1.3.1). The edits used are based on computing the following variables at branch level. The weights for the previous period are used in computing these variables.

T1=100,0*WORK_HOUR(August)/WORK_HOUR(June)
S1=WORK_HOUR(August)-WORK_HOUR(June)

T3=100,0*SUM_WAGE(August)/SUM_WAGE(June)
S3=SUM_WAGE(August)-SUM_WAGE(June)

T4=100,0*HOURLY_WAGE(August)/HOURLY_WAGE(June)
S4=HOURLY_WAGE(August)-HOURLY_WAGE(June)

T5=100,0*EMPLOYED(August)/EMPLOYED(June)
S5=EMPLOYED(August)-EMPLOYED(June)

These ratios and differences at branch level are listed in ascending order. On the basis of these lists the acceptance limits for the checks at branch level are determined. The checks which are used at branch level are of the following type:

If (Tx<lower limit_Tx & Sx<lower limit_Sx) |
(Tx> upper limit_Tx & Sx>upper limit_Sx) then flag the branch.

4.2.4.3 The SAS-application

For the test a simple prototype was created using SAS on the PC. The prototype was built using the SAS/AF software. It contains a number of menus to be filled in by the user.

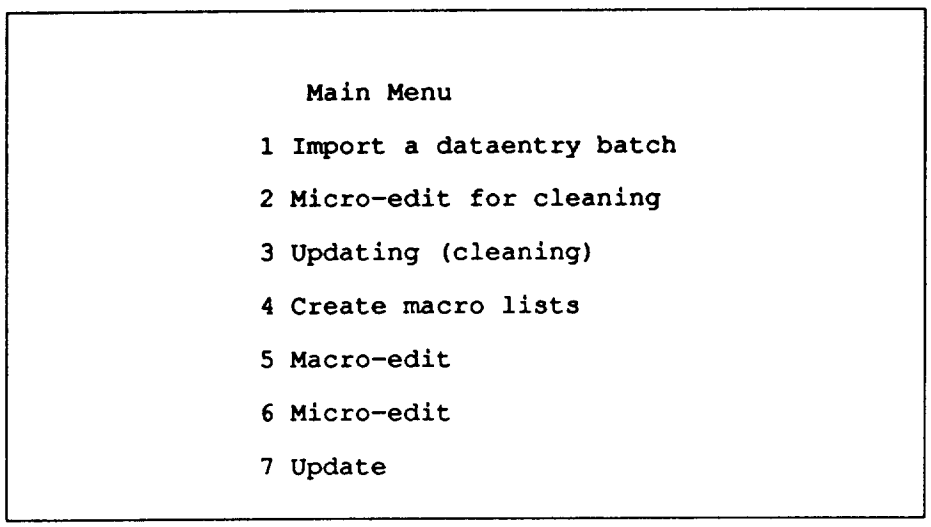
Figure 4.2-1 shows the main menu. The editing is supposed to be done for a number of batches. Each batch is composed by the number of forms which have been data-entered at the same time. This reflects the production system where the forms arriving up to a specified date are data-entered at the same time. To edit a batch the user passes through the different alternatives on the menu.

The second and third alternatives on the menu were not used in the

evaluation. The "cleaning" of the records was moved to alternative 6.

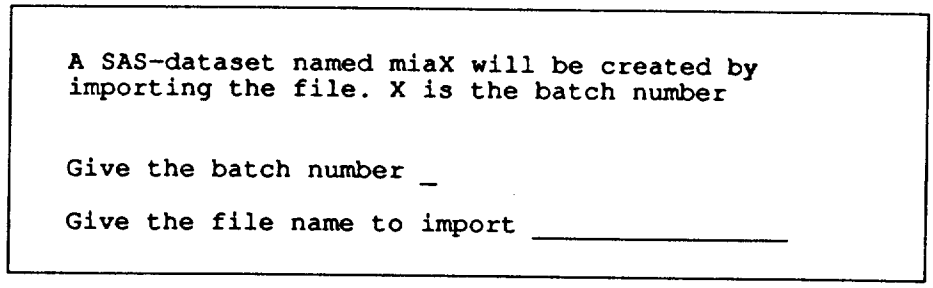
The purpose of the "cleaning" is to handle records with item nonresponse.

FIGURE 4.2-1:



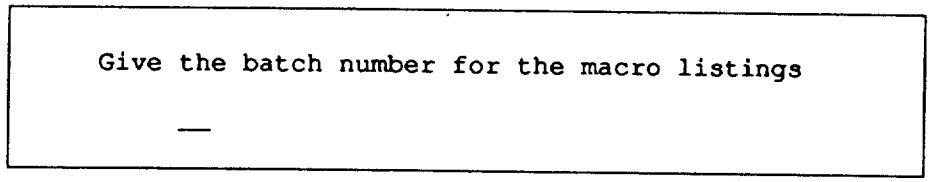
The first thing to do is to import the data entry file from a diskette into the SAS-system. At the same time two of the variables are imputed namely Number of operation days for the month and Number of operation days for the period when they are missing. This is done by choosing alternative 4 on the Main menu and then entering the batch number and the file name on the menu displayed in Figure 4.2-2. This will create a SAS-job which will import the data to a SAS-dataset and impute the variables Number of operation days for the period and Number of operation days for the month. As a result a list of the imputed records and a SAS-dataset will be produced.

FIGURE 4.2-2:



When the batch has been imported the next step is to create the macro lists which shall be used to determine the acceptance limits for the macro editing. This is done by choosing alternative 4 on the Main menu and then entering the batch number on the menu displayed in Figure 4.2-3. As a result a SAS-job is created that will produce the desired lists.

FIGURE 4.2-3:

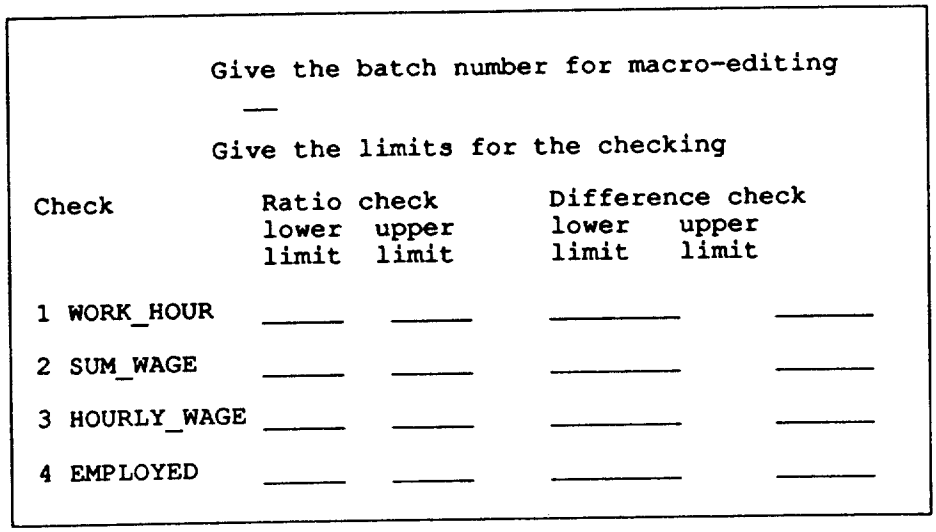


On the basis of the lists it is possible to determine the acceptance limits for the macro-editing. Choose alternative 5 on the Main menu and then enter the acceptance limits together with the batch number on the menu displayed, see Figure 4.2-4. This will create a SAS-job that will produce a list of the flagged branches and a SAS-dataset.

The SAS dataset will contain all the variables from the input and new variables indicating whether the record belongs to a flagged branch and in that case which check has caused the flagging.

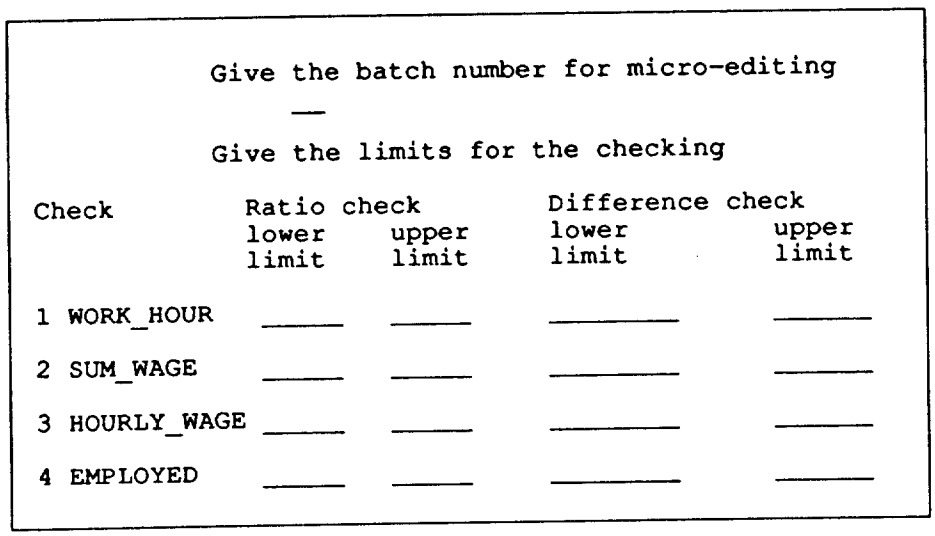
Lists on the establishment level are also produced here. The lists are the same as on branch level but the differences and the ratios are calculated on establishment level. Only establishments belonging to a flagged branch for that variable appear on the list. This is a change compared with the previous study made on the same survey. In the earlier study all establishments belonging to a flagged branch were listed for all variables (see Granquist (1988)).

FIGURE 4.2-4:



These lists are used for determining the acceptance limits on the micro level. When the limits have been determined you choose alternative 6 on the main menu and the menu in Figure 4.2-5 will be displayed. Here you enter the number of the batch together with the acceptance limits for the checks on micro level. As a result a SAS-job will be created that produces a new SAS-dataset containing all previously existing variables together with flags from the micro editing. The job will also produce a list of all flagged records.

FIGURE 4.2-5:



On the basis of the error list a decision about which records to correct can

be made. Then you choose alternative 7 on the main menu. This will result in a menu, in which you enter the batch number, and thereafter the flagged records will be displayed one by one on the screen. Here it is possible to correct the record before continuing with the next flagged record.

4.2.4.4 The Results

A simulation of the editing was done on the survey for August 1989. The data was checked against the June 1989 survey.

The data for macro-editing was divided into two batches the first one with 1090 records and the second one with 1961 records.

Before the data was edited the variables Number of operation days for the period and Number of operation days for the month were imputed. When those variables had no value they received a value based on the value for the previous month. In the first batch 25 records were imputed and in the second batch 26 records were imputed.

After the imputation the macro lists were created. These were used for determining the acceptance limits for the editing rules at macro level.

Eight different lists were produced. The lists contain the difference between the present and the previous survey at branch level, and the ratio between the present and previous survey. These lists were produced for the variables Number of worked hours, Sum of wages, Hourly wages and Number of employed. This gives two lists for every variable one sorted in ascending order by the value of the difference and one sorted in ascending order by the value of the ratio.

On the basis of the lists the limits for the acceptance interval on macro level were determined. The limits are presented in Table 4.2-1. The results of using these limits are presented in Table 4.2-2.

TABLE 4.2-1 shows the acceptance limits for the macro-editing on branch level

VARIABLES/LIMITS	BATCH 1		BATCH 2	
	RATIO	DIFFERENCE	RATIO	DIFFERENCE
NUMBER OF WORKED HOURS				
LOWER LIMIT	87 %	-4000	84 %	-14 000
UPPER LIMIT	120 %	4000	140 %	25 000
SUM OF WAGES				
LOWER LIMIT	87 %	-300 000	89 %	-300 000
UPPER LIMIT	113 %	150 000	89 %	1 000 000
HOURLY WAGES				
LOWER LIMIT	95 %	-3.0	95 %	-4.0
UPPER LIMIT	105 %	3.0	110 %	6.0
NUMBER OF EMPLOYED				
LOWER LIMIT	95 %	-40	90 %	-100
UPPER LIMIT	105 %	40	105 %	100

The complete material contained 88 different branches. Not all branches were represented in both batches as can be seen in Table 4.2-2.

TABLE 4.2-2 shows the number of branches and flagged branches per checking rule and batch at the macro-editing on branch level.

VARIABLE	BATCH 1		BATCH 2	
	NUMBER OF BRANCHES	NUMBER OF FLAGGED BRANCHES	NUMBER OF BRANCHES	NUMBER OF FLAGGED BRANCHES
Number of worked hours	78	13	86	13
Sum of wages	78	15	86	13
Hourly wages	78	19	86	7
Number of employed	78	17	86	9
All checking rules	78	40	86	24

For the flagged branches, lists were produced of the same type as earlier but now on establishment level instead of branch level. The lists contain all establishments belonging to a flagged branch for the variable.

The numbers of establishments belonging to a flagged branch is presented in Table 4.2-3.

TABLE 4.2-3 shows the numbers of establishments belonging to a flagged branch per checking rule.

VARIABLE	BATCH 1	BATCH 2
Number of worked hours	175	302
Sum of wages	89	308
Hourly wages	249	194
Number of employed	204	159

On the basis of these lists the acceptance limits for the editing on establishment level were determined. The limits are presented in Table 4.2.4.

TABLE 4.2-4 shows the acceptance limits for the editing on establishment level.

VARIABLES/LIMITS	BATCH 1		BATCH 2	
	RATIO	DIFFERENCE	RATIO	DIFFERENCE
NUMBER OF WORKED HOURS				
LOWER LIMIT	85 %	-10 000	70 %	-14 000
UPPER LIMIT	150 %	10 000	130 %	8 000
SUM OF WAGES				
LOWER LIMIT	75 %	-500 000	76 %	-700 000
UPPER LIMIT	140 %	400 000	150 %	700 000
HOURLY WAGES				
LOWER LIMIT	85 %	-10.0	80 %	-10.0
UPPER LIMIT	115 %	10.0	115 %	10.0
NUMBER OF EMPLOYED				
LOWER LIMIT	85 %	-40	83 %	-40
UPPER LIMIT	115 %	50	110 %	20

These rules were applied to the material together with a number of validity checks. The result of this editing is presented in Table 4.2-5 below. It should be noted in the table that an establishment may have been flagged by more than one checking rule. This means that the total number of flagged

establishments is less than the sum of the different checking rules.

4.2.4.5 Comparing the Macro-editing and the Production Editing

The flagged establishments in the macro-editing were imputed with the edited values from the production. When the record had been corrected in the production the same corrections were applied to the macro-edited record.

TABLE 4.2-5 shows the number of flagged and corrected establishments after editing on establishment level.

CHECKS	BATCH 1		BATCH 2	
	FLAGGED	CORRECTED	FLAGGED	CORRECTED
# Worked hours	13	15	23	31
Sum of wages	20		22	
Hourly Wages	16		21	
Number of employed	18		26	
Validity checks	22	21	53	50
All	80	36	118	81

After the corrections had been applied estimates were made for the items on branch level.

TABLE 4.2-6 shows the number of aggregates by the total relative difference in per cent of the estimates. The figures within parentheses show the outcome of an earlier study.

DIFFERENCE	WORKERS	HOURS	PAY-ROLL	WAGES/HOUR
0	72 (78)	45 (77)	38 (66)	36 (68)
0.0 - O.1	1 (4)	6 (2)	6 (12)	15 (8)
0.1 - 0.4	10 (3)	10 (6)	14 (4)	18 (8)
0.5 - 0.9	3 (1)	6 (1)	6 (5)	6 (3)
1.0 - 1.9	0 (1)	9 (1)	15	7 (1)
2.0 - 2.9	1	3	2	3
3.0 - 3.9	1 (1)	2 (1)	1	1
4.0 - 4.9	0	2	2 (1)	1
> 4.9	0	5	4	1

When comparing the results from this study with the results from the previous study it can be seen that the deviations are larger in this study.

A closer look at the branches with a deviation greater than 5% shows that there are 5 branches which cause those deviations. For those branches it should be sufficient to correct one record per branch. Then the deviations should be less than 5%.

The reason why these records were not flagged at the macro-editing is that the difference and ratio for the branch is inside the acceptance interval. This means that the branch will not be flagged and the establishments for that branch will not at all be checked in the micro-editing.

4.2.4.6 Further Editing

To try to find the records behind the big deviations the data was micro-edited once more. At this micro-editing ratio checks were applied to all records which did not belong to a flagged branch. The following ratios

where used in the checks:

T1=100,0*WORK_HOUR(August)/WORK_HOUR(June)

T3=100,0*SUM_WAGE(August)/SUM_WAGE(June)

T4=100,0*HOURLY_WAGE(August)/HOURLY_WAGE(June)

T5=100,0*EMPLOYED(August)/EMPLOYED(June)

The ratio checks were of the following type:

If (Tx<lower limit_Tx OR Tx> upper limit_Tx) then
flag the establishment

Tests where made with different limits for the ratio checks. The results of these tests are presented in Table 4.2-7.

TABLE 4.2-7 shows the number of flagged and thereof corrected establishments at different limits for the ratio checks.

Upper limit	Lower limit	Number of flagged establishments			Number of corrected establishments		
		Batch 1	Batch 2	Total	Batch 1	Batch 2	Total
500	20	0	15	15	0	12	12
400	25	3	19	22	2	15	17
300	33	9	25	36	4	16	20
200	50	45	68	113	17	29	46

The changes of the estimates which these corrections led to are presented in Table 4.2-8; 4.2-9; 4.2-10 ; 4.2-11.

TABLE 4.2-8 shows the number of deviations between data edited in production and in the experiment as a function of the absolute percentage deviation for the estimates on branch level after further micro-editing with ratio < 20 or ratio > 500 for not flagged branches.

ABSOLUTE PERCENTAGE DEVIATION	WORKERS	HOURS	PAY-ROLL	HOURLY WAGES	SUM
0	72	47	40	39	198
0.0 - O.1	1	5	6	16	28
0.1 - 0.4	10	12	16	20	58
0.5 - 0.9	3	9	8	5	25
1.0 - 1.9	0	10	14	6	30
2.0 - 2.9	1	3	2	1	7
3.0 - 3.9	1	1	0	0	2
4.0 - 4.9	0	0	1	1	2
> 4.9	0	1	1	0	2

TABLE 4.2-9 shows the number of deviations between data edited in production and in the experiment as a function of the absolute percentage deviation for the estimates on branch level after further micro-editing with ratio < 25 or ratio > 400 for not flagged branches.

ABSOLUTE PERCENTAGE DEVIATION	WORKERS	HOURS	PAY-ROLL	HOURLY WAGES	SUM
0	72	48	41	40	201
0.0 - O.1	1	6	6	18	31
0.1 - 0.4	10	12	16	18	56
0.5 - 0.9	3	8	8	4	23
1.0 - 1.9	0	10	14	6	30
2.0 - 2.9	1	3	2	1	7
3.0 - 3.9	1	1	0	0	2
4.0 - 4.9	0	0	1	1	2
> 4.9	0	0	0	0	0

TABLE 4.2-10 shows the number of deviations between data edited in production and in the experiment as a function of the absolute percentage deviation for the estimates on branch level after further micro-editing with ratio < 33 or ratio > 300 for not flagged branches.

ABSOLUTE PERCENTAGE DEVIATION	WORKERS	HOURS	PAY-ROLL	HOURLY WAGES	SUM
0	72	48	41	40	201
0.0 - O.1	1	7	6	16	30
0.1 - 0.4	10	12	17	20	59
0.5 - 0.9	3	8	8	4	23
1.0 - 1.9	0	8	12	6	26
2.0 - 2.9	1	4	3	1	9
3.0 - 3.9	1	1	0	0	2
4.0 - 4.9	0	0	1	1	2
> 4.9	0	0	0	0	0

TABLE 4.2-11 shows the number of deviations between data edited in production and in the experiment as a function of the absolute percentage deviation for the estimates on branch level after further micro-editing with ratio < 50 or ratio > 200 for not flagged branches.

ABSOLUTE PERCENTAGE DEVIATION	WORKERS	HOURS	PAY-ROLL	HOURLY WAGES	SUM
0	73	51	43	42	209
0.0 - O.1	1	7	11	16	35
0.1 - 0.4	9	14	17	20	60
0.5 - 0.9	3	7	7	4	21
1.0 - 1.9	0	5	8	5	18
2.0 - 2.9	1	3	2	1	7
3.0 - 3.9	1	1	0	0	2
4.0 - 4.9	0	0	0	0	0
> 4.9	0	0	0	0	0

As can be seen from the tables records which caused the largest deviations are flagged when the acceptance limits of the ratio checks are less than 25

and greater than 400. A continuation of the checking with smaller acceptance interval lessens the deviations but at the same time the number of flagged records will increase.

The selected strategy combining the macro-editing with ratio checks with very wide acceptance intervals for records belonging to a non-flagged branch will give the flags and corrections presented in Table 4.2-12.

TABLE 4.2-12 shows the number of flagged and corrected establishment after batch and type of flag. For the macro-edited data with ratio check for ratio less than 25 or ratio greater than 400.

CHECKS	FLAGGED ESTABLISHMENTS			CORRECTED ESTABLISHMENTS		
	BATCH 1	BATCH 2	TOTAL	BATCH 1	BATCH 2	TOTAL
Number of worked hours	13	23	36	15	31	46
Sum of wages	20	22	42			
Hourly wages	16	21	37			
Number of employed	18	26	44			
Validity checks	22	53	75	21	50	71
Ratio checks	3	19	22	2	15	17
Total	83	137	220	38	96	134

A comparison between the results in the production and in the experiment is presented in Table 4.2-13.

TABLE 4.2-13 shows the number of edited, flagged and corrected establishment at the production and at the experiment.

	EDITED ESTABLISHMENTS	FLAGGED ESTABLISHMENTS		CORRECTED ESTABLISHMENTS	
		NUMBER	PERCENT	NUMBER	PERCENT
Production	3051	1087	36	306	28
Experiment	3051	220	7	134	61

The macro-editing used caused a reduction of the number of flagged establishments with 80 percent. 61 percent of the flagged establishments were corrected compared with 28 percent at the production.

This reduction of the number of flags is probably slightly greater than if the method were to be used in ordinary production because then the editing would be divided into more batches and this would lead to an increase in the number of flags from the macro-editing.

4.2.4.7 Summary and Conclusions

A prototype, developed in PC-SAS, for editing the Survey of Employment and Wages (SEW) with the Aggregate Method is described and the studies carried out on the SEW data from August, 1989 are reported.

The realization of the Aggregate Method was somewhat modified as compared to the original method (see 4.2.3). Instead of forming an error file from the aggregate check consisting of all questionnaires of all aggregates which had failed at least one edit, the records of a flagged aggregate were given a specific signal, telling which of the four variables that did not pass the edit. The check on the micro level was then applied only to those questionnaires which belonged to the aggregates which had failed the check for that variable. This version of the method may imply a small reduction in the number of flagged data, as fewer records are checked on the micro level than in the realization described previously.

The result of this study was a reduction in the number of flagged data by nearly 80 per cent.

However, the loss in quality was slightly higher than in the studies carried out by the main frame prototype, due to the modification, the unusually large number of questionnaires in the second processing round and/or to the wider acceptance interval. It was found that this loss in quality was caused by a few large errors, which did not cause the aggregates to be flagged.

How to overcome the detected weakness of the method was the subject for further studies. An additional editing for records belonging to non-flagged branches was developed (see Tables 4.2-6 - 4.2-11). The edit was a simple ratio check with very wide acceptance bounds.

The strategy of combining the macro-editing with a ratio micro-edit check for records of the non-flagged aggregates was very successful, which is shown by the tables 4.2-12 and 4.2-13.

This finding led to questioning the aggregate checks method as such. The only advantage of the method is that they can save storage or computer time.

However, when there are no problems with either the storage capacity or the computer cost, the aggregate checks can be skipped. The method is still a macro-editing method but the term "Aggregate Method" may not be adequate. When the tails of the distribution of the check function is provided with Box-Plots (which is recommended), this method is called the Box-Plot Method.

4.2.5 The Aggregate Method for Implementation in EDP systems

The method can be used for repetitive surveys where the same population sample is used repeatedly or on surveys where it is possible to have checks based on computing ratios or differences between variables collected at the same time. The method for repetitive surveys is based on the fact that it is possible to compare the answers between the previous and the current survey period to detect suspicious records. The description is based on the method used and presented in Lindström (1990).

The first operation is joining the file from the previous period with the file from the current period.

The joined file is thereafter used to create an aggregated file. The aggregation is performed on the survey values multiplied by the weight (according to the sample design).

For the aggregated file the differences and the ratios between the current and the previous value are calculated.

These ratios and differences are then listed in ascending order. With one separate list for every ratio and difference.

The ratios and differences presented on these lists are then manually inspected to determine the acceptance limits in the checks made on the aggregated level.

The acceptance limits are then entered into checks of the following type:

```
IF      (ratio > upper_limit or ratio < lower_limit or
         difference > upper_limit or difference < lower_limit
THEN    signal
```

The result of this checking will be a file with one record per domain containing Boolean variables that indicate if the domain is suspicious.

This file is joined with the original file.

After the join the ratios and differences are calculated for records belonging to domains that have been marked as suspicious.

These ratios and differences are then listed in ascending order. With one separate list for every ratio and difference.

The ratios and differences presented in these lists are then manually inspected to determine the acceptance limits in the checks made on the micro level.

The acceptance limits are then entered into checks of the following type:

```
IF      (ratio > upper_limit OR ratio < lower_limit OR
         difference > upper_limit OR difference < lower_limit)
    AND  domain_suspicious
THEN     signal
```

The records which have been signaled are then listed. The listed records are then manually reviewed to find records which should be updated. The update procedure of the system is then applied i found necessary.

4.3 STATISTICAL EDITS (THE HIDIROGLOU-BERTHELOT METHOD)

4.3.1 Introduction

4.3.1.1 References

The Hidiroglou-Berthelot Method (the HB-Method) is described as a micro-editing method in Hidiroglou-Berthelot (1986). The method is a ratio check inspired by Tukey's Explorative Data Analysis (EDA) methods (see Tukey (1977)). In the original paper by Hidiroglou and Berthelot it is considered as a solution to some problems connected with the traditional ratio-check method.

At Statistics Canada it is known as "Statistical Edits", and has been adapted to several surveys, e.g. the Delivery, Stock and Order Survey, (Lalande (1988 a,b)), the Current Shipment, Inventories and Orders Survey (Tambay (1986)), and the Wholesale-Retail Survey (Berthelot (1983)).

As a macro-editing method it is reported in Granquist (1991) and Höglund (1989).

4.3.1.2 Contents

First there is a short description of the method, including a few findings from various studies of the method. The description is essentially drawn from the corresponding chapter in Granquist (1991). It is intended to serve as an introduction and short overview of the most important features of the method.

Then follows an edited version of Höglund (1989) which is a rather explanatory and detailed description of the method and of an evaluation on data of the Survey of the Delivery and Orderbook Situation (DOS). It may serve as a guide for marketing and implementing the method in statistical agencies.

Finally there is a description of the method as an EDP function (from Lindström (1990 b)).

4.3.2 Overview of the Hidiroglou-Berthelot Macro-Editing Method

4.3.2.1 Explanatory Description

The Hidiroglou-Berthelot Method (HB-Method) is a ratio method, for which the bounds are automatically calculated from the data to be edited. The method uses the robust parameters median, quartiles (Qi) and interquartile ranges (DrQi) instead of the mean and standard deviation to prevent the bounds from being influenced by single outliers. Then the lower (l) and the upper (u) bounds should be:

$$l = R_{MEDIAN} - k * D_{rQ1}$$

$$u = R_{MEDIAN} - k * D_{rQ3}$$

However, such a straightforward application of the ratio method has two drawbacks,

i) the outliers on the left tail may be difficult to detect
ii) the method does not take into account that the variability of ratios for small businesses is larger than the variability for large businesses

The HB-Method solves these drawbacks by a symmetric transformation followed by a size transformation.

The symmetric transformation

$$S_i = \begin{cases} 1 - R_{MEDIAN} / R_i \ , & 0 < R_i < R_{MEDIAN} \\ R_i / R_{MEDIAN} - 1 \ , & R_i \geq R_{MEDIAN} \end{cases}$$

The size transformation

$$E_i = S_i * (MAX (X_i (t) \ , \ X_i (t+1)))^{U}$$
$$0 \leq U \leq 1$$

E_{Q1} , E_{Q3} are the first and third quartiles of the transformation E

$$D_{Q1} = MAX (E_{MEDIAN} - E_{Q1} , | A * E_{MEDIAN} |)$$

$$D_{Q3} = MAX (E_{Q3} - E_{MEDIAN} , | A * E_{MEDIAN} |)$$

which gives the lower and upper limits of the checks:

$$l = E_{MEDIAN} - C * D_{Q1}$$

$$u = E_{MEDIAN} + C * D_{Q3}$$

A is considered a constant, equal to 0.05, which means that there are only two parameters, U and C, which have to be set in advance to get the method to run. The real reason behind the symmetric transformation is to get rid of one parameter. We have found that the parameters are not very sensitive. The same values can be used for many variables of a survey.

Figure 4.3-1 below may explain the transformations and the method. The figure has been produced by the prototype of the Box Method for the editing of the Survey of Employment and Wages (SEW). The acceptance limits have been calculated for a few values of the parameters U and C. This method can be applied when implementing the HB-Method forming an operation in the production process. That way the acceptance bounds would be completely determined by the data to be edited.

It should be noted that to make the method a macro-editing method we only had to inflate the keyed-in values of the variable X, in principle by the inverted sample fraction.

4.3.2.2 Findings

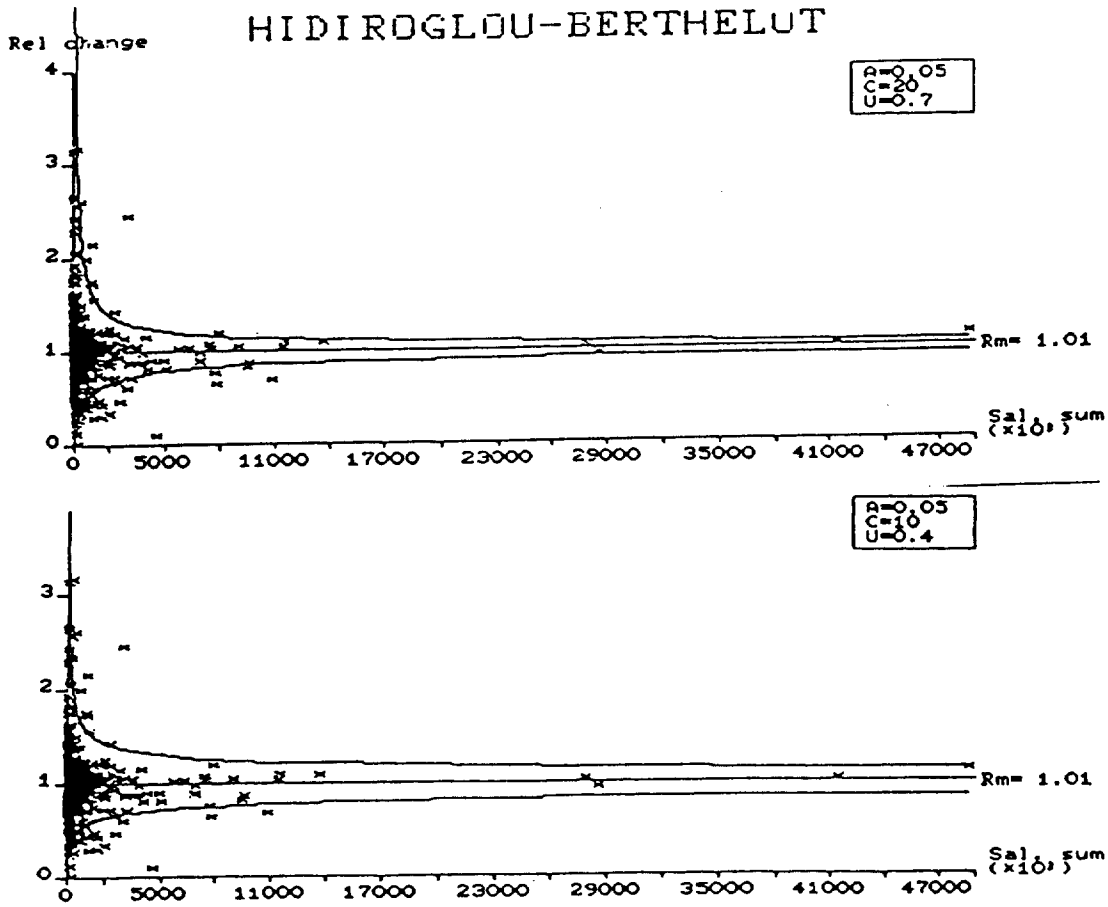
This is an excellent method. It can well compete with the Top-Down Method. In a comparison with the Aggregate Method on data from the Survey of Employment and Wages, the HB-Method was found to be superior. It has also been evaluated in connection with experiments on Consumer Price Index Survey (CPI) data, (see Wickberg (1991)) with overwhelming results. The hit rate of the edit was very high and all important errors were found.

The parameters U and C are not very sensitive. The same values can be used

for many variables of a survey which indeed makes the method easy to use.

The protection against (the A^*E_{median} term) detecting too many non-outliers when the E_is are clustered around a single value with only a few deviations does not work when the values are clustered around **1**, which was a finding from the experiments on CPI data.

FIGURE 4.3-1 shows the acceptance limits of the H-B-Method on SEW data, calculated for a few values of the parameters U and C.



4.3.3 The Evaluation of the HB-Method on DOS Data

4.3.3.1 Edits in Periodic Surveys

Numerous methods have been proposed for detecting divergent observations in large periodical surveys. Some suggest that the problem should be treated as a hypothesis-testing problem, either with or without the assumption of a certain distribution for the data. Other methods use ratios of current period data to previous period data and set upper and lower bounds according to some rule which may depend on the distribution of the ratios or the data. Observations outside the bounds are considered to be divergent.

Usually the methods try to make use of the distribution of the ratios in the construction of bounds around the mean or the median value either using the standard deviation or the quartiles. Most of the suggested methods have some drawbacks, mainly because the variables in business-surveys usually exhibit very skewed distributions.

4.3.3.2 The Hidiroglou-Berthelot Edit

The Hidiroglou-Berthelot edit is a heuristic method that has been developed using parts of several other methods. By using information provided by the data itself the intention is to identify all big changes regardless of whether it is an increase or a decrease.

Given data for a variable from two consecutive periods,

$(x_i(t) , x_i(t+1))$ $i = 1,2,...,n$,

the individual relative change for each element is defined as

$R_i = x_i(t+1) / x_i(t)$.

Hidiroglou-Berthelot (1986) claim that to be able to find and treat both increases and decreases in the same way R_i has to be transformed in the following manner:

$$S_i = \begin{cases} 1 - R_{median} / R_i , & 0 < R_i < R_{median} \\ R_i / R_{median} - 1 , & R_i \geq R_{median} \end{cases}$$

where R_{median} is the median of the R_i ratios.

Half the number of S_i's are less than zero and the other half greater than zero. However according to Hidiroglou- Berthelot the transformation ensures an equally good detection of divergent observations in both tails of the distribution. The transformation does not provide a symmetric distribution of the observations. This is perhaps more obvious if S_i is rewritten as:

$$S_i = \begin{cases} (R_i - R_{median}) / R_i , & 0 < R_i < R_{median} \\ (R_i - R_{median}) / R_{median} , & R_i \geq R_{median} \end{cases}$$

To make use of the magnitude of the observations a second transformation is performed:

$$E_i = S_i * (MAX(x_i(t), x_i(t+1)))^U.$$

U is by Hidiroglou-Berthelot proposed to be a value between 0 and 1. The E transformation is a rescaling of the S's that keeps the order and sign of the elements. It makes it possible to put more importance on a relatively small change in a "large" element than on a relatively small change in a "small" element. The choice of the U-value governs the importance associated with the magnitude of the data. U=1 gives full importance to the size of the element x_i and U=0 gives no importance at all to its size. In the latter case $E_i=S_i$. The effects of different choices of U is illustrated in Example 4.3-1.

EXAMPLE 4.3-1. R_{median} = 1.25 and A= 0.05

$X_i(t)$	$X_i(t+1)$	R_i	S_i	$E_i(U=0.1)$	$E_i(U=0.4)$	$E_i(U=0.9)$
1	5	5	3	3.52	5.71	12.77
10	5	0.5	-1.5	-1.89	-3.77	-11.91
100	5	0.05	-24	-38.4	-151.43	-1514.30
1000	5	0.005	-249	-496.82	-3946.38	-124795.62
10000	5	0.0005	-2499	-6277.20	-99486.98	-9948698.19
100000	5	0.00005	24999	-79053.78	-2499900.00	-790537792.47
5	1	0.2	-5.25	-6.17	-9.99	-22.35
5	10	2	0.6	0.76	1.51	4.77
5	100	20	15	23.77	94.64	946.44
5	1000	200	159	317.25	2519.98	79788.77
5	10000	2000	1599	4016.51	63657.34	6365733.66
5	100000	20000	15999	50593.28	1599900.00	505932802.98

The example shows that negative E_i's (and S_i's) indicate a decrease and positive E_i's (and S_i's) an increase.

The greater U becomes, the greater the dispersion of E will be.

The E_i's are distributed around zero and those E_i's that are too small/big are considered as possible outliers.

In this context the definition of an outlier according to Hidiroglou-Berthelot is an observation "whose trend for the current period to a previous period, for a given variable of the element vector **x(t)**, differs significantly from the corresponding overall trend of other observations belonging to the same subset of the population".

Hidiroglou-Berthelot construct upper and lower limits for an interval around E by means of the following quantities:

$$D_{Q1}=MAX\{ E_{median} - E_{Q1}, \mid A^* E_{median} \mid \}$$

$$D_{Q3}=MAX\{ E_{Q3} - E_{median}' \mid A^* E_{median} \mid \}$$

A is an arbitrary value suggested by Hidiroglou-Berthelot to be 0.05 .

E_{median}, E_{Q1}, E_{Q3} are the median and the first and third quartiles of the transformation E.

The $A*E_{median}$ term is a protection against detecting too many non-outlier when the E_i's are clustered around a single value with only a few deviations. That is, when $E_{median} - E_{Q1}$ or $E_{Q3} - E_{median}$ are less than $A*E_{median}$.

In Example 4.3-1 above, with U = 0.4 and A = 0.05, E_{median} = -1.13 which means that if E_{Q1} is in the interval [-1.1865,-1.13] then $E_{median} - E_{Q1} \leq A* E_{median}$= 0.0565 and thus D_{Q1}=0.0565.

All values outside the interval

$$\{ E_{median} - C * D_{Q1} ; E_{median} + C * D_{Q3} \}$$

where C is constant, are treated as outliers. Increasing the value of C gives a wider interval and a lower number of outliers.

The point in using the quartiles instead of standard-deviations is to avoid too much influence from the outlier.

EXAMPLE 4.3-2

Using the values of Example 4.3-1 again, with U = 0.4 and A = 0.05 gives

E_{median} = (-3.77+1.51)/2 = -1.13

E_{Q1} = -3946.38

E_{Q3} = 2519.98

D_{Q1} = -1.13 - (-3946.38) = 3945.25

D_{Q3} = 2519.98 - (-1.13) = 2521.11

C=10 gives the interval

(-1.13 - 10 * 3945.25 ; -1.13 + 10 *2521.11) = (-39453.63 ; 25209.97).

In Example 4.3-1 two observations in each tail would be classified as a

possible outlier.

If C=30 the interval is (-118358.63 ; 75632.17) and only one observation in each tail would be classified as a possible outlier.

4.3.3.3 Application of the Hidiroglou-Berthelot Edit

In the application of the H-B Method, three parameters, **A, U** and **C**, had to be estimated. Hidiroglou-Berthelot do not give any indications about the choice of the parameters other than that **A** should be very small and $0 \leq U \leq 1$.

A large number of combinations of parameter values had to be tried in order to see what happened to the material. The material used for the study contained both the raw-data set and the final data set edited according to the currently used method. Each element whose value had been changed was considered to be an outlier. Note that the meaning of the word outlier here is not exactly the same as in the definition given in Hidiroglou-Berthelot (1986).

The data could thus be separated in two populations, one population of outliers and one population of non-outliers. This was exploited when estimating the parameters. In fact, the problem was treated as a classification problem (Anderson (1958)). The attention was focused on finding some combination of the parameters, **A, U** and **C**, that would minimize the probability of misclassification of an element. For that purpose the following points were defined.

(i) As many as possible of the outlier should be correctly classified.

(ii) As few as possible of the non-outlier misclassified.

These two points contradict each other and a loss function would have been useful. Unfortunately this was not possible since no information about the consequences of misclassification was available.

A third aspect of interest was that

(iii) the identified outlier ought to have both a large impact on the estimates and a monthly change that clearly diverge from previous month.

The method of this paper consider (i) and (ii) in the first stage and in a

second stage point (iii).

A grid of different combinations of the parameters **U** and **C** was run through. The parameter **A** was fixed at 0.05 as suggested by Hidiroglou-Berthelot.

For each combination the edit identified a number of observations as possible outliers which were flagged. The frequency of correctly identified outliers:

(# outliers correctly identified by the H-B Method)/(total # outliers)

gave a measure that was used when evaluating the method.This measure can be regarded as the empirical probability of

"good" classification = 1 - Pr(misclassification).

After all the combinations of **U** and **C** had been tried, one was chosen for the examination of the effect on the other variables (see Table 4.3-2).

4.3.3.4 Data and Results of the Evaluation Study

A *Experimental Data*

The data used for estimating the parameters of the Hidiroglou-Berthelot edit came from the monthly Delivery and Orderbook Survey mentioned earlier.

There were six variables of interest included in the raw data file : Domestic Delivery, Export Delivery, Domestic Order, Export Order, Domestic Stock and Export Stock. The variable chosen for the adaption of the parameters, Domestic Delivery, was the one having most non-zero observations left after excluding missing values since such observations have to be taken care of apart.

The data used in the H-B-Method were the original values reported in the month of January-88 and the final values of December-87 where the latter had been checked with the current edit (the Top-Down procedure, see 4.4.3.3). Comparison between the January-88 data edited with the currently used method and the corresponding raw data of the same month identified the population of outliers.

For the variable "Domestic Delivery", the outlier population consisted of the

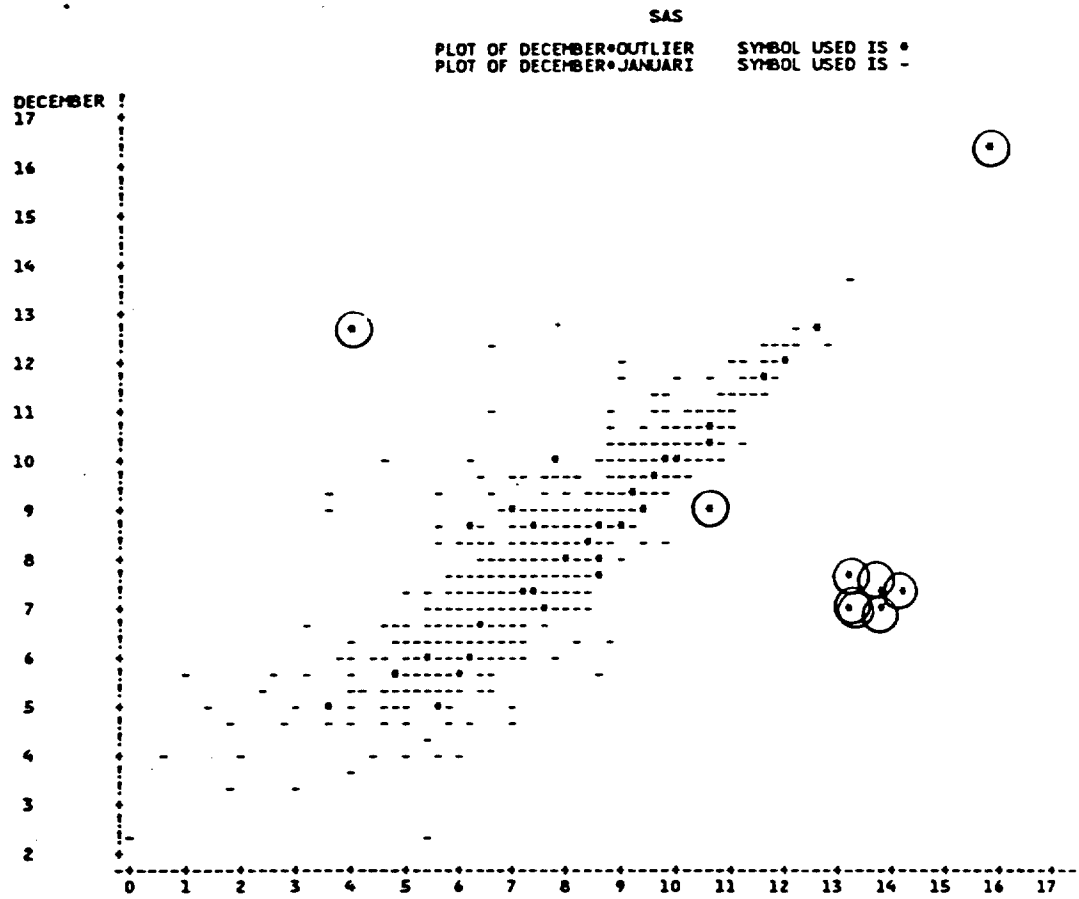
38 observations in Table 4.3-1 and the non-outlier population of 1511 observations (see Figure 4.3-2).

TABLE 4.3-1 shows the outlier population of the variable "domestic deliveries" (SEK 1000's)

OBS	DECEMBER	JANUARY		
		OLD	NEW	CHANGE
1	14648285	7403131	7403	-7395728
2	1560	1605000	1605	-1603395
3	1032	973693	974	-972719
4	1500	925000	925	-924075
5	1302	902805	903	-901902
6	962	593636	594	-592942
7	1926	501675	502	-501173
8	8473	38010	25328	-12682
9	110893	107891	101520	-12682
10	31745	42201	37201	-5000
11	5000	8300	5200	-3100
12	21930	19013	16730	-2283
13	3465	3084	1007	-2077
14	2879	5495	3650	-1845
15	22839	22400	20977	-1423
16	7344	11758	10840	-918
17	1651	1601	1076	-525
18	4072	4061	3691	-370
19	127	262	114	-148
20	8839	1200	1130	-70
21	434	480	462	-18
22	831	664	649	-15
23	11990	9790	9780	-10
24	4834	4404	4398	-6
25	301	116	111	-5
26	1882	4995	4994	-1
27	6245	5658	5659	1
28	274190	271455	271456	1
29	47500	38509	38539	30
30	148	35	80	45
31	1046	1860	1936	76
32	14551	14000	14169	169
33	5383	463	1032	569
34	20400	2500	3600	1100
35	173860	165000	167261	2261
36	6180	1702	4131	2429
37	6000	6000	14300	8300
38	294600	54	54000	53946

The sum of all changes: 12 997 728 (SEK 1000's)

FIGURE 4.3-2 shows a plot of the logarithmed values of the variable Domestic Delivery for the two successive periods. The outlier are marked with '*' and the outlier found by the H-B edit (U=0.4, C=41) are marked ' * '.



B *Estimation of the Parameters*

According to the method described in 4.3.3.3 several combinations of different values of the parameters A, U and C were tried. Changing A, for some fixed value of U and C, had no impact at all on the probability of misclassification. In this particular case

$E_{median} - E_{Q1} > | A * E_{median} |$ and

$E_{Q3} - E_{median} > | A * E_{median} |$ for small values of A.

Therefore A was set to 0.05, as suggested by Hidiroglou-Berthelot, which

simplified the further computations. The remaining parameters U and C were systematically varied to each other. The results are shown in Table 4.3-2.

TABLE 4.3-2 shows the *number of correctly identified outliers*, and separated by "/" the *total number of observations classified as outliers by the H-B Method* for different values of the parameters U and C.

C	U=.0	U=.1	U=.2	U=.3	U=.4	U=.5	U=.6	U=.7	U=.8	U=.9
47	7/21	7/20	7/20	7/22	7/23	9/30	9/32	10/37	10/43	10/51
45	7/21	7/20	7/20	7/22	7/24	9/31	9/33	10/38	10/43	10/54
43	7/23	7/21	7/22	7/22	9/28	9/33	9/34	10/39	10/46	10/56
41	7/24	7/23	7/22	7/25	9/28	9/33	9/34	10/40	10/47	10/58
39	7/26	7/23	7/22	7/25	9/29	9/33	10/37	10/43	10/47	11/61
37	7/28	7/25	7/23	7/26	9/30	9/35	10/38	10/43	10/48	11/61
35	7/28	7/25	7/24	7/27	9/33	9/35	10/39	10/44	10/52	11/64
33	7/29	7/25	7/26	8/29	9/33	10/38	10/41	10/44	10/53	12/67
31	7/31	7/27	7/29	8/29	9/36	10/39	10/42	10/47	11/58	13/73
29	7/31	7/31	7/31	8/31	10/38	10/40	10/43	11/49	11/61	13/75
27	7/32	7/31	8/32	8/31	10/39	10/41	11/45	11/52	12/67	13/82
25	7/35	7/34	8/32	11/45	11/45	11/48	11/57	13/72	13/84	15/98
23	7/37	7/38	8/33	10/41	11/45	11/47	11/56	12/63	13/77	13/89
21	7/39	9/42	9/38	10/43	11/46	11/51	12/58	12/70	13/81	14/97
19	8/43	9/43	10/42	10/45	11/50	12/56	12/62	12/76	13/86	15/102
17	9/48	9/44	10/46	11/49	12/54	12/62	12/70	12/82	14/98	15/110
16	9/52	10/49	10/52	12/55	12/62	12/67	12/81	13/92	14/105	16/124

As can be seen, no combination of the parameters could identify more than about 40% of the 38 outlier. An additional number of combinations of U and C, for U > 1 and C > 47, were also tried without any diverging results.

C The Outliers

What kind of outliers were then found? The sum of the absolute values of the changes shown in table 1 equalled 12 997 728. This was taken as a base for comparison with the results reported in B above.

TABLE 4.3-3 shows the impact of changes found by the HB-Method.

Number of changes found	Accumulated sum of changes found	Accumulated sum relative to the sum of all changes
7	5 550 152	42.7 %
8	5 562 834	42.8 %
9	12 958 565	99.7 %
10	12 959 665	99.71 %
11	12 960 234	99.71 %
12	12 960 304	99.71 %
13	12 960 305	99.71 %
14	12 962 566	99.73 %
15	12 964 995	99.75 %
16	12 966 840	99.76 %

Considering the three aspects outlined in 4.3.3.3 together with a restriction on the amount of observations appointed by Hidiroglou-Berthelot for a check the number of combinations diminished considerably. For convenience combinations given by U=0.4 or 0.5 and C between 15 and 43 would be preferable.

Another point of interest is what the observations flagged by the H-B Method looked. This question is answered by Table 4.3-4.

TABLE 4.3-4 shows the observations flagged by the H-B Method using U=0.4 and C=41.

OBS	DECEMBER	JANUARY	OBS	DECEMBER	JANUARY
1*	14648285	7403131	15	22773	103
2*	1560	1605000	16	11434	789
3*	1032	973693	17	16166	1100
4*	1500	925000	18	61536	6040
5*	1302	902805	19	105933	8310
6*	972	593536	20	156455	7568
7*	1926	501675	21	11363	293
8*	8473	38010	22	19849	483
9*	294600	54	23	8412	34
10	657	7063	24	248	3
11	315	4985	25	66730	723
12	5500	250	26	264900	700
13	10396	34	27	18006	1478
14	5382	274	28	15074	663

All of these observations flagged by the H-B edit have been exposed to a relatively large change.

With the currently used method the first 9 of these observations, marked with a '*', as well as 29 others were checked and changed (see the outlier population in Table 4.3-1).

D *Effects on the Other Survey Variables*

Using one of the best combinations, U=0.4 and C=41, as a check of the effect on the other variables the results of Table 4.3-5 were produced.

TABLE 4.3-5 shows the impact on the other variables edited by the HB-Method with the values on the parameters C and U which were used for the variable "deliveries to the domestic market".

VARIABLE	DELIVERIES		ORDERBOOK		STOCKS	
	Domestic	Foreign	Domestic	Foreign	Domestic	Foreign
Number of changes	38	19	109	72	31	30
Number of HB changes	9	4	6	6	4	4
Number of flags	28	25	30	35	12	10
Sum of all changes	12 997728	1 157332	9 080144	825626	4 316568	6 096551
Sum of HB changes	12 958562	1 029825	8 701388	485480	4 209380	5 856540
Relative the sum of all changes	99.7 %	88.98 %	95.82 %	58.80 %	97.52 %	96.06 %

Looking at the percentages of changes represented by the outlier that were found, the results are with one exception in agreement with the current method. Obviously there are a few observations in each population of outlier that are responsible for the largest changes.

E Comments

The purpose of this study was to adapt the Hidiroglou-Berthelot edit to a Swedish material.

Hidiroglou-Berthelot edit succeeds in its objective of finding observations that have been submitted to a relatively large change from one period to a succeeding one.

Some points that deserve mentioning are :

1) The edit has to be performed only once to identify all suspect observations.

2) The material itself supports the information needed.

3) There are two possibilities, the parameters U and C, to control the number of observations to be flagged.

4) Little time and efforts are spent on checking observations of minor importance.

Points 2 and 3 show the periodical adaption and flexibility that exist within the limits of the edit.

4.3.4 Description of the HB-Method for Implementation in EDP Systems

The HB-method is a ratio method for which the bounds are automatically calculated from the data to be edited.

To calculate the lower and upper bound for the ratios checks the following is done.

Using the original file the median for the ratio is calculated. The file containing the median is then joined with the first file. A new file containing the old variables and the new variables S and E is then created.

$$S = \begin{cases} 1 - R_{MEDIAN} / R \,, & 0 < R < R_{MEDIAN} \\ R / R_{MEDIAN} - 1 \,, & R > R_{MEDIAN} \end{cases}$$

$$E = S * (MAX(X_i, X_{i+1}))^U$$

The constant U can be choosen in different ways but it should be between 0 and 1.

The next operation is to compute the first and third quartiles and the median for E.

These variables are joined to the file from the previous step to be used for the following calculations. The constant A is recommended to be 0.05.

$$D_{Q1} = MAX(E_{MEDIAN} - E_{Q1} , | A * E_{MEDIAN} |)$$

$$D_{Q3} = MAX(E_{MEDIAN} - E_{Q3} , | A * E_{MEDIAN} |)$$

Now the upper and lower bounds for the ratio check can be calculated. The constant C has to be decided in advance.

$$l = E_{MEDIAN} - C * D_{Q1}$$

$$u = E_{MEDIAN} - C * D_{Q3}$$

These values are then inserted in the ratio checks in the following way:

IF $X_{t+1}/X_t < l$ or $X_{t+1}/X_t > u$ THEN signal error

4.4 THE TOP-DOWN METHOD

4.4.1 Contents

The Top-Down Method is first presented as it is implemented in the main-frame production system of "The Survey of Delivery and Orderbook Situation" (DOS). The production system is written in APL. The description is an abbreviated merged version of Granquist (1987) and Granquist (1991). It also contains a presentation of DOS and its processing system, including the alternative micro-editing system.

Then there is a report of a prototype developed in PC-SAS for editing the same survey. It was developed to save machine costs and to facilitate the system maintenance because there is a general lack of APL competence at Statistics Sweden. The description of the work with this implementation of the Top-Down Method is rather detailed. It follows Lindblom (1990).

Finally there is a description of the method as an EDP function, taken from Lindström (1990 b).

4.4.2 The Survey and Its Main Frame Editing Procedures

4.4.2.1 Introduction

The idea behind the Top-Down Method is to sort the values of the check functions (which are functions of the weighted keyed-in values) and start the manual review from the top or from the bottom of the list and continue until there is no noticeable effect on the estimates.

The method is described as it is applied in the DOS production system. The generalization is obvious.

4.4.2.2 The Delivery and Orderbook Situation Survey (DOS)

DOS is a monthly sample survey of enterprises. There are about 2000 reporting units, kind of activity units, sampled once a year from the Swedish register of enterprises and establishments. It estimates changes in deliveries and the orderbook situation (changes and stocks) every month for both the domestic and the foreign market for the total Swedish manufacturing industry and 38 branches (classified according to the Swedish standard industrial classification of all activities).

The questionnaire is computer printed. The reported values for the two previous months are preprinted for the six variables. The questionnaire thus contains three columns. If the respondent cannot give data for the calender month he has to indicate data about the period to which the reported data refer.

The entry of the questionnaire data is carried out in three batches every production cycle by professional data-entry staff. The last questionnaires are entered directly interactively and then checked by the editing program, which contains checks of formal errors on micro level.

The whole production process was thoroughly revised, and the new system replaced the old one (from 1970) in 1986.

4.4.2.3 The Editing and Imputation Process

Within the new system there are two editing procedures, one automatic adjusting procedure and one automatic imputation procedure. As the correction processes are the same irrespective of which of the editing procedures is applied, they are described first.

A *The Automatic Correction Procedures*

Data are adjusted automatically if the report period differs from the present calender month. This is performed before the editing and thus the adjustments will be checked. In the old system these adjustments were carried out manually.

Imputations are only made for non-response (unit as well as partial non-response). The imputation procedure utilizes data reported by the unit for the six previous months which are adjusted by means and trends from the branch to which the unit belongs. The imputations form part of the estimation procedure and are never reported back to the respondents nor used in imputations for succeeding months. They are carried out after the editing, and thus they are not checked.

B The Micro-editing Procedure (MIEP)

MIEP was supposed to be the main editing procedure. However, it was used only in the first run on the survey for which it was supposed to be regularly used.

MIEP is a traditional record-checking procedure. Such procedures are used in practically all surveys carried out by Statistics Sweden. Data classified as erroneous or suspect according to specified checks are printed on error lists or displayed on the screen. The error messages are then scrutinized manually. Very often they involve telephone calls to the respondents. The corrections are entered interactively and then again checked by the editing rules. However, the number of corrections (detected errors) is relatively small.

The system has an "Okey"-function (key) which makes it possible to approve original data or changes found to be correct, though they do not pass all the checks.

Validity checks, consistency checks and ratio checks (against the previous month) are used. The acceptance regions are intervals, which can be up-dated any time during the editing process.

When the new production system was run for the first time, the MIEP was not a success. It produced so many error messages that the subject matter specialists realized that they had neither resources nor energy to handle all the error messages. Especially as they knew by experiences from the old system that only a small percentage of the flagged data indicated detectable errors.

They had built up the checks on basis of all the experience and subject matter knowledge they had gained during years of work with the survey. The procedure was flexible, user-friendly and easy to fit to the data to be edited. In spite of all that the procedure did not work. Of course this was a

great disappointment.

4.4.2.4 The Top-Down procedure

The basic and simple idea behind the Top-Down Method is to study those observations which have the greatest impact on the estimates.

The procedure is governed by an inter-active menu program, written in APL. The in-data file is successively built up by the records of the three batches from the data-entry stage. There are three functions to select the records to be studied, i.e.

i) the 15 highest positive changes
ii) the 15 highest negative changes
iii) the 15 greatest contributions

which for every variable can be applied to the total and to the 38 branches. For a selected function and domain of study, the screen shows the following list for the 15 records of the in-data file sorted top-down:

IDENTITY	DATA	WEIGHT	WEIGHTED VALUE	SUM
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...
...	...	...	...	...

Having all the top 15 records on the screen the operator can select any of the records to be shown with its entire content and then examine it to see whether an error is committed. If an error is identified, he can up-date the record directly on the screen and immediately see the effects. The record usually loses its position on the top 15 list, another record enters the list and the sum value is changed. This scrutinizing of the top 15 list goes on until the operator can see that further editing will only have negligible effects.

Theoretically all the records can be inspected manually in this way. In practice only a few records of each of the three batches of the entered records are scrutinized entirely.

A *Experiences*

A study expressly designed to compare the Top-Down Method with a corresponding micro-editing method has not yet been carried out. However, the implementation of the Top-Down Method in the DOS processing system, as can be read above about the new micro-editing system, has made such an evaluation study unnecessary.

The Top-Down procedure, which had been developed as a complementary or reserve procedure to the new micro-editing procedure, had to be taken in use at once.

The staff is very satisfied with the new editing method. It is continuously educating the staff on the subject matter and on the production and presentation problems of the survey. It is the user, who by his growing knowledge and skill focuses his work on the most important errors in the data of the studied period. In traditional editing programs the user is dominated by these errors and distributions of earlier periods. By this Top-Down procedure, the user does not have to handle a great number of unnecessarily flagged data, and he has the satisfaction of being able to see the effects of his work immediately.

Since the first processing with the macro-editing method, the number of records for manual review has decreased slowly. The subject-matter statisticians have become convinced that there is no need for editing on the industry level. The Top-Down lists are now only produced at the total manufacturing industry level. Though there still seems to be a certain amount of over-editing it is doubtless the most rational editing procedure at Statistics Sweden.

According to Anderson (1989a) the method is also considered as the most efficient out-put editing method in use at the Australian Bureau of Statistics.

B *Conclusion*

We have shown that the Top-Down Method can be used as an editing method during the processing of a survey without losses in quality and timeliness. The method can reduce the verifying work by 50-75 per cent.

The subject-matter clerks are very satisfied because they feel in control of the editing task and can see the effects of their work.

For practical reasons the method should not be applied to more than ten variables of a survey at the same time. The number of variables may gradually grow, according as the subject matter statisticians learn about the effects of the editing procedure.

4.4.3 The PC-SAS Prototype

The SAS system is used as software, especially the two modules:

SAS / AF - to build the menu-system

SAS / FSP - to edit records

The main purpose of the macro-editing method Top-Down is to concentrate the editing to the most serious errors. Information recieved from the respondents is studied by ranking the records top down after their impact on the estimate.

It is possible to perform three different editing functions with this prototype:

- Largest contribution to the estimate of the total

- Greatest (absolute) change between two months

- Greatest (absolute) deviation in a consistency check

These three functions can be performed on the whole population or just for one of the 38 branches. Naturally it is possible to choose the month(s) and variable(s) to be studied.

The 15 establishments with the largest impact on the estimate are shown on the screen, one on each row. Column headings are:

* identity
* the number of the branch
* value of the variable
* raising factor
* raised value

The operator then marks on the screen the establishment to be more closely examined. The information received from that establishment for the past 4 months is then shown on the screen.

Prospective errors can be verified immediately and if a record is updated a new top 15 list will be presented on the screen. This is because an updated record usually loses its position on the top 15 list and another record replaces it. Therefore the effect of the verifying can be seen directly on the ranking list. Consequently it is possible to deal with the largest errors one by one as they appear.

The operator changes domain of study or finishes editing when the prospective errors no longer affect the estimate.

The survey includes about 2000 records with 6 variables of study and a few other variables. The performance of the PC depends strongly on the number of records used in the calculations. Therefore it is important to minimize the number of records used for each function and domain of study. For example, if branch is specified then only those records belonging to that branch will be included in the further calculations.

Producing the first list of the top 15 records for a selected function and domain of study means entering data and creating a data-set with the appropriate records, calculating sums/differences and sorting the data set top down.

This part of the application requires the most time, about 50 seconds if branch is specified and about 2 minutes if not. The PC used for this prototype was a Special Instrument with a 286 processor and 1 MB expanded memory.

After the first list of the top 15 records is shown on the screen only the 30 records which have the largest impact on the estimate are kept in the created data set to improve the performance. This is because if an error is identified and updated the record usually loses its position on the top 15 list and another record enters the list and the sum/difference value is changed. Theoretically all the records can be inspected but in practice only a few records are inspected before the operator changes to a new domain of study.

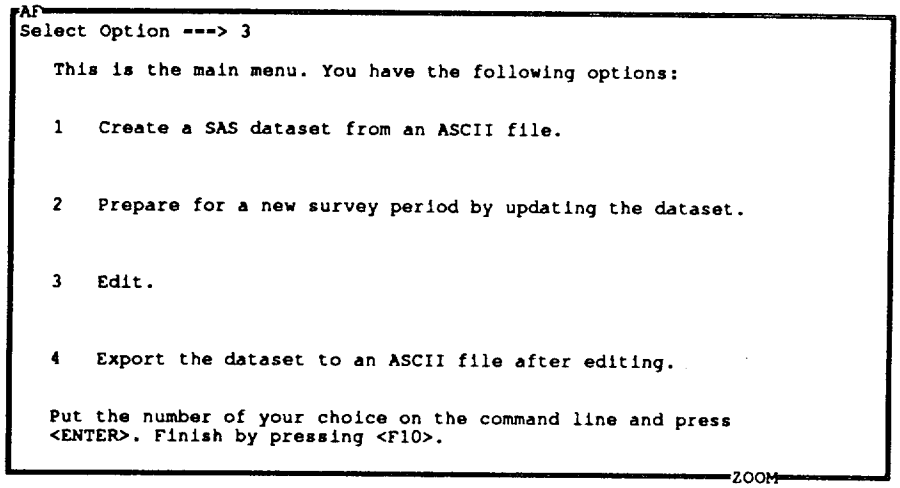
Therefore it is sufficient to have a data set containing only the 30 records which have the largest impact on the estimate because records with lower

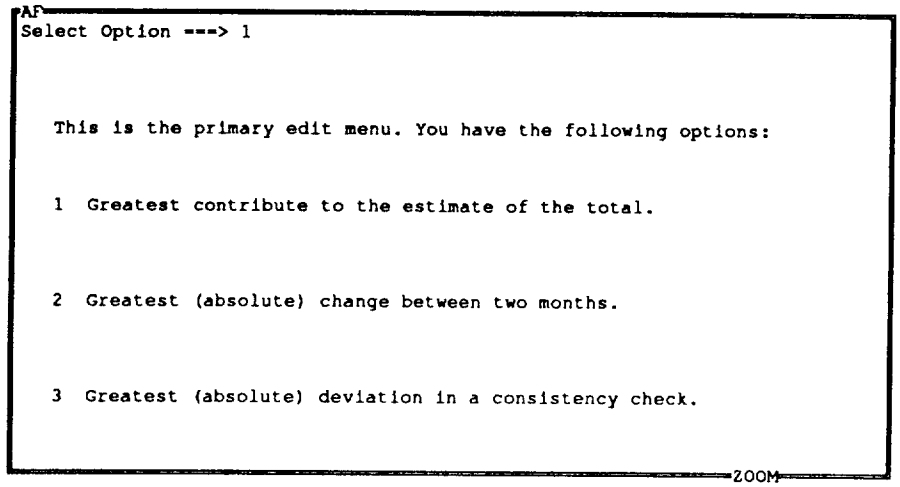
impact will never be candidates for the top 15 list.

The gain is when a new top 15 list for the same function and domain of study is to be produced after updating because the sorting procedure then involves at the most 30 records. Sorting requires a lot of time and depends naturally on the number of records to be sorted. But when the operator changes domain of study and/or function, data must be entered again and new calculations must be made.

When a record is updated the sum/difference is only changed with the actual correction in the record before the new top 15 list is presented. That is the sum/difference is not recalculated for all records in the data set.

A short demonstration of the application is given below:

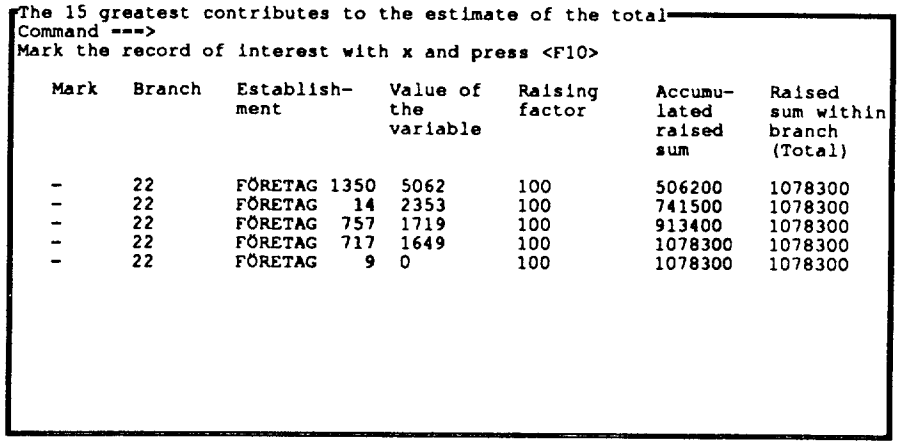




```
AF
Command --->

 Fill in the number of the branch (All branches just press <ENTER>)

 22


 Fill in the name of the variable:

 levh                      LEVH = Domestic Deliveries
                           LEVE = Foreign Deliveries
                           ORDH = Domestic Order Book
                           ORDE = Foreign Order Book
                           STOH = Domestic Order Stock
                           STOE = Foreign Order Stock

 Fill in the number of the month:  3 = This month
                                   2 = Last month
 3                                 1 = Two months ago

 Press <ENTER>
```

```
The 15 greatest contributes to the estimate of the total
Command --->
Mark the record of interest with x and press <F10>

  Mark   Branch   Establish-    Value of    Raising     Accumu-     Raised
                  ment          the         factor      lated       sum within
                                variable                raised      branch
                                                        sum         (Total)

  x      22       FÖRETAG  757  151900      100         15190000    16096400
  -      22       FÖRETAG 1350  5062        100         15696200    16096400
  -      22       FÖRETAG   14  2353        100         15931500    16096400
  -      22       FÖRETAG  717  1649        100         16096400    16096400
  -      22       FÖRETAG    9  0           100         16096400    16096400

                                                                    ZOOM
```

```
FSEDIT S.ORDER
Command --->                                                        Obs   7

Report number:        Establishment:   Branch:      Raising factor
     20382609          FÖRETAG  757        22             100

 Period:   8708          8709          8710          8711

 LEVH:     138000        149300        162300        151900

 LEVE:     0             0             0             0

 ORDH:     138000        149300        162300        151900

 ORDE:     0             0             0             0

 STOH:     0             0             0             0

 STOE:     0             0             0             0

 Enter your changes and press <F10> or press just <F10>
 (Move between the fields with the <TAB> key)
```



The 15 greatest contributes to the estimate of the total
Command ===>
Mark the record of interest with x and press <F10>

Mark	Branch	Establish-ment	Value of the variable	Raising factor	Accumu-lated raised sum	Raised sum within branch (Total)
-	22	FÖRETAG 1350	5062	100	506200	1078300
-	22	FÖRETAG 14	2353	100	741500	1078300
-	22	FÖRETAG 757	1719	100	913400	1078300
-	22	FÖRETAG 717	1649	100	1078300	1078300
-	22	FÖRETAG 9	0	100	1078300	1078300

4.4.4 The Top-Down Method for Implementation in EDP Systems

The Top-Down method is an interactive technique for error localization and correction. The method works best with small surveys with a small number of variables. The description is based on the example presented in 4.4.2.4.

The method is implemented by using a number of screens which are displayed for the user. The user interacts with the system by entering answers to the screens.

The first screen displayed is a screen for selecting the check function to be used on the survey. Example of check functions are the main contributor to the estimate, the greatest difference between a variable at two different survey periods etc.

After selecting the check function to be used a screen is displayed where the user can decide which domain the check function should be applied to and which variable to use the check function on.

After this screen has been filled in the following computations are made. From the file containing all the records a selection is made of the records belonging to the domain of interest.

The check function is applied to this file. In the case of the main contributor to the estimate, the estimates are computed, in the case of the differences

those are computed etc. The absolute value of the check function is then used to sort the file in ascending order.

Then the fifteen first records on this file are displayed on the screen. On the screen the following variables are displayed: identity of record, value of variable, weight, estimate for the domain, cumulative estimate for the domain.

On the screen the user can select a record for updating by typing an x in front of the record. The record will then be displayed and the user has the possibility to update the record, the record will be updated both on the file used for presenting the list and on the original file.

After having updated the record new values are computed for the estimate and the cumulative estimate for the domain. A sort using the value of the check function is preformed once again. Thereafter the screen displaying the fifteen first record on the file is displayed again. On this screen the updated record should either have disappeared or have changed its position.

The updating now continues until the user is satisfied. Thereafter it is possible either to finish the editing or to choose a new check function.

4.5 OTHER METHODS

4.5.1 Introduction

This section presents an outline of a Box-Plot method and the Box Method both considered as potential macro-editing methods at Statistics Sweden.

The Box Method is a graphical editing method. Presently, there is a great interest in "graphical editing" methods. Prototypes for graphical editing are under development in Australian Bureau of Statistics, U.S. Bureau of Labor Statistics and Statistics Sweden. The concept is under consideration by various other statistical agencies, among them U.S. Department of Agriculture and Statistics Canada. Because of the developing state of this editing method, there is hardly any papers on this topic. However, one published paper is Hughes et al (1990), presented at the Sixth Annual Research Conference of the U.S. Bureau of the Census and discussed in Kovar (1990) at that conference.

4.5.2 The Box-Plot Method

4.5.2.1 Introduction

The concept of Box-Plots was introduced by Tukey (1977). The Box-Plot Method as a micro-editing method is reported in Anderson (1989 b), and is suggested as a macro-editing method in the discussion on the Aggregate method in Granquist (1991, 6.3) and in 4.2.2 above.

Anderson (1989 a) reports an experiment carried out on data from the Australian Bureau of Statistics (ABS) survey "Average Weekly Earnings" (AWE). ABS extended the bounds of every ratio check used in that survey to

$$(Q_1 - 3*I_{QR} ; Q_3 + 3*I_{QR})$$

where Q_1, Q_3 and I_{QR} are respectively the first and the third quartile and the interquartile range. This means that the manual review work became limited to "extreme" outliers according to the definition given by Hoaglin et al (1983).

The study indicates that 75 per cent of the resources for the manual reviewing of flagged data from the whole editing process could be saved by limiting the manual review to "extreme outliers". It is proved in the report that the remaining errors in data had no significance at all on the estimates of the survey used for the experiment. Actually, Anderson used the same method for evaluation as in the studies related above.

In the latter version of the report, Anderson suggests that the lower and upper bounds of the checks should be set on the basis of a manual analysis of box-plots of the check functions. Then the bounds can be modified by taking into account outliers near the bounds for extreme outliers. Above all, the survey staff will by such a procedure get full control of responsibility for the data editing.

4.5.2.2 Description of the method

The distributions of the check functions of the weighted keyed-in values are displayed as box-plots. Then the acceptance intervals for the checks are set by the staff on the basis of these graphs and put into the regular error-detecting program.

By Anderson (1989 a) and the studies on the Aggregate Method reported in this paper, the definitions of extreme outliers may serve as guidelines for efficient limits. The only difference to the method reported in Anderson (1989 b) is that the values should be weighted by (approximately) the inflation factor (according to the sample design) to be a macro-editing method of the type which is the subject of this report.

4.5.3 The Box Method

4.5.3.1 Introduction

The Box Method is a graphical macro-editing method under development at Statistics Sweden. A first version of a prototype for the Survey of Employment and Wages is expected by March 1992.

The basic principles are to utilize computer graphics to visualize the distribution of the check function of the weighted data and the interactivity of a computer to get indications of when to stop the manual verifying work. The method may be considered as a combination of a generalized Box-Plot method and the Top-Down Method.

4.5.3.2 Description of the method

The keyed-in data are weighted and then put into the check function. Any mathematical expression may be used as a check function. The values of the function are plotted on the screen and acceptance regions of any shape can also be provided. The reviewer draws a box around the observations he wants to review. On the screen, the data then appear on in advance selected items of the records belonging to the data points inside the box. For every check function the user can select the items of the records to be displayed. A change is entered inter-actively and some data (statistics) on the impact of the change will be displayed.

The method may also be used as a tool to find appropriate values of acceptance regions for other editing methods (e.g. the HB-Method), as shown by Figure 4.3-1 in 4.3.2.

5 DISCUSSION OF MICRO-MACRO METHODS

5.1 CONTENTS

First, a summary of the characteristic features of the specific macro-editing methods discussed in chapter 4 is given in 5.2. Then, in 5.3 this kind of macro-editing methods are compared with traditional micro-editing methods from an operational point of view. It is stressed that this kind of macro-editing methods is a way to improve corresponding micro-editing methods. In 5.4 and 5.5 alternative methods to rationalize the manual review in editing processes are presented.

However, almost all the methods discussed so far only detect randomly appearing errors in data, thus having a negligible impact on quality. To improve the quality substantially, the methods have to focus on systematic errors, e.g. the so-called in-liers mentioned in the description of the process reported in 5.5. The systematic errors or the bias in reported data is the subject of the Response Analysis Surveys reported in 5.6.

The chapter concludes with a brief summary of the whole paper and an overall statement of the merit of the macro-editing methods as compared to traditional editing methods.

5.2 BRIEF OUTLINE OF THE CHARACTERISTICS OF THE METHODS

The basic principle of all the reported macro-editing methods is that the acceptance regions are determined solely by the distributions of the received observations of the check functions. The keyed-in values of the variable to be checked are weighted by the expansion factor before they are put into the check function.

In the Aggregate, Box-Plot, HB and Top-Down methods all the values of the check function are sorted by size.

The tails of the distributions are displayed and analyzed in the Aggregate and the Box-Plot methods in order to set the acceptance bounds for the checks. In the HB-method the acceptance limits are set automatically.

In the Top-Down and the Box methods the effects of the detected errors on

the estimates "determine" how far the manual review should go. In the Top-Down Method the manual review work starts with the extreme values and goes towards the median value, while in the Box Method the records for manual reviewing are selected by the reviewer from a graphical display of the values of the check function. This selection may be supported by guidelines (acceptance regions) displayed in the same graph.

The choice of method should be based on the number of variables to be edited by the macro-editing method and on how the staff wishes to work.

5.3 MACRO-EDITING VERSUS MICRO-EDITING METHODS

Macro-editing is not a new concept. It has always been used in data editing, but only as a final checking. Another term is out-put checking.

What is new is that such out-put check methods can be used in data editing procedures in the same way as micro-editing methods and that they have proved to be much more efficient than traditionally applied micro-editing procedures. Here reported studies reduce the work by 35 - 80 per cent.

Macro-editing methods of the type described in this paper may be considered as a statistical way of providing micro-editing checks with efficient acceptance limits. The limits are based only on the data to be edited. The methods bring a kind of priority thinking to the verifying work, and the data are edited according to their impact on the estimates. The macro-editing methods solve the general problem inherent in micro-editing methods, i.e. that they produce too many error signals without giving any guidance as to how the resources of the verifying work are to be allocated. We have seen that with micro-editing procedures even very large errors are not always detected, due to the large number of flagged data.

However, there are no limits to the number of cases that can be selected for a manual review. The reviewer can select all the cases he deems necessary. The difference to micro-editing methods is that the cases are selected in priority order, i.e. according to the impact they may have on the estimates. The selection is mainly done by the reviewer, which means that he governs and has the full responsibility for the whole scrutinizing work. In micro-editing procedures, the reviewer is dominated by the computer and cannot see the effects of his work.

Both procedures focus on randomly appearing negligence errors and utilize the same principle to indicate outliers or extreme observations. Micro-editing procedures flag data by criteria fixed in advance, based on historical data, while macro-editing procedures focus on data, which at that very moment and relative to the estimates are the most extreme ones.

5.4 SELECTIVE MANUAL REVIEWING

Latouche & Berthelot (1990) launch the idea of using a score function to identify records requiring manual review in a typical editing process. The report presents an experiment carried out on the Canadian Annual Retail Survey and indicates that only a third of the records flagged by a typical editing process need manual review.

5.4.1 The Idea

"Selective micro-editing" is the term used in Kovar (1991) for limiting the manual review of flagged records to the most important ones. Only those flagged records with a potentially large impact on the estimates, based on a score-function, will be manually reviewed.

The score function gives a score to each suspicious record on the basis of the following criteria:

i) the size of the unit
ii) the potential impact of the error on the estimate
iii) the survey weight
iv) the number of flags for that record, and
v) the relative importance of each flagged item.

5.4.2 The Study on Canadian Annual Retail Survey Data

In order to study the score functions, the data collection and data capture process was simulated using data from the 1987 Canadian Annual Retail Trade Survey (CARTS). The score functions were evaluated against the final 1987 data using the same method as applied in the studies of the macro-editing methods reported in this paper (see 4.1.2).

After a record has been flagged as suspicious, a manual review is possible. A manual review was simulated by replacing all the 1987 reported data of a questionnaire flagged for manual review by the corresponding 1987

tabulated data.

For any given number of records flagged (review rate), an estimate $Y_{i,87}$ of the total for the full sample was computed using 1987 reported values for non-flagged units, and 1987 released values for flagged units. The absolute relative discrepancy between $Y_{i,87}$ and the total $Y'_{i,87}$ that was released in 1987 constituted the pseudo-bias:

$$\text{absolute pseudo-bias} = 100 * (|Y_{i,87} - Y'_{i,87}|)/(Y'_{i,87})$$

Different review rates, 0, 17, 34, 50 and 100 per cent, using the same score function, were applied to the sample.

5.4.3 Results

For all review rates the pseudo-bias grew small for the frequently reported variables. There is a significant decrease in the standard error in going from a 17 to a 34 % of review. But little is gained in going from 34 to 100 %. Thus, as a result of the study, a 34 % review rate is recommended.

TABLE 5-1 (reprint from Latouche & Berthelot), shows the overall pseudo-bias, obtained with a 34% review rate and the percentage of time the variable is reported.

VARIABLES	% PSEUDO-BIAS	% TIME REPORTED
NET SALES	-0.63	100
GROSS COMMISSION	13.91	5
RECEIPTS FROM REPAIRS	-1.43	46
RECEIPTS FROM RENTALS	-0.08	10
RECEIPTS FROM FOOD SERVICES	2.68	2
OTHER OPERATING REVENUE	-13.20	15
TOTAL NET SALES AND RECEIPTS	0.18	100
NON OPERATING REVENUE	-1.46	22
OPENING INVENTORY	-0.87	100
CLOSING INVENTORY	-0.34	100
PURCHASES	-2.23	99
SALARIES	-0.92	99

5.4.4 Conclusion

The standard error of the discrepancy was not significantly reduced when more than one-third of the remaining errors were reviewed. For infrequently reported variables, quality could be improved by increasing their importance weight in the score-function.

Although the simulation was run only for the Annual Retail Trade Survey, Latouche & Berthelot believe it is representative of business surveys in general.

The study shows that it is sufficient to review a limited number of units to ensure the same data quality as obtained when reviewing **all** the records flagged by the computer review. If this approach is used in production, the editing process becomes faster and valuable resources can be saved or reallocated to other purposes without significantly affecting the data quality.

5.5 IMPROVING COST-EFFECTIVENESS AND QUALITY BY TARGETING THE EDITS ON SERIOUS ERROR TYPES

Mazur (1990) reports on developing more efficient edits in a micro-computer environment for the Livestock Slaughter Data Survey at the U.S. Department of Agriculture.

5.5.1 Background

A census of federally inspected livestock slaughter plants is conducted each week using a one page mail questionnaire. U.S. Department of Agriculture, (USDA), livestock slaughter inspectors in the plants report daily number of head killed, and weekly weight totals.

Several limitations in the old editing system prompted the research project. Range checks with predetermined bounds were used which identified too many data incorrectly as outliers. Another problem were inliers, for example when the same values were reported every period.

5.5.2 The Idea

The main goal of the project was to create a statistical edit, unique for each plant, by utilizing the plant's historic data to define edit limits for that plant.

Tukey's biweight (see Hoaglin et al (1983)) was finally selected to calculate edit limits, as it worked well in varying data distributions.

The report covers cases where outliers and inliers (suspicious values in the middle of the data) were found.

5.5.3 Results

The new edit system resulted in substantial cost savings. Compared to what it would have cost on the current edit system, the new system achieved a cut of approximately 75 %, according to the author. However, part of the cost savings were certainly due to changes in the computer environment, which are not mentioned in the report. Nor does Mazur report whether there was any significant improvement in quality due to the new inlier checks.

Another improvement is that the system can tell the proportion of imputations being done with weight data, and make the management aware of this very substantial proportion.

Adding to that, the survey staff is very happy with the new system, as it makes their jobs much easier and takes less time.

5.5.4 Conclusion

The study reports savings of some 75 % of the editing costs, improved quality, and more information from the editing system, all to a great extent due to using statistical methods for the checks. The author claims that the idea seems to be applicable to a series of USDA surveys.

5.6 RESPONSE ANALYSIS SURVEYS FOR CONTROLLING RESPONSE ERROR

Werking et al (1988) report from the development of an ongoing record check program, which uses computer-assisted telephone interviewing (CATI) for controlling response error and a fully automated touchtone data collection system for employers to quickly and easily self-report their data.

5.6.1 Background

The authors state that inter-active editing is particularly effective in surveys where data are compared over different reporting periods. However, the

bounds of the checks used are usually too broad to detect many definitional inconsistencies and thus, *insufficient for identifying and correcting consistently reported errors.*

The CATI collection program in the Current Employment Survey (CES) allows on-line editing of reported data, including range checks, internal consistency checks, and longitudinal editing checks. However, even with CATI improvements, this inter-active data review will generally catch only gross errors, and will offer little protection against widespread, systematic small errors, or large errors consistently reported over time.

5.6.2 The Approach

Current research demonstrates that a record check survey conducted with the respondents can identify and correct many of the systematic and consistent errors which traditional editing misses. For example, a range check for average hourly earnings for production workers in a particular industry would most likely not detect the exclusion of both overtime hours and pay. The range checks are intended to detect data which are unreasonable relative to similar firms (by industry and size), and must therefore be broad enough to accommodate normal differences among firms.

The aim of the Response Analysis Survey (RAS) is to identify and correct for bias that may be present in ongoing periodical establishment surveys. Therefore the RAS instrument is designed to review the survey definitions against the firm's recordkeeping system, thereby permitting to identify differences in definitions, and discrepancies to be corrected in the future.

5.6.3 The Study

The effect of the RAS on bias in the Current Emplyment Survey, (CES), was measured in two ways.

First, separate estimates were made of the characteristics of interest for the test group (those participating in a RAS) and the entire CES sample before and after conducting the RAS.

Second, indirect measures of bias were obtained for each data element by tabulating the types and number of adjustments agreed to for the RAS sample.

The following finding should be noted in particular, because it is a very

striking argument against those who justify intensive micro-editing with many and tight checks in order to "ensure" every single record to be correct.

"**Over half of the establishments studied did not entirely apply the survey definitions**, primarily due to the recordkeeping system at the establishment."

5.6.4 Results

TABLE 5-2 (reprint from Werking et al) shows the average changes in reported data:

VARIABLE	AVERAGE PERCENT CHANGE SEP.85 TO SEP.86	
	RAS UNITS	CONTROL GROUP
EMPLOYMENT	1.3 (3.9)	4.8
WOMEN	6.0 (2.4)	6.9
PRODUCTION WORKERS	5.3 (2.2)	5.8
PRODUCTION WORKERS HOURS	8.8 (4.2)	6.7
PRODUCTION WORKERS EARNINGS	10.7 (3.2)*	1.6

( ) Standard error
* Significantly different from the control group at 2 SE's

The RAS sample showed a significantly larger change in production worker earnings than did the entire CES sample. There were no significant differences for the other statistics. An examination of the average changes for the RAS sample based on the expected direction of the change indicated that the units changed their reporting habits. While the sample sizes were to small to produce significant results were consistent with the supposition that the adjustments were large relative to economic changes.

5.7 CONCLUDING SUMMARY

As indicated by this report and confirmed in Granquist (1991), it is evident that about 70 % of the manual review work in computer assisted editing processes is unnecessary. This holds true for almost all surveys with quantitative data conducted by statistical agencies all over the worlds. This

explains the success of "selective editing" and the macro-editing methods treated in this paper.

However, the methods do not imply that the quality will be improved essentially. Quality can only be considerably improved by editing if the editing is focused on present misunderstanding errors. Such errors cannot (in principle) be detected by either the macro-editing or the micro-editing methods normally used by statistical agencies. This kind of errors have to be detected by other types of methods, which are quite clear from the experiences presented in Mazur (see 5.5) and above all in Werking et al (see 5.6).

What have been proved in this paper is that the kind of macro-editing methods described here certainly

provide a much more efficient way than traditional micro-editing methods of reaching the same "quality" standard

and that they may release resources for editing misunderstanding errors.

REFERENCES

Anderson, K. (1989): "Draft, Output Edit Study, Average Weekly Earnings", Statistical Services Branch, Australian Bureau of Statistics, September 1989.

Anderson, K. (1989 b): "Enhancing Clerical Cost-Effectiveness in the Average Weekly Earnings", Draft, Australian Bureau of Statistics, Statistical Services Branch, 9 November 1989.

Anderson, T.W. (1958): "An Introduction to Multivariate Statistical Analysis", Wiley & Sons, pp 126-153.

Berthelot, J.-M. (1983): "Wholesale-Retail Redesign, Statistical Edit Proposal", Technical Report, Statistics Canada.

Bethlehem, J.G., Hundepool, A.J., Schuerhoff, M.H., Vermeulen, L.F.M. (1989): "BLAISE 2.0 An introduction", Central Bureau of Statistics, The Netherlands, February 1989.

Boucher, L., Statistics Canada (1991): "Micro-editing for the Annual Survey of Manufactures, What is the Value Added ?", Presented at the Annual Research Conference, Washington,DC, March 20, 1991.

Chinnappa, N., Collins, R., Gosselin, J.-F., Murray, T.S. and Simard, C., (1990): "Macro-editing at Statistics Canada, A Status Report", prepared for the Advisory Committee on Statistical Methods, September 1990.

Cochran, W. G. (1963): "Sampling Techniques", Second Edition 1963.

DEFS (1990): Subcommittee on Data Editing in Federal Statistical Agencies, (1990): "Data Editing in Federal Statistical Agencies" Statistical Policy Office, Working Paper 18, May 1990.

EDIT 78 (1982): "Description of EDIT 78", Statistical Commission and Economic Commission for Europe, Statistical Computing Project, Data Editing Joint Group, SCP/DE/WP.12, April 1982.

Fellegi, I.P. and Holt, D. (1976): "A Systematic Approach to Automatic Edit and Imputation", Journal of the American Statistical Association, Vol.71, 17-35.

Ferguson, Dania P. (1989): "Review of Methods and Software Used in Data Editing", SCP2/DE/WP.33 (U.S. Department of Agriculture, National Agricultural Statistics Service), October 1989.

Forsman, G. (1991:a): "Olika felorsakers betydelse för granskningskostnaderna och skattningarnas kvalitet - En fallstudie på SCBs finansstatistik", GRANSK-PM Nr 22, 1991-02-05.

Forsman, G. (1991:b): "Handledning för effektstudier av granskning", GRANSK-PM Nr 24, 1991-08-25.

Granquist, L. (1982): "On Generalized Editing Programs and the Solution of the Data Quality Problems", Statistical Commission and Economic Commission for Europe, Statistical Computing Project, Data Editing Joint Group, SCP/DE/WP.17.

Granquist, L. (1984:a): "On the Role of Editing", Statistisk Tidskrift 1984:2

Granquist, L. (1984:b): "Data Editing and Its Impact on the Further Processing of Statistical Data", Workshop on the Statistical Computing Project, Budapest, 12-17 November 1984, Invited paper.

Granquist, L. (1987): "Macroediting - The Top-Down Method", Statistics Sweden, Report 1987-04-09.

Granquist, L. (1988:a): "On the Need for Generalized Numeric and Imputation Systems", Statistical Commission and Economic Commission for Europe, Seminar on Statistical Methodology, Geneva, 1-4 February 1988, CES/SEM.23/R.10

Granquist, L. (1988:b): "Macroediting - The Aggregate Method", Statistics Sweden, Report 1988-08-18.

Granquist, L. (1991): "A Review of Studies on Impact of Data Editing on Estimates and Quality", Draft, Statistical Commission and Economic Commission for Europe, Work Session on Statistical Data Editing, Working Paper No.3 Geneva, 28-31 October 1991.

Greenberg, B., Petkunas, T. (1987): "An Evaluation of Edit and Imputation Procedures Used in the 1982 Economic Censuses in Business Division", 1982 Economic Censuses and Census of Governments, Evaluation Studies.

GRUS (Undated): GRUS, Manual, Statistics Sweden, SCB/P.DBM (Swedish).

Hidiroglou, M.A. and Berthelot, J.-M. (1986): "Statistical Editing and Imputation for Periodic Business Surveys", Survey Methodology, June 1986, Vol 12. No 1. pp 73-83.

Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (1983): "Understanding Robust and Exploratory Data Analysis", John Wiley & Sons, ISBN 0-471-09777-2.

Hughes, P.J., McDermid, I., and Linacre, S.J. (1990): "The Use of Graphical Methods in Editing", Proceedings of the Sixth Annual Research Conference of the Bureau of the Census, Washington D.C.

Höglund Davila, E. (1989): "Macroediting - The Hidiroglou-Berthelot Method (Statistical Edits)", Statistics Sweden, Report 1989-03-28.

Kovar, J.G. (1990), Discussion: "Session on Editing", Proceedings of the Sixth Annual Research Conference of the Bureau of the Census, Washington D.C..

Kovar, J., (1991): "The Impact of Selective Editing on Data Quality", Statistical Commission and Economic Commission for Europe, Work Session on Statistical Data Editing, Working Paper No.5, Geneva, 28-31 October 1991.

Lalande, D. (1988:a): "Système de détection des données aberrantes-SIO Documentation-MACRO", Technical Report, Statistics Canada.

Lalande, D. (1988b): "Système de détection des données aberrantes-SIO Documentation-MICRO", Technical Report, Statistics Canada.

Latouche, M. and Berthelot J.-M., (1990): "Use of a Score Function for Error Correction in Business Surveys at Statistics Canada", Draft, International Conference on Measurement Errors in Surveys, Tucson, Arizona, November 12, 1990.

Linacre, S.J. and Trewin, D.J.: "Evaluation of Errors and Appropriate Resource Allocation in Economic Collections", Undated, Internal paper, Australian Bureau of Statistics.

Lindblom, A.: "A review of the macro-editing procedure Top-Down", Data Editing Joint Group Product No. SCP2/D.12/f, June 1990.

Lindström, K.: "A macroediting method application developed in PC-SAS", Data Editing Joint Group Product No. SCP2/D.11/f, May 1990.

Mayda, J.E., Whitridge, P. and Berthelot, J-M. (1990): "An Integrated Approach to Editing", J.E. Mayda, Statistics Canada, 11-D, R.H. Coats Building, Ottawa, Ontario, Canada, K1A 0T6.

Mazur, C., (1990): "A Statistical Edit for Livestock Slaughter Data", SRB Research Report Number SRB-90-01, October 1990.

Pierzchala, M. (1990): "A Review of the State of the Art in Automated Data Editing and Imputation", Journal of Official Statistics, Vol. 6, No 4, 1990.

Pons Ordinas, Juan (1988): "Proceso de Macroedición, Analisis y Transferencias, Macro-Micro en la Encuesta Nacional, Desagregación en Cascada de Tablas de Series", Instituto Nacional de Estadística, España, Documento de Trabajo, Diciembre 1988.

Pritzker, L., Ogus, J. and Hansen, M.H., (1965): "Computer Editing Methods - Some Applications and Results", Bullentin of the International Statistical Institute, Belgrade, 1965.

Pullum, T.W., Harpham, T. & Ozsever, N., (1986): "The Machine Editing of Large Sample Surveys: The Experience of the World Fertility Survey", International Statistical Review, Volume 54, Number 3, December 1986.

Tambay, J.L. (1986): "Study of Outlier in the C.S.I.O., Regional Offices", Technical Report, Statistics Canada.

Tukey, J.W. (1977): "Exploratory Data Analysis", Addison-Wesley Publishing Company, ISBN 0-201-07616-0.

Wahlström, C. (1990): "Granskningens effekter - En studie av SCBs finans-statistik", F-METOD NR 27, 1990-02-26.

Villan Criado, I. and Bravo Cabria, M.S. (1990): "Procedimiento de depuración de datos estadísticos", Seminario internacional de estadística en Euskadi, 1990.

Werking Georg, Tupek Alan, Clayton Richard, (1988): "CATI and Touchtone Self-Response Applications for Establishment Surveys", Journal of Official Statistics, 1988-4.

Wickberg, I. (1990): "Makrogranskning med Hidiroglou-Berthelots metod och med aggregatmetoden utvärderad mot produktionsgranskningen av månadsstatistiken över sysselsättning och löner under november 1987 för industriarbetare", GRANSK-PM Nr 20, 1990-06-25.

Wickberg, I. (1991): "Ytterligare studier av granskningsmetoder på KPI-data avseende mars 1991", GRANSK-PM NR 25, 1991-08-30.

"1985 -1986 Retail Census Edit Evaluation Study", undated, unnamed. Internal paper, Australian Bureau of Statistics.