5 Chapter Censuses

Chair: Clyde Tucker, Bureau of Labor Statistics

Olivia Blum ◆ Eliahu Ben-Moshe

Jan Thomas ◆ David Thorogood

Frank van de Pol ◆ Bert Diederen

5

Chapter

Automated Record Linkage and Editing: Essential Supporting Components in Data Capture Process

Olivia Blum and Eliahu Ben-Moshe,
Central Bureau of Statistics, Israel

Abstract

The data capture process of the Israeli 1995 Census of population and housing is based on an optical character recognition (OCR) technology. The data capture system has been designed and developed while bearing in mind short run targets and long-run goals. The short-run targets are concerned with the data capture itself. These include getting an accurate and reliable file in a relatively short period of time; decreasing the subjective, human component in census data capture; and simplifying control and quality assurance processes. The long-run goals address future needs and uses. These include shifting from microediting to macroediting, to avoid overediting and to make editing more efficient; allowing for reprocessing the data starting with the raw, input data file; and linking census records with previous census files and administrative records. The use of macroediting and the ability to return to a basic file permit recreation of the main census file on a different basis, as future needs become clear. An accurate and reliable file, with the exact values given by the respondents, is also necessary for reprocessing. The search for additional support for the data capture process was motivated by these concerns and the limitations of OCR.

Although automatic processes are embedded in an OCR system, it lacks as a substitute for a human eye-brain mechanism in two main respects. First, the mathematical function used in the recognition process does not specify the whole scope of handwriting styles. As a result, the reliability of the OCR values varies.

Second, enumerators' errors hamper the automatic definition of the process units. Correcting these errors involves altering values in ways beyond the OCR capabilities. Record linkage has been incorporated into the data capture process as external support for OCR in determining values in the questionnaires' fields. This benefits the data capture process and enhances the final file.

From an Optical Reader to an Optical Data-Capture System

In spite of the visible merits of the optical character recognition (OCR) facility, it is not a full substitute for a human eye-brain mechanism. Moreover, the system has been developed under restricted budget; thus, the full potential capabilities have not been realized.

Consequently, supporting components have been added to the optical reader to transform it into an optical data capture system. These components have come to specificly address the following **limitations of the optical reader**:

❒ The recognition process does not cover the whole scope of hand-written styles. Therefore, the reliability of the OCR-suggested values is variable, meaning that relying on the OCR as the only source of identification can be error prone.

In order to utilize the process and channel resources to where they are needed, each OCR value is accompanied by a status, indicating its level of confidence (Super-Sure, Sure, Doubt or Fail). These statuses determine the nature of the treatment needed in the following stage.

❒ The census processing units are records of an individual, an household, or an enumeration area (EA). A record is defined once it is exclusive, exhaustive and unique, meaning that it contains ALL the data of only ONE unit and that there are no duplicate records.

The need to define a record arises when the enumerator's fields on the questionnaire carry wrong values. Hence, defining a processing unit implies altering values in the enumerator's fields. This task is beyond the capabilities of the OCR.

❒ The optical reader has been designed to identify bar codes, preprinted numbers, handwritten numbers and marks (Xs). Because of the limited potential market for recognition of Hebrew letters -- and, therefore, high development costs -- the OCR does not include alphabetic character recognition.

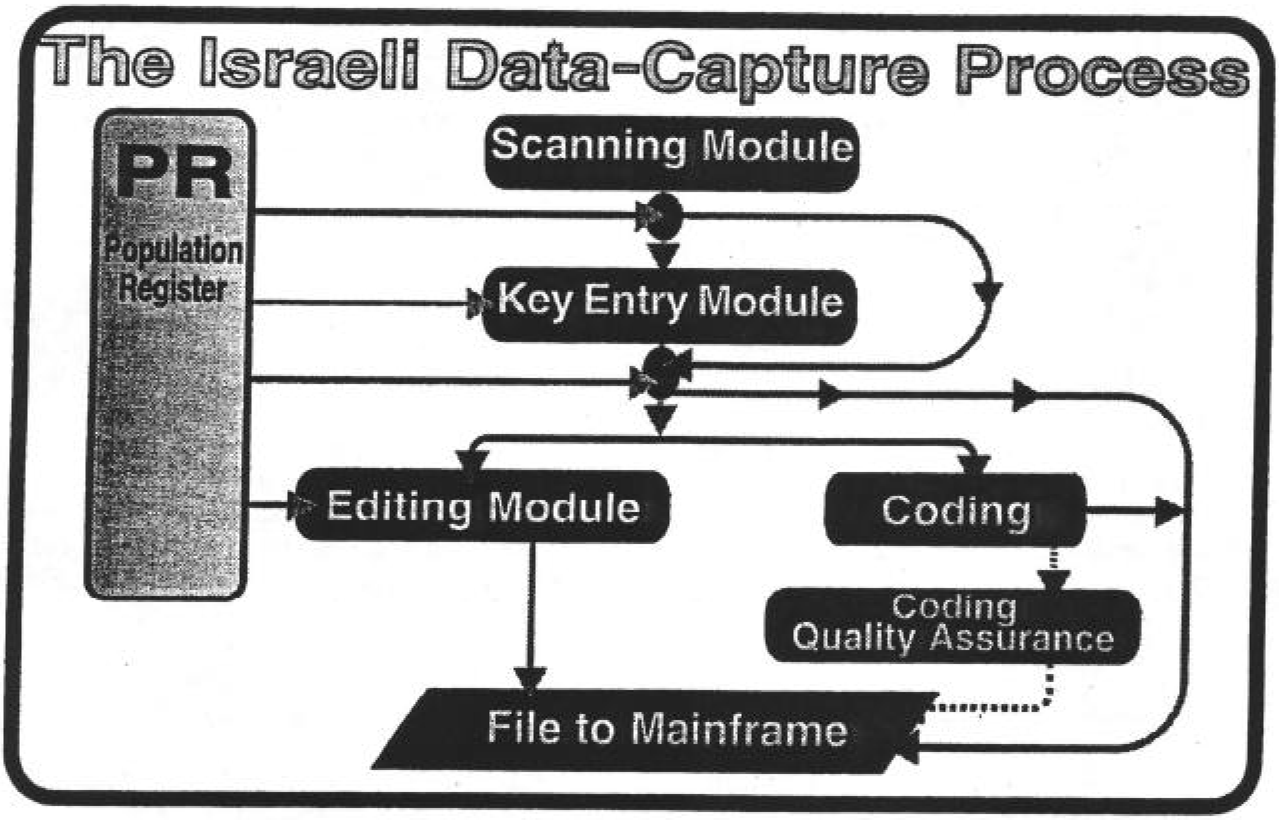
These limitations, along with the requirement to have a raw file, dictate a non-conventional data capture process (see next page).

This process has several unique interrelated features:

❒ Each component in the process supports and relies on adjacent components.

❒ There are no homogeneous stages, in a sense that each one includes operations that traditionally belong to other stages. For example, throughout key entry procedure, edit checks are performed and the individual processing units are defined by an automatic record linkage with the population register. Both tasks are traditionally editing operations, but, in this instance, they support the key entry stage.

❒ In addition to the intertwined stages, the population register of Israel has been added to serve as an external support throughout the process.



- ❒ The definition of editing during the data capture process is altered altogether; the conventional editing tasks in which data are changed, deleted or added are postponed and take place in the central computer, on a macro level. Editing within the data capture process consists of three assignments:

  - ◆ Correcting enumerators' errors, in order to define those processing units that have not been defined automatically.
  - ◆ Corroborating or correcting values in respondents' fields, when captured values fall out of an expected range or when they produce logical contradictions; and
  - ◆ Coding residual, open categories, where answers are in not optically recognized, alphabetic characters.

The description of the data capture system is beyond the scope of this paper; however, a description of **the flow of data** throughout the process is needed. The first step in transferring the written paper into a magnetic medium is done in the **scanning module**. It is not a sheer scanning procedure but also includes Form Drop Off (FDO), where all preprinted non-unique texts, numbers and graphics are removed from the image of the questionnaire; Optical Character Recognition (OCR), which is performed on marks and numbers written on the questionnaire, compression of the data and inserting it into the database. At this point, the paper questionnaire is repacked and stored.

In the **key-entry module**, the system looks for two sources of identification in order to confirm a field's value. The four available sources in the order of application are: OCR, the Population Register (PR), 1st key-entry, and 2nd key-entry. The Israeli PR contains data on nearly all Israeli citizens and permanent residents (over 95 percent of the population). Here it serves to corroborate the OCR suggested values of variables included in both records, PR and census. By so doing, human intervention is reduced significantly, since even doubtful values that are usually directed to a relatively intensive key-entry process can skip this step altogether. It also serves as an independent source for defining an individual record; once a census record is linked to the PR, it is considered as a defined reference record. This operation is taking place throughout the data capture process: in the key-entry and editing stages, before each round within each stage, and after it (key-entry is performed in, at most, two rounds while editing can be performed in up to four rounds).

The **editing module** deals with non-conventional tasks, while it omits the traditionally defined editing tasks. The editors correct errors in enumerators' fields in order to complete the definition of processing units that have not been defined automatically. They also correct or validate values assigned in previous stages. It should be noted that no editing, in its traditional form, is performed. As was explained at the beginning of the paper, respondents' answers are not changed, even if logical contradictions are embedded in them.

A parallel activity to the editing step is **coding**. There are three types of coding in the system: addresses, occupation, and economic activity. The last two are followed by a quality control process.

At the end of the data entry process, an ASCII file is formed and sent to the main frame, and the images of the questionnaires of each enumeration area, along with the ASCII values of the fields, are saved on an optical disk.

Concluding Remarks

The data capture system of the 1995 Israeli census started in February 1996 and will be completed in August 1996. Although data capture is still in process, there are two important points to note:

- The questionnaires of 1.6 million households are captured by 123 workers. They are expected to complete the task in less than 140 working days.
- The current error rate in the raw data file, as measured by the out-going quality checks, is 0.558 percent. The permissible error rate is 1 percent.

The reduction of human involvement in the data capture process and the high quality of the raw data file are the outcome of careful planning, using the advantages offered by the technological improvements, avoiding or solving problems that have plagued the traditional data capture process, and anticipating -- and, therefore, giving solutions for -- the new system-related problems. ■

5

Chapter

Editing and Imputation Research for the 2001 Census in the United Kingdom

Jan Thomas and David Thorogood,
Office for National Statistics, United Kingdom

Abstract

The United Kingdom Census Offices are working on a development programme for the 2001Census. This paper outlines the research being undertaken on editing and imputation as part of this programme.

New methodological and technical developments are being investigated to see if they offer improvements over previous systems. These include the possible application of neural computing to imputation, and the use of generalised editing software to create editing rules. Research will also address the extent to which editing and imputation processes can be integrated to reduce the occurrence of inconsistencies caused by imputation. The impact on the editing and imputation systems of possible changes to other aspects of the census procedure will also be considered.

Editing and Imputation Research for the 2001 Census in the United Kingdom

Jan Thomas and David Thorogood,
Office for National Statistics, United Kingdom

Introduction

The three UK Census Offices are working on a development programme for the next decennial population census in 2001. The Census Offices are:

❒ the Office for National Statistics (ONS) in England and Wales;
❒ General Register Office for Scotland; and
❒ Northern Ireland Statistics and Research Agency.

The same editing and imputation systems will be used in both England and Wales, and Scotland. Northern Ireland may adopt different processing systems.

The census has traditionally used a system of enumerator delivery and collection of forms to/from households. The possibility of asking respondents to return completed forms by post will be examined in a census test in 1997. Changes such as these may mean that different demands are made on the editing and imputation systems.

As in any data collection exercise census data will contain errors. "Tidying up" the data helps to check the validity of the entry, and ease computer processing. Editing is performed on the raw data which is received from the public. This will contain missing answers, answers which are inconsistent with others on the form, or coded answers which are outside of a pre-defined range.

Imputation aims to fill in gaps in data caused by missing answers and items rejected as invalid or inconsistent by edit checks.

The roles of editing and imputation systems are closely linked. In the 1991 Census, certain imputed items were inconsistent. Checks will be incorporated into the imputation process for 2001 to avoid this. Re-editing the data after imputation is problematic as this may lead to "looping" with data repeatedly failing the post-imputation edit. The option of closely integrating edit and imputation processing will be investigated as this may offer an efficient way of ensuring consistency.

Editing

Policy

Editing should be considered an integral part of the data collection process. In addition to the role of fixing errors, editing can also play a valuable part in gathering intelligence about the census process.

The overall editing (and imputation) policy is to make the minimum number of changes to the database, whilst ensuring that it is complete and error free (as defined).

A review has taken place of the statistical and operational requirements for editing systems and processes for 2001. The main findings were:

- the methods chosen need to be practical and statistically sound. Editing must not cause bias or distortion in the data;
- the methods must allow pre-determined Data Quality levels to be met;
- the processes need to provide a complete, consistent, comprehensive, valid dataset.

There are a number of stages at which editing can take place within the overall process. These are:

- clerical editing by manual scrutiny of the forms;
- at the Data Capture stage (in the processing office), with simple stand alone checks built into the Data Capture software;
- at the Coding stage, when certain decisions may need to be made, for example, preferences where two answers are given to a question but only one is allowed;
- a Post Capture/Coding main edit process to carry out checks within records and between records, and ensure consistency of the database; and
- within or post-imputation checking to ensure that the imputation process has not created inconsistencies.

In summary, errors are:

- ***invalid (out of range)***: relatively simple for any data capture system to spot;
- ***missing data***: i.e., no answer given. A missing code needs to be supplied to identify such cases, again at data capture; and
- ***inconsistent data***: i.e., answers to questions that conflict with one other. These can be:

  definite such as a 1 year old married person;
  less definite such as a 15 year old married person; or
  doubtful such as a 60 year old student.

Inconsistencies can occur:

- **within records** (of the type described above);
- **between person records in the same household** -- for example, relationship to person 1 ticked as husband or wife, but person 1 having ticked single; or

- ❐ **between households** -- for example, if there are two households within one building, one ticks use of bath as shared, and the other ticks use of bath as exclusive.

Editing Options

The edit options being considered for 2001 fall under the following main headings.

1991 "Edit Matrices" Approach

In the 1991 Census, the edit system checked the validity of data and performed sequence and structure checks. Invalid, missing and inconsistent items were identified for the imputation process. The editing process filled in a few missing items. The edit matrices were constructed so as to consider every possible combination of values for relevant items and to give the action, (if any) required should that combination arise, by making the least number of changes.

Simplified 1991 Edit Matrices

An assessment of the 1991 approach is underway to identify areas of excessive complexity that can be simplified. In 1991, it was found that missing items are usually genuinely missing and so could go straight to imputation (providing a quality check takes place to ensure that there are no quality assurance problems, for example, with software). Only inconsistencies need to be handled at editing stage.

Stand Alone Edits at Data Capture

Within record checks are being defined which could be carried out at the data capture stage. There is a school of thought which says that the sooner errors are detected and eliminated the better. However, it may prove more effective to eliminate all inconsistencies as one main process.

Interactive Editing

Simple logic and data validation edit checks could take place via clerical intervention, using software packages to load the main database with correct values.

Fastpath Editing

After capture editing could be carried out on "closed" questions (those covered by tick box answers) only. This would produce an earlier partial database. The "hard to code" questions such as occupation could follow later. Such an approach would only be adopted if there was a clear customer requirement for information on certain "easy to process" questions.

Selective Editing

Selective editing prioritises which fields should be edited and then applies edits to those priority fields only.

The selective editing approach is to calculate a score for each field with one or more detected errors. If the score is low it is expected that correction would have little impact on the resulting edits.

Generalised Editing Rules

Specific software, designed and developed to generate editing rules is available. Generalised software systems have great advantages when compared with *ad hoc* applications: they obviously reduce application costs, but, more importantly, they allow the correct application of given methodologies to each suitable situation. The system considers the edits for all fields simultaneously and the response reliability of each field is assessed. The explicit edits needed to edit combinations of these fields are then automatically generated, even if the relationships between the fields are very complex. Work is planned to investigate the use of these generalised systems.

Editing and Imputation Combined

Imputation needs to take place on a consistent database, but can itself cause inconsistencies. If, as a result of new approaches, fewer errors have to be dealt with at main edit time, then imputation and editing could be carried out in conjunction, with a final consistency check at the end of processing.

Imputation

Policy

The Census Offices have access to all available information at individual record level, and so are best placed to guide the imputation of missing data. The imputation system removes the need for "missing" or "not-stated" categories in statistical outputs, which can be inconsistently used and interpreted by users. It is therefore accepted that some form of imputation must be undertaken by the Census Offices prior to the production of outputs.

Imputation Options

Three options are being considered. These are:

- ❒ Donor imputation (primarily hot deck systems as in the 1991 Census);
- ❒ Neural networks; and
- ❒ Multi-level modelling (MLM).

Of these, a hot deck system or a neural network solution are the most likely to be adopted. However, the MLM approach is being investigated further to see if it can be used, in whole or in part. The boundaries between the various types of system are not always clear, particularly those between different forms of donor imputation.

It is intended that the imputation options will be trialed using a common set of test data from the 1991 Census. This should assist with the comparison of results.

It is possible that different variables might be imputed using systems of different types and/or differing levels of sophistication. For example, a complex hot deck or neural network system might be used to impute key variables (age, sex, marital condition), with other variables imputed using a simpler system.

Some imputation can be carried out by the editing system itself, in cases where only one code is possible. For example, the marital status of a one year old person can only be single.

Donor Imputation

In donor imputation methods a value is selected from a valid record (the donor) and copied to fill in the missing item(s) of another record (the recipient). Donor methods offer the benefit of imputing plausible values as they are copied from real records. However, these values are not always consistent with other parts of the recipient record. Differences between types of donor method centre mainly on how the donor is selected. These methods are outlined below. Although the most likely donor option for 2001 is a sequential hot deck, aspects of the other methods below might be adopted.

Sequential Hot Deck

The imputation system used in 1991 was based on the hot deck method developed by Felligi and Holt (1976).

A series of tables were designed reflecting the relationship between the variable to be imputed and other census variables. For example, it was known from the 1981 Census and intercensal tests that a good indication of the number of cars available to a household could be found from housing tenure, the number of people in the household, and whether the accommodation was in a permanent building or not. The imputation table for "number of cars" therefore held the observed values for number of cars available to households with all combinations of these reference variables. For example, if the number of cars was missing for a particular household, the most recently processed record with the same tenure, building type and number of persons as the recipient, was selected as the donor. The number of cars available to the donor household was copied to the recipient.

Fifty separate imputation tables (or **decks**) were used: 13 for household items; 24 for persons in households; one for communal establishments; and 12 for persons enumerated in communal establishments. For each cell in the imputation table, a series of values were held. These were updated continuously, with new values being taken from the most recently processed wholly valid record, and the oldest in that series of values being discarded. When an item required imputation, the newest value in the appropriate cell series was copied to take the place of the missing value. However, if this value had already been used to impute, the next oldest was used. For household variables, such as tenure and accommodation type, each cell held a series of 3 values, whereas cells in tables used to impute individual variables, such as a person's age and sex, held 6 values.

This hot deck method is known to work. It makes efficient use of computer processing capacity as each data file is read once only, although there is an additional storage requirement to hold data in both the imputation deck and main data file.

However, hot deck systems can be complicated and time consuming to program. This is closely related to the number and complexity of the imputation decks which are included in the system.

If more than one case in succession contains missing items, certain donor values may be used several times. The likelihood of this is reduced by storing several donor values in each class. Where there are many classes, the likelihood of having to re-use donor values increases, particularly where there is much missing data.

Simplified Hot Deck

This is essentially the 1991 system reduced in complexity, with fewer imputation decks and cells. There could be significant reductions in the time and resources need for system development, and there may also be reductions in the computer processing and storage resources required.

As simpler imputation classes are used, each cell should tend to be updated more frequently. There is therefore less chance of re-using donors. This can be seen as a trade-off: as classes are simplified, donors may be less similar to the case to be imputed but reuse of donors is less likely.

Hierarchical Hot Deck

Here, the data file is sorted into a much larger number of detailed imputation classes in a hierarchical structure. If no suitable donor is found at the finest level of the classification, classes can be collapsed into broader groups until a donor is found. A pattern of "hard" and "soft" class boundaries can be programmed into the hierarchical structure, e.g., to ensure that an item is always imputed from a donor of the same age group, even though the area of residence classes may be collapsed.

Hierarchical hot deck imputation frequently allows items to be imputed from very similar cases. However, the method is less efficient in its use of computing resources compared with sequential hot deck imputation. The system development process is difficult and time consuming, with possibly little benefit. For these reasons it is unlikely that the method would be of direct use for 2001. Aspects of the method, such as collapsible class boundaries for certain key or hard to impute variables, could be of possible use.

Statistics Canada -- New Imputation Methodology (NIM)

The NIM system is being investigated to see if aspects of this could be used. NIM allows forward searching within a data file to select a suitable donor record. This differs from the hot deck system used in the UK in 1991 which could only select donors from already processed records. This system may offer a way of avoiding re-use of donor values in areas where there is sparse data. NIM may also offer more plausible imputations.

Neural Networks

Unlike traditional computing approaches which need to be explicitly programmed, neural computers automatically learn solutions from the data itself. A neural computer can be taught and can learn about personal and household profiles provided in the census data in order to impute missing values.

Initially, the neural computer will go through this learning process, commonly known as "training." By using analysis tools, a model which has learnt profiles from the data may be analysed to show the relationships it has learnt.

Neural computing can impute by forming a model using examples showing how the imputed variable is related to the other variables, and then applying this model to cases where a value for the imputed variable is not known. One model is constructed for each variable to be imputed, where the model takes the form of a function that takes the known data as an input, and delivers the imputed value as an output.

The output is the most likely value based on the relationship of the imputed variable to all other variables. It differs from traditional modelling approaches in that the output is initially in the form of a probability distribution. This is the distribution which would be achieved if successive imputations on identical records were carried out. The selection of the imputed value is made by the use of a random number generator weighted by the probability distribution.

An initial "proof of concept" trial was very successful, and a further trial has begun. This second trial will address operational aspects as well as further examining the statistical quality of outputs. It is claimed that neural computing is reliable and cost effective. Such a system would need far less large scale development and programming work. However, there will be a need to carefully integrate the neural network with other parts of the processing system.

Other problems relate to difficulties in understanding the concepts behind the system and explaining these to users of census data. There would also need to be a considerable transfer of skills from neural technicians to Census Office staff to allow the efficient development and operation of the neural network and interfacing systems.

Multi-level Modelling Imputation

This model-based imputation method is being researched by Dr. R. D. Wiggins of City University, London. ONS have supplied anonymised test data from 1991 to assist in this project.

In multiple imputation (Rubin, 1987), missing data are replaced by two or more feasible values which represent the distribution of possible values. The most plausible imputation value is then selected from this distribution. The multi-level modelling work being undertaken by Wiggins is a complicated statistical technique which aims to go beyond existing multiple imputation. The key advantage is that the method does not require the generation of multiple data sets. It is aimed to produce consistent and efficient estimates together with their standard errors. The method can be used for both continuous and categorical variables. Multi-level modelling may require extensive expert intervention during processing and may therefore be unsuitable for the census operation.

Key Considerations

Statistical Quality

The great majority of Census records do not contain missing data and pass the various edit checks which are applied to them. However, the level of missing data varies from one small area of the country to another, and also from one population sub-group to another. For example, the level of missing data is higher in inner cities and among the elderly population.

For certain geographical areas and variables, it is likely that only sparse data will be available. The imputation methodology selected will need to cope adequately with this, making full use of what data are available.

System Development and Operation

In addition to the statistical quality of the output, the probable development and operational requirements in terms of time, staff and computing of each option will be considered. The census database is large and the data needs to be processed quickly. It is important that the edit and imputation methods selected allow efficient use of computer storage and processing capacity. This consideration may become less important as computing technology develops over the next 5 years, but nevertheless it cannot be ignored.

Operation of the edit and imputation systems must be largely automatic with little need for expert intervention during processing. This means that more complex interactive modelling approaches to imputation will be unsuitable, even though they may produce output of satisfactory quality.

It is planned to set up a data quality management system (DQMS) for the 2001 Census. The editing and imputation systems will need to output to the DQMS readily understandable information on their processes, by different variables, areas, and population sub-groups.

Edited and imputed items were not marked in 1991 as it was felt that this would increase the size of the database too much. However, there is a duty to report on the quality of data and, as such, the recording of edit/imputation actions is an essential requirement for 2001. Some options (such as neural networks) offer this automatically but, even for those that do not, developments in computer storage and processing power mean that this should be less of a problem in 2001.

Processing Order and Geography

The editing and imputation systems need to fit into the overall processing strategy which is currently being defined. In previous censuses, records have been processed "in batch" with all records from a geographical area processed together. This can cause bottlenecks in the processing, and there is ongoing work to decide whether "free-flow" processing could be adopted, with census forms processed on receipt. It is likely that imputation would have to wait until sufficient forms from an area had been returned to allow the creation of an accurate imputation model or to ensure that sufficient geographically close records were available as donors. This is irrespective of the actual methodology used. The characteristics of people returning forms early are likely to differ from those returning them late. Accurate imputation, whatever method is selected, will need to reflect this if a free-flow approach were to be adopted.

Geographical variations in the relationships between variables must be reflected by the imputation method. In 1991, this was allowed for by processing groups of contiguous small areas together. It would still be necessary to do this if an alternative donor method were used. Similarly, a model-based approach would need to allow for changes in the model to reflect geographical differences.

Absent Households/One Number Census

In 1991, records were imputed for households that were wholly absent at the time of the Census, and for other identified households which did not return a Census form. This requirement will remain in 2001. The imputation system selected must be able to cope with the imputation of these missing records as well as missing data items. It may be necessary or preferable to use different methods for these essentially different tasks.

A related development for 2001 could be the adoption of the One Number Census (ONC) approach to disseminating census and validation data. A ONC approach means only adjusted (for coverage) census results are output, as opposed to census tables and separate coverage correction factors. One way of doing this might be to impute missing person and household records of a number and type indicated by the validation estimates. These are people and households **not** identified by the enumerator, but estimated to exist by the use of alternative sources (such as administrative registers) or a follow-up survey. This would represent a significant increase in the role of the imputation system. It has not yet been decided if a ONC approach to dissemination will be adopted.

Disclosure Control

There are proposals that some form of additional imputation should be undertaken as a disclosure control technique. By deleting and re-imputing valid records, additional uncertainty of identification and matching is introduced which may reduce the need for other disclosure control techniques to be used. Although there are problems with this approach to disclosure control, this use of imputation will be considered.

Reference

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

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys, New York, Wiley. ■

5

Chapter

A Priority Index for Macro-Editing the Netherlands Foreign Trade Survey

Frank van de Pol and Bert Diederen,
Statistics Netherlands

Abstract

A macro-editing index for the Netherlands Foreign Trade Survey is described. This index is intended to trace errors by merely confronting current publication figures with the publication cell's history. A provisional experiment was carried out to determine the index's power in pointing to errors in the data. Information on pseudo errors was obtained from the present record-oriented editing process. Despite the big variability in many of the time series concerned, first results show a good association between index values and pseudo errors.

A Priority Index for Macro-Editing the Netherlands Foreign Trade Survey

Frank van de Pol and Bert Diederen,
Statistics Netherlands

Introduction

The Netherlands Foreign Trade Survey (NFTS) used to be a complete enumeration of all trade above a certain threshold, until in 1993 the EC put an end to restrictions on trade within the EC. Because of this, administrative customs data on the Netherlands trade with EC countries were replaced by survey data with the usual low response rates, ranging from 44 percent for small firms to 85 percent for large firms. Therefore, a shift of attention to increasing the response rate is necessary. Resources for this will be found in a complete redesign of the NFTS, especially in more efficient data editing.

Editing the 2 million transactions that come in each month, mostly via electronic data interchange, will be reduced to the bare minimum of valid value checks. The main instrument to trace errors will be macro-editing. An index is designed to prioritize inspection of the 58,000 nonempty publication cells. Only publication cells that deviate clearly from the value we expect from the past will be checked for errors.

A special feature of the data that shapes the priority index is that these many detailed trade time series contain lots of zero observations. Therefore our index is built on two distances, one relating the observed value to zero and the other relating the observed value to the expected non-zero value. The probability to observe a zero is also taken into account. A section on the data gives some more detail about the data. The next section on macro editing contains the formula we chose, and the section on tracing errors by macro editing gives results on comparisons of edited and unedited data. The final section gives the conclusions.

The Data

According to EC regulations, firms have to use a very detailed classification of goods and countries. Publication of monthly imports and exports in this amount of detail for the Netherlands turns out to give quite irregular figures for most publication cells. Berends et al. (1995) recently decided to publish monthly data on a somewhat more aggregated level, with about 50 country groups and about 800 product categories. When data on value and weight are aggregated to publication cells, the data matrix as displayed in Table 1 is obtained.

Table 1.--Notation for the NFTS Data; Import in Month *t*

	country 1, product 1	country 1, product 2	country c, product p	country C, product P	unit-total
unit 1	v_{111} q_{111}				v_{1++} q_{1++}
unit 2					
unit *u*	v_{u11} q_{u11}		v_{ucp} q_{ucp}		v_{u++} q_{u++}
unit *U*					
publication cell total			v_{+cp} q_{+cp}		

Rows are firm units and columns are publication cells. For firm unit *u* the import datamatrix holds value v_{ucp} and quantity q_{ucp} of product *p* from country *c* in month *t*. For exports, there is another matrix of the same format. In fact, both matrices are split in two submatrices, one for EC trade and one for non-EC trade. Although the number of publication cells has been reduced, these data matrices still hold hundreds of thousands of non-empty cells, as Table 2 shows.

Table 2.--The Size of the Datamatrix of the Netherlands FTS

	EC imports	EC exports	Non-EC imports	Non-EC exports
Country groupings, *C*	10	10	44	44
Product categories, *P*	798	797	796	796
Nonempty publication cells	6,681	7,538	17,688	26,088
Firm units, *U*	25,694	20,184	? 2,000	1,099
Nonempty matrix cells	347,356	273,109	? 100,000	83,598

?: Educated guess.

We want to use macro-editing principles to trace errors in these many publication cells (Granquist, 1994, 1995). The editors will look for the current publication figures that deviate most from what is expected, and look for errors in those columns of the datamatrix only. In the following we will describe the index we use and present some first results on its ability to trace errors.

A Macro-Editing Index Which Takes Zero Observations into Account

An index was devised to quantify the deviation between current and expected publication values. It should be approximately uniformly distributed between 0 and 100 in order to enable an interpretation in terms of a percentage of the cells. Moreover, we have to take account of the fact that one third of the publication cells that are nonempty for yearly figures, will remain empty with monthly figures.

The same index will be applied to value and quantity, for imports and for exports. Therefore we simplify notation to x_{tuc}, with t for month, u for firm unit and c for publication cell. A publication total is written as x_{t+c}. An unedited value is written with a prime, as x'_{tuc}.

An unedited publication value x_{t+c} might be in error, when it differs a lot from the value we expect, $\hat{x}_{t+c}$, on the basis of exponential smoothing of the cell c history (Michels, 1996; Siver and Peterson, 1985). However, with a very stable time series, differences are sooner suspect than with a variable one. Therefore we standardized the difference with the standard deviation of the time series concerned, $s_{x_{tc}}$, thus obtaining as distance measure

$$d_{tc} = \left|x'_{tc} - \hat{x}_{tc}\right| / s_{x_{tc}}.$$

Next, we observed that many time series have zero observations in some months, which greatly boosts the standard deviation and thus makes the d_{tc} less sensitive for an outlying current observation in zero-ridden time series. As a consequence, too small, but non-zero observations will not be noticed with this distance measure. To avoid this, time series are considered to have two regimes, zero observations and non-zero observations. An observation should be compared with the zero and the non-zero regime of the time series. We use on the one hand the distance measure between the observed value, x'_{tc}, and the predicted non-zero value, $\hat{x}_{tc} | x_{tc} \neq 0$,

$$d^{\varnothing}_{tc} = \frac{x'_{tc} - (\hat{x}_{tc} | x_{tc} \neq 0)}{s_{(x_{tc}|x_{tc}\neq 0)}},$$

and, on the other hand, the distance between the observed value and zero,

$$d^{0}_{tc} = \frac{\left|x'_{tc} - 0\right|}{s_{(x_{tc}|x_{tc}\neq 0)}}.$$

These two distances are combined into a single measure using the probability to observe a zero, p_o, as predicted from the time series' history,

$$M^{*}_{tc} = v^{*}_{tc}\left[\left(1 + d^{0}_{tc}\right)^{\hat{p}_o} \times \left(1 + d^{\varnothing}_{tc}\right)^{(1-\hat{p}_o)} - 1\right],$$

with value $v^*_{tc} = \max(v'_{tc}, \hat{v}_{tc})$. This M^*_{tc} measure has a minimum of 0 and no maximum. Its distribution is left skewed and has a long tail to the right. To obtain a more evenly distributed measure between 0 and 100 we transform it with

$$M_{tc} = \frac{100\ M^*_{tc}}{median\ (M^*_{tc}) + M^*_{tc}} .$$

With this macro editing index some tests were carried out, which will be presented in the next section.

Tracing Errors by Macro-Editing

Before using real data we first did some testing with artificial data. Table 3 shows four time series of six months, each with non-zero mean 100, but with varying amounts of zero observations.

Table 3.--M_{tc} Values for Four Series and Four Current Observations, 0, 50, 100, and 200

Month	Series 1	Series 2	Series 3	Series 4
January	105	105	0	0
February	110	110	110	0
March	95	0	0	0
April	105	100	100	100
May	95	95	0	0
June	90	90	90	0
p_o	0	0.17	0.5	0.83
$\hat{x}_{tc} \vert x_{tc} \neq 0$	100	100	100	100
$s(x_{tc} \vert x_{tc} \neq o)$	7.07	7.07	8.16	10[1]
M_{tc} (July= 0)	66.5	54.8	26.8	6.6
M_{tc} (July= 50)	50.0	50.0	46.2	41.3
M_{tc} (July=100)	0	7.8	26.8	47.4
M_{tc} (July=200)	66.5	66.5	63.2	58.5

[1] With only one non-zero observation $s\,(x_{tc} \vert x_{tc} \neq o)$ was set to $(\hat{x}_{tc} \mid x_{tc} \neq 0)/10$.

As estimator for the predicted non-zero value, the mean of the non-zero values in the time series was used. The probability of a zero was estimated as the proportion zeroes in the time series. In a time series without zeroes and with mean 100, an observation of 200 turns out to be as alarming as an observation of 0 ($m_{tc} = 66.5$). When the series has a zero in it, an observation of 200 still has $m_{tc} = 66.5$, using the distance measures $d^{\otimes}_{tc}$ and d^{o}_{tc}, which treat zero observations separately. A distance measure, which treats zero as an ordinary observation, would, due to a higher standard deviation of the series, wrongly consider 200 as less exceptional than our measure does.

An observation of 100, which is the mean non-zero value of all series considered, gives index M_{tc} = 0 when the time series has no zeroes in it. The more zeroes the time series has, the higher the index value will be for an observation of 100 ($M_{tc} = 7.8$ for $p_o = 0.17$, $M_{tc} = 26.8$ for $p_o = 0.5$ and $M_{tc} = 47.4$ for $p_o = 0.83$).

In order to test the macro-editing index with more realistic data, an arbitrary sample was drawn from the unedited data of August 1995. A limited number of products and countries of varying type was selected from the 44 country groupings and 798 product groupings available. Moreover, only large firms were included in the sample, which holds about 2000 firm units. This test sample was then matched with its history, time series of 24 months length, and with the corrected data after data editing.

In this file we computed pseudo errors, that is the difference of unedited values minus edited values. Our major concern was to test whether the macro editing index M_{tc} would be able to trace publication cells with large pseudo errors. If so, editing of non-EC trade will rely on this index for error detection

Table 4.--Relation Between Macro-Editing Index (Vertical) and Absolute Edit Size (Horizontal); Observed Non-zero Import Values of a Selection of Publication Cells in August 1995

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
$0<M_{tc}\leq 5$	14							13														12
$5<M_{tc}\leq 10$	9									13												8
$10<M_{tc}\leq 15$	6		100				9					10										6
$15<M_{tc}\leq 20$	5			67						13		10										5
$20<M_{tc}\leq 25$	4					8			17	13												4
$25<M_{tc}\leq 30$	5								8						33							4
$30<M_{tc}\leq 35$	2					8																2
$35<M_{tc}\leq 40$	3					8	9															3
$40<M_{tc}\leq 45$	3					8																3
$45<M_{tc}\leq 50$	3					23		13		25												4
$50<M_{tc}\leq 55$	3									25												3
$55<M_{tc}\leq 60$	3											10										3
$60<M_{tc}\leq 65$	3					8																3
$65<M_{tc}\leq 70$	3						9				100			50								3
$70<M_{tc}\leq 75$	3						18	25				10										4
$75<M_{tc}\leq 80$	4						9		17			10			33							5
$80<M_{tc}\leq 85$	3					8	9	13				33										4
$85<M_{tc}\leq 90$	7			33								10				25						6
$90<M_{tc}\leq 95$	6					15	18	13	25			10	50			50		50	50			8
$95<M_{tc}\leq 100$	11					15	18	25	33	13		30	50	50		25	100	50	50	100		14
Total %	85	.		1		2	2	2	2	2	.	2	.	.	1	1	.	.	.	.		100
Total freq:*	465	0	1	3	0	13	11	8	12	8	1	10	2	2	3	4	1	2	2	2	0	550

Labels absolute edit size in Dutch guilders (= $0.68):

0: 0
1: 1- 10
2: 10- 100
3: 100-500
4: 500-1000
5: 1000-5000
6: 5000-10,000
7: 10,000-20,000
8: 20,000-50,000
9: 50,000-70,000
10: 70,000-100,000
11: 100,000-200,000
12: 200,000-500,000
13: 500,000-700,000
14: 700,000-1,000,000
15: 1,000,000-2,000,000
16: 2,000,000-5,000,000
17: 5,000,000-7,000,000
18: 7.000.000-10.000.000
19: 10,000,000-50,000,000
20: ≥50,000,000

from June 1996 onward. EC trade is planned to follow in January 1997. An Oracle database is being built to guide the editors from a suspect publication cell to a suspect firm unit, a cell in Table 1. Underlying transactions can be looked up and corrected.

Table 4 shows a cross-table of non-zero reported import values with the absolute size of the pseudo errors on the horizontal axis and the M_{tc} index of all 550 non-zero publication cells considered on the vertical axis. The export counterpart concerns 772 non-zero publication cells and is shown in Table 5.

Table 5.--Relation Between Macro-Editing Index (Vertical) and Absolute Edit Size (Horizontal); Observed Non-zero Export Values of a Selection of Publication Cells in August 1995

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
$0<M_{tc}\le 5$	16					10	6		4			7										13
$5<M_{tc}\le 10$	7					3	19															6
$10<M_{tc}\le 15$	5							5														4
$15<M_{tc}\le 20$	4							5	12			7	11									4
$20<M_{tc}\le 25$	4					10		5	4													4
$25<M_{tc}\le 30$	5							5														4
$30<M_{tc}\le 35$	2					3		9	4		9				33							2
$35<M_{tc}\le 40$	5					3			4													4
$40<M_{tc}\le 45$	3							5		13		7										3
$45<M_{tc}\le 50$	4					7	6		8			13										4
$50<M_{tc}\le 55$	3						6	5	4			7				13						3
$55<M_{tc}\le 60$	3							5			9											3
$60<M_{tc}\le 65$	3					7	6	5				7										3
$65<M_{tc}\le 70$	4							9			9											3
$70<M_{tc}\le 75$	3					3	13	9	4	38			11									4
$75<M_{tc}\le 80$	3					14	6					7	22									4
$80<M_{tc}\le 85$	3					14	6	5	12	13	18	7	11	33	33							4
$85<M_{tc}\le 90$	6					3	6		4		9	13	11									5
$90<M_{tc}\le 95$	6					7		9	12		9		11				33		50			7
$95<M_{tc}\le 100$	11				100	14	25	23	31	38	36	27	22	67	33	88	67		50	100	100	15
Total %	79				.	4	2	3	3	1	1	2	1	.	.	1	.		.	.	.	100
Total freq.	612	0	0	0	10	29	16	22	26	8	11	15	9	3	3	8	3	0	2	3	1	772

Labels absolute edit size in Dutch guilders (≈ $0.68):

0: 0	6: 5000-10,000	12: 200,000-500,000	18: 7,000,000-10,000,000
1: 1- 10	7: 10,000-20,000	13: 500,000-700,000	19: 10,000,000-50,000,000
2: 10-100	8: 20,000-50,000	14: 700,000-1,000,000	20: ≥50,000,000
3: 100-500	9: 50,000-70,000	15: 1,000,000-2,000,000	
4: 500-1000	10: 70,000-100,000	16: 2,000,000-5,000,000	
5: 1000-5000	11: 100,000-200,000	17: 5,000,000-7,000,000	

We first focus on the frequency distribution of the edits that were made in the present, transaction-record oriented approach. It turns out that in most of the publication cells no edits have been done, 85 percent for imports and 79 percent for exports. This is the column labeled '0' in tables 4 and 5. Few cells have an accumulated effect of less than dfl.1000, which probably means that no very small corrections are carried out. Most edited cells, about 12 percent for import values and about 18 percent for output values, have undergone medium sized alterations, that is less than half a million guilders.

About one quarter of these medium sized edit effects are predicted by a macro editing index larger than 90. When editors would be looking at cells with an index value larger than 80 percent only, about 50 percent of the medium sized edits will take place. For high impact edits, causing publication cell changes of more than half a million, the "hit rate" comes close to 100 percent.

Tables 6 and 7, finally, refer to those publication cells that firm units reported to be zero in August 1995. Few zero observations are altered in the present editing process, in our sample 3 percent for import values and 1 percent for export values. These few cases are well predicted by the macro editing index.

Table 6.--Relation Between Macro-Editing Index (Vertical) and Edit Absolute Size (Horizontal); Observed Zero Import Values of a Selection of Publication Cells in August 1995

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
$0<M_{tc}\leq 10$	24																					23
$10<M_{tc}\leq 20$	7																					7
$20<M_{tc}\leq 30$	8																					8
$30<M_{tc}\leq 40$	5																					5
$40<M_{tc}\leq 50$	6																					6
$50<M_{tc}\leq 60$	5																					5
$60<M_{tc}\leq 70$	5																					5
$70<M_{tc}\leq 80$	7						100															7
$80<M_{tc}\leq 90$	11					50																11
$90<M_{tc}\leq 95$	8				100																	8
$95<M_{tc}\leq 100$	13					50								100				100	100	100		14
Total %	97				1/2	1	1/2							1/2				1/2	1/2	1/2		100
Total freq.	262	0	0	0	1	2	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	270

Labels absolute edit size in Dutch guilders (= $0.68):

0: 0
1: 1- 10
2: 10-100
3: 100-500
4: 500-1000
5: 1000-5000

6: 5000-10,000
7: 10,000-20,000
8: 20,000-50,000
9: 50,000-70,000
10: 70,000-100,000
11: 100,000-200,000

12: 200,000-500,000
13: 500,000-700,000
14: 700,000-1,000,000
15: 1,000,000-2,000,000
16: 2,000,000-5,000,000
17: 5,000,000-7,000,000

18: 7,000,000-10,000,000
19: 10,000,000-50,000,000
20: ≥50,000,000

Table 7.--Relation Between Macro-Editing Index (Vertical) and Edit Absolute Size (Horizontal); Observed Zero Export Values of a Selection of Publication Cells in August 1995

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
$0<M_{tc}\leq 10$	18																					17
$10<M_{tc}\leq 20$	11																					11
$20<M_{tc}\leq 30$	8																					8
$30<M_{tc}\leq 40$	7																					7
$40<M_{tc}\leq 50$	6																					6
$50<M_{tc}\leq 60$	5																					5
$60<M_{tc}\leq 70$	8																					8
$70<M_{tc}\leq 80$	8																					8
$80<M_{tc}\leq 90$	7																					7
$90<M_{tc}\leq 95$	8																					8
$95<M_{tc}\leq 100$	14												100				100		100			15
Total %	99												1/2				1		1/2			100
Total freq.	338	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2	0	1	0	0	342

Labels absolute edit size in Dutch guilders (= $0.68):

0: 0	6: 5000-10,000	12: 200,000-500,000	18: 7,000,000-10,000,000
1: 1- 10	7: 10,000-20,000	13: 500,000-700,000	19: 10,000,000-50,000,000
2: 10-100	8: 20,000-50,000	14: 700,000-1,000,000	20: ≥50,000,000
3: 100-500	9: 50,000-70,000	15: 1,000,000-2,000,000	
4: 500-1000	10: 70,000-100,000	16: 2,000,000-5,000,000	
5: 1000-5000	11: 100,000-200,000	17: 5,000,000-7,000,000	

Concluding Remarks

The results in the previous section showed that our macro-editing index will trace almost all large errors in the foreign trade data, but a sizable proportion of the mid-sized errors will remain undetected. Although most of these errors are too small to have a serious effect on publication figures, we still would like to boost the performance of our index with the larger ones.

One improvement will be that we will use a better time series model to predict the current value. In this preliminary experiment we simply used the arithmetic mean of the non-zero values in the previous 24 months. First results with exponential smoothing (Silver and Peterson, 1985; Michels, 1996) promise to be better.

Secondly, a robust estimate for the variability of the series would make the index more sensitive with series which behave relatively stable, but have one extreme outlier in their history.

Finally, we consider an alternative for the distance measures we presently use. Instead of dividing the absolute difference between an observed and a predicted value by a measure of the variability of the series, one could subtract 1.96 times the standard error of the predicted value,

$$d_{tc}^{\varnothing *} = \max\left\lfloor 0 , \left|x'_{tc} - (\hat{x}_{tc}|x_{tc}\neq 0)\right| - 1.96 s_{(\hat{x}_{tc}|x_{tc}\neq 0)} \right\rfloor .$$

With this measure, macro editing would only look at significant deviations from the expected value. The main problem with this measure is that an estimate of the standard error of the predicted value is not available for the case of exponential smoothing. Moreover, this measure will be 0 for all publication cells with nonsignificant deviations. This property is at variance with our wish to have a measure which has a nearly uniform distribution.

References

Berends, M. L. Visser; R. Janssen; and G. Slootbeek en N. Nieuwenbroek (1995). Onderzoek bij de Statistiek van de Internationale Handel (Eindrapport), Centraal Bureau voor de Statistiek, Heerlen.

Granquist, L. (1994). Macro-Editing -- A Review of Methods for Rationalizing the Editing of Survey Data, United Nations Statistical Commission and Economic Commission for Europe: Statistical Data Editing, vol. 1, *Methods and Techniques*, United Nations, Geneva, pp. 111-126.

Granquist, L. (1995). Improving the Traditional Editing Process. In Cox, Binder, Chinnappa, Colledge, Knott, (Eds.), *Business Survey Methods,* John Wiley, New York, pp. 385-401.

Michels, P. (1996). Voorspellingen Met Effeningsmethoden voor Controle/Correctie bij de Internationale Handel, Internal Research Note, Statistics Netherlands, Heerlen.

Siver, E. and R. Peterson (1985). Decision Systems for Inventory Management and Production Planning, second edition, Wiley, New York. ■