

STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 10
English only

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Prague, Czech Republic, 14-17 October 1997)

Item 3 of the provisional agenda

NATIONAL REPORT: UNITED KINGDOM

Submitted by the Office for National Statistics, United Kingdom

¹ Prepared by Jan Thomas, Paul Smith and Mohammed Yar.

I. INTRODUCTION

1. This report contains a summary of the research being carried out into editing imputation and data quality by two divisions of the Office for National Statistics (ONS) in the UK. Census Division are researching into the methodology for the next national census of population, which will take place in 2001. Methods and Quality Division have the responsibility for providing methodological advice to other areas in the ONS, and its Business Survey team provide consultancy services on statistical methodology to the Business Statistics group in the ONS.

II. THE CENSUS IN THE UK

2. The UK Census Offices are working on a development programme for the next decennial population census in 2001. This paper outlines the research being undertaken on editing and imputation as part of that programme, and gives some details about the Census Test, which has taken place this year, to try out new methodological options with the aim of improving coverage, quality and cost-effectiveness.

The 1997 Census Test

3. The new options which have been tested include a postal method of collection of census forms. If this test is successful it will allow us to target resources where they are most needed in the fieldwork. In statistical terms, the 1997 Census Test is a designed experiment. The statistical analysis is primarily concerned with the evaluation of the key drivers: enumeration methodology, form style and form content. The design of the census questionnaire has been improved to make the form clearer and easier for the public to complete. In addition to the key drivers, automated data capture and coding systems and a new method of planning enumeration areas, with the introduction of customised maps, the use of pre-printed address lists and updated information from local authorities are the other major new procedures which have been tested.

4. Processing of the data and the test evaluation are currently underway, but it is planned that the test evaluation will be complete by March 1998.

Data Capture and Coding

5. One of the primary activities of the 1997 Census Test is to trial a prototype data capture and coding system. The key elements being tried for the first time include automatic scanning and recognition of forms completed by the public, capturing and automatically coding the data with interactive coding query resolution. This has involved in-house development as well as the integration of a number of software packages. The Test processing operation will also be controlled by an automatic processing control system working at census form level. The approach for 1997 has been designed after carrying out extensive research and a series of mini-trials over the past 2 years. The outcome of the 1997 Census Test will help to determine the processing strategy for the 2001 Census, and to decide the approach for the procurement of a fully integrated processing system, with the opportunity to refine the chosen system in a major dress rehearsal in 1999.

Editing

6. Given that we will adopt the Felligi-Holt principles, then the choice of an edit system is a practical issue, as the system needs to be compatible with the overall data processing information systems and output strategies. The UK Census Offices are currently considering the relative merits of single stage versus multi-stage implementation for the between questions edits, to guide the development of an edit policy.

7. A disadvantage of single stage data editing is timeliness as it requires data on all the variables simultaneously, including the data on hard to code questions, which takes much longer to become available. Another disadvantage of single stage editing is potential complexity of the edit system. All the variables, ranging from relationships within a household to occupation and industry, would need to be considered simultaneously and this would greatly complicate the edit process.

8. However, single stage editing has the potential to preserve the structure of the data (joint distributions or relationships among variables) better than multistage editing because you are using all the available information.

Imputation

9. The Neural Network Imputation Trial - A separate paper has been submitted to this work session fully describing the results of the trial of neural imputation which was concluded this year. The neural network imputation method largely met the criteria for evaluating operational performance. However on the statistical side, it failed to demonstrate that it was superior to the hot-deck method which was used in the 1991 Census, (although the hot deck results were not very good either).

10. Despite these disappointing results, the Census Offices consider the research into neural imputation worth pursuing, and are considering a separate research programme, in conjunction with academics from the University of Southampton, to look at refining the approach for possible application within the Office for National Statistics, or for the 2011 Census.

11. **Donor Imputation** - We are considering a donor system in which there is only one donor record for all missing values in a recipient record. This will, in principle, improve the preservation of the marginal and joint distributions. The donor approach can be improved by the use of advanced statistical techniques such as CHAID (a regression-tree based binary segmentation method) or multinomial logistic modelling in the selection of the matching variables. The hot-deck method only searched for donors among records in one direction. It may well be that a more suitable donor can be found from the records next to the recipient in the other direction. It is also proposed to use some kind of statistical distance function to measure the closeness of the recipient and donor households to ensure the selection of the most suitable donor. These can be 'tuned' to reflect both 'accurate individual imputations' and 'accurate distributional imputations'.

12. **Multinomial logistic regression** - For the purposes of providing a benchmark to the other processes we are looking at this approach. Using a subset of 1991 census data, (the same set that was used in the neural trial) we are using multinomial logistic regression

techniques to research whether such techniques could be practicably used to impute missing census variables. Models are being created using the Catmod procedure within the SAS system as this is the only procedure which accommodates multivariate, nominal variables. Just six variables are being trailed as response variables at the present time, while all other census variables are being considered as explanatory variables.

Data Quality

13. A Data Quality Management Programme has been put in place to co-ordinate the data quality aspects of the 2001 Census. The programme aims to identify the most important sources of error in the data, to provide advice on any corrective action that can be taken and to form the basis of commentary to users to accompany census output statistics.

14. A paper describing progress of the development of the Data Quality Monitoring System (DQMS) for the 2001 Census, is described in a separate paper to this conference.

III. METHODS AND QUALITY

Editing and Valuation in Business Surveys

15. The last year has seen greater awareness of editing and detection and treatment of unusual observations in ONS business surveys. There have been several main areas where work has been taken forward.

16. **Winsorisation** - The application of winsorisation in business surveys has benefited from further experience and of using the method with real data. The experiences of using this method in practice for surveys and of applying the methods in the context of editing sampling frames are the subject of a separate paper at this work session.

17. **Electronic questionnaires** - The ONS has actively been pursuing the use of electronic questionnaires which contain validation checks, to help reduce the number of recontacts of contributors, and to improve the quality of the information received. The feedback from this process suggests that the contributors like having the main queries at the time of data entry, so that they can be immediately followed up, as this saves time in form-filling and responding to subsequent queries. The ONS plans to extend this service over the next few years (it is currently available only on diskette) so that it becomes available from a web site.

18. **Electronic data capture** - This is done by several means, but the most important point to come to light over the last year has been the level of recontact following scanning. On a structural (complex) business survey, >90% of forms returned failed at least one edit rule. This has resulted in a high level of recontact and a consequent drain on resources in survey areas. It is planned to review the editing procedures and rules in these surveys to see whether savings of resources can be made, and to introduce more automated procedures.

19. **Imputation of frame variables** - The annual structural business surveys are being put together to form a new large-scale sample survey, the Annual Business Inquiry. Since this will replace a periodic census which was used to maintain the employment variables on the ONS's sampling frame, (the Inter-departmental business register), the issue of how to update values on the

frame has been raised. Several methods are to be investigated, including regression imputation to correspond with the anticipated estimation method for the new survey.

For further information please contact:

For Census:

Jan Thomas
ONS
Segensworth Road
Titchfield
PO15 5RR
UK
Tel +44 1329 813296
Fax +44. 01329 813532
Email: jan.thomas@ons.gov.uk

For Methods and Quality:

Paul Smith
ONS
Cardiff Road
Newport
NP9 1XG
UK
Tel +44 1633 813436
Fax +44 1633 813166
Email paul.smith@ons.gov.uk