

CONFERENCE OF EUROPEAN STATISTICIANS

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

Item 5 of the provisional agenda

DATA EDITING AND IMPUTATION AT STATISTICS DENMARK

Submitted by Statistics Denmark¹

¹ Prepared by Birger Madsen.

I. INTRODUCTION

1. This paper provides a summary of the work that is being done at present at Statistics Denmark within the topics of data editing and imputation. During the last two years, a working group has had the task of developing guidelines for a complete strategy of data editing and imputation in Statistics Denmark as well as describing the existing data editing and imputation systems.

II. DATA EDITING: ERROR DETECTION

2. The focus in this area has mainly been on “Macro Editing techniques”, (see Granquist (1994)). These include the following methods among others:

- The Aggregate Method
- The Top-Down Method
- The Hidiroglou-Berthelot Method

3. The former two methods are widely in use in several statistical agencies. They are to be considered as general principles rather than exact methods. In short, they are characterized by the fact that they are output controls, i.e. they focus on detecting errors in table cells (aggregates) respectively in individual records which have a high impact on estimates. These principles are in use e.g. in the new Danish Quarterly Statistics of Earnings.

4. The Hidiroglou-Berthelot Method is probably less well known. It is in short a kind of scoring method (other similar methods exist), useful in periodic surveys. The rationale behind the method is that ratio of the current value to the previous value of some key variables.

5. Instead of studying directly the ratio between successive values of key variables, these ratios are transformed in order to:

- (a) assure that equal importance is given to small ratios and large ratios, i.e. that increases and decreases are detected equally efficiently;
- (b) ensure that larger units are detected more efficiently than smaller units, i.e. that a small deviation in a large unit may be more important than a larger deviation in a small unit.

6. The advantages of this method are that it is flexible because it is possible to control how much importance is to be given to large units and how many outliers are to be detected. Furthermore, it is objective because data themselves yield the information that is necessary, once the above-mentioned parameters have been fixed- no subjective element is involved. Finally, the method can easily be implemented using standard statistical software.

7. The Hidiroglou-Berthelot Method has recently been implemented in the Danish VAT Statistics (total turnover in non-agricultural industries).

8. The methods of macro editing apply mostly to surveys in which a few important numeric variables are of primary importance. In complex surveys many numeric variables may exist and the question is that of designing a complete set of edits for identifying outliers

and errors in data. These may include ratio edits, balance edits or more complex edit rules, see for example Cotton (1993).

9. In the near future it is expected that neural networks will provide a useful alternative to the traditional editing methods, especially in surveys where a large number of non-numeric variables exist, e.g. in Labour Force Surveys. Very few general methods for error detection in this type of survey exist. See Nordbotten (1995) for some studies of using neural networks in data editing and imputation.

III. DATA EDITING: ERROR CORRECTION AND IMPUTATION

10. One of the most obvious methods for error correction and imputation is to recontact the respondent. This is actually carried out frequently at Statistics Denmark to some extent in more business surveys. In addition to providing data in the present survey, this also has the desirable effect of increasing data quality in future surveys. Since much attention has been paid to the question of reducing the response burden, other methods are to be used increasingly in the future.

11. Most of the methods that have been studied at Statistics Denmark use some kind of donor principle, i.e. some way a "similar" respondent is found that provides the data needed.

12. In Madsen (1996) a very simple, hierarchical approach is described for finding a donor respondent in the Danish Labour Force Survey using only information from register variables. A preliminary study has been carried out using neural networks on the same data - more will be done in this area. Neural networks are presumably the most general and flexible method of data editing and imputation today, but their application within data editing and imputation is still at its infancy. Statistics Denmark are doing studies on the performance of neural networks in all areas of data editing and imputation at present.

13. Other methods which are applicable to numeric variables only are based on clustering techniques. These methods are implemented in GEIS from Statistics Canada which is considered an effective tool in surveys where a complete set of edit rules can be constructed (see Cotton, 1993, for details).

References

Granquist, L. (1994). Macro editing: A Review of some Methods for Rationalizing the Editing of Survey Data. Statistical Data Editing, Volume No. 1: Methods and Techniques. UN/ECE, 1994.

Nordbotten, S. (1995). Editing and Imputation by Means of Neural Networks. Statistical Journal of the UN/ECE, No. 12, 1995.

Cotton, C. (1993). Functional Description of the Generalized Edit and Imputation System (GEIS). Business Survey Methods Division. Statistics Canada (1993).

Madsen, B. (1996). Statistical Match in the Danish Labour Force Survey. Paper presented at UN/CES Work Session on Statistical Data Editing, Voorburg 4-7 November 1996.