

CONFERENCE OF EUROPEAN STATISTICIANS

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

Item 3 of the provisional agenda

**REPORT ON SOME NEW DATA EDITING METHODS
IN AUSTRIAN STATISTICS**

Submitted by the Austrian Central Statistical Office (ÖSTAT)¹

¹ Prepared by Alois Haslinger.

I. INTRODUCTION

1. This paper deals with some data editing methods in Austrian statistics which were developed and implemented during the last two years. Two of these projects are of particular interest. First, efforts to rationalize the editing of invalid codes of goods in trade statistics are described. The other project concerns the automatic imputation of missing answers in the Labour Force Survey.

II. INTRASTAT

2. Until 1994, the Austrian Foreign Trade Statistics used to be a complete enumeration based on customs data. With the entrance of Austria into the EU, the administrative customs data on Austria's trade with the EU countries were replaced by survey data (INTRASTAT), only statistics about trade with non-EU-countries are further compiled from administrative customs data (EXTRASTAT).

3. Although only transactions of enterprises with a innercommunity dispatch- or an arrival-value above 1.5 million ATS per year have to be reported for INTRASTAT, still 8 million transactions come in each year. 80 % of the reports are collected electronically via tapes, diskettes or mailbox, but 20 % on paper forms. Only half of the paperforms can be read with OCR, the other half has to be captured manually. Both data capture and editing of this large volume of data is a very costly and time consuming task.

4. The micro-editing of INTRASTAT reports occurs in 3 steps or modules:

(a) Editing of structure information (head of the report): mainly validity checks of the identification number of reporting units (VAT-number), month and year of report import/export and document number are performed. The forms are also checked for duplicates;

(b) Editing of qualitative variables of the micro-records: this phase consists of validity and consistency checks of all qualitative variables of individual transactions;

(c) Editing of quantitative variables of the micro-records: this phase consists of ratio edits of the variables amount invoiced, statistical amount and weight.

5. Although the quality of reports was improved due to intensive follow-ups and contacts with the respondents, still about 7% of all micro-records were erroneous in the first quarter of 1997. The errors are mainly concentrated in those records which are filled out with paper and pencil. At the start of INTRASTAT, almost all errors had to be manually reviewed. In the meantime each of the three steps of editing is automated, e.g. 7/8 of all records with an erroneous or missing qualitative variable are imputed automatically with a batch program. The automatic editing and imputation concerns records whose export or import value is below a certain threshold.

6. According to EU regulations, firms have to use a very detailed classification of goods (combined nomenclature), consisting of about 10.000 codes, each of 8 digits length. These

numerical codes are especially error-prone in paper forms. Either the code may be written incorrectly or a specified code is scanned and not all digits are recognized correctly. Both errors together result in a proportion of 4% invalid codes of scanned forms (compared to 0.2% invalid codes in the electronically transmitted forms). Until 1996 all invalid codes had to be edited interactively. Now a programme is tested which replaces invalid codes by valid ones.

6. The programme has the following logic: for each invalid code (consisting of at least 6 digits) a similar valid one is searched for which differs by the digits on at most 2 positions. The search for a similar valid code is restricted to codes which already have been used in the past by a reporting unit. If the search in this area fails, the search is expanded to all codes whose first two digits have already been used by the reporting unit. In case of a failure, the search is expanded to all 10.000 possible codes. If at any point in this sequence there are several solutions, then the code is selected which differs from the invalid code only at a position(s) as far to the right as possible. If the last criterion still finishes in a draw, a code with a higher digit on a given position is preferred to a code with a lower digit on the same position.

7. This programme was tested with a batch of records whose code numbers previously had been edited interactively. For 60% of the invalid codes the programme substituted the same valid codes as in the interactive editing. During the interactive editing phase there seems to have been a tendency to change largely the digits having their positions to the right. It is not yet clear whether the programme or the editors come nearer to the truth in the remaining 40% of differing imputations. The programme is tested now for records with a small statistical value and will go in production in case of a successful test.

8. The editing process will be further developed to restrict the interactive editing to the important records and to increase the capabilities of online interactive editing of this important records. As far as paper forms are concerned most of them should be scanned. The images can be retrieved already now in the edit-application without manually searching for them.

III. LABOUR FORCE SURVEY (LFS)

9. The annual EU Labour Force Survey has been part of the Microcensus (MC) in Austria since March 1995. The MC is a quarterly multi-purpose sample survey carried out since 1967 by means of face-to-face interviews in about 1% of all dwellings (n=30.000 dwellings containing about 70.000 persons). The survey covers the noninstitutional population of Austria (once a year a supplementary mail sample survey of persons living in institutions is carried out). The basic questions are on the size, the equipment and the costs of dwelling and refer to the demographic, the occupational and the educational status of the inhabitants. The MC was primarily designed to produce timely estimates of the demographic and social composition of the Austrian population, their families, households and dwellings.

10. The MC consists of a stable obligatory part - the so-called basic programme according to the Federal Statistics Act - and of a flexible voluntary part which is devoted to quarterly varying special programmes such as living conditions, travelling customs etc. The LFS is conducted in March of each year as a voluntary special programme and consists of a detailed inquiry into the employment situation (to meet the ILO-OECD-EU requirements for the

measurement of unemployment).

11. Since the LFS-part of the MC-survey is not mandatory there are usually about 10% of persons of the MC-sample who do not answer any of the LFS-questions (unit nonresponse). About 50% of the person-records have item-nonresponse, most of them between 1 and 3 empty questions. The MC-dataset contains a sample weight, which is used for the production of statistical tables. This weight is the product of a two-stage stratified estimation procedure and a raking procedure to fit the MC-results to the distribution of the Austrian population estimates according to age, sex and province, and according to nationality and province. This sample weight is produced on the basis of the records of the obligatory basic program. It was the request of ÖSTAT and EUROSTAT to use the same sample weight for the production of tables of the MC obligatory variables as well as for tables of the LFS variables. To get consistent tables from both parts of the survey it was necessary to impute the missing data of the LFS. Another reason for imputation was to prevent biased results due to differing nonresponse in different subgroups of the population.

12. For the imputation of missing answers to LFS questions a distance function was used, which measures the similarity or distance of a record to possible donor records. The choice of a distance function was influenced by the methodology used in the Canadian NIM edit and imputation project (Bankier, 1994). Generally, two data records can be considered as similar when the objects corresponding to the records have similar characteristics in some qualitative or quantitative variables. If one considers I_q qualitative and I_n numeric variables then the distance between two records R and S can be measured by

$$D(R,S) = \sum_{i=1}^{I_q} w_{qi} D_{qi}(R,S) + \sum_{i=1}^{I_n} w_{ni} D_{ni}(R,S)$$

For the LFS $D_{qi}(R,S) = 1$, if the value of the qualitative variable i of record R is not equal to the value of the same variable of record S; otherwise $D_{qi}(R,S) = 0$. $D_{ni}(R,S) = 1$ if $|R_{ni} - S_{ni}| > \text{maxdiff}$; otherwise $D_{ni}(R,S) = 1 - (|R_{ni} - S_{ni}| / \text{maxdiff})$.

13. The values w_{qi} and w_{ni} are weights according to the importance of variables qi and ni . For the LFS imputation in 1995, 4 qualitative variables (*sex, marital status, household position and employment status*) and the numeric variable *age in years* were used for the distance function. Age got the strongest weight, the other variables had equal importance. The parameter *maxdiff* for age was taken as 8 years.

14. Unit nonresponse records of the LFS were imputed using hot deck imputation of the values of a donor minimizing the distance function. The number of times a record had already been used as donor was stored, so that in the case of no unique solution the donor which had been used most rarely got priority.

15. The imputation of item nonresponse followed a similar path: for a given record with partial nonresponse all those records were used as donors, which had a valid answer on those positions that were missing on the given record. From this set of possible donors that record was selected which minimized the distance function.

16. After execution of the imputation the distributions of the imputed and not imputed values of some variables were compiled and compared. In most cases there were no great

differences, but in some cases the differences were striking, e.g. the share of persons getting unemployment benefits was significantly higher among imputed answers than among original answers.

References

BANKIER, M.(1994): Imputing Numeric and Qualitative Variables Simultaneously, Statistics Canada

BURG, T.(1997): Imputation fehlender Werte im Labour Force Survey, to appear probably in Österreichische Zeitschrift für Statistik.