WP. 43
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and**
**ECONOMIC COMMISSION FOR EUROPE**

**CONFERENCE OF EUROPEAN STATISTICIANS**

**Work Session on Statistical Data Editing**
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (viii): Selective and macro editing

# IMPLEMENTING SELECTIVE EDITING AND IMPUTATION METHODS IN FOREIGN TRADE STATISTICS

**Supporting Paper**

Prepared by Thomas Helmert, Federal Statistical Office, Germany

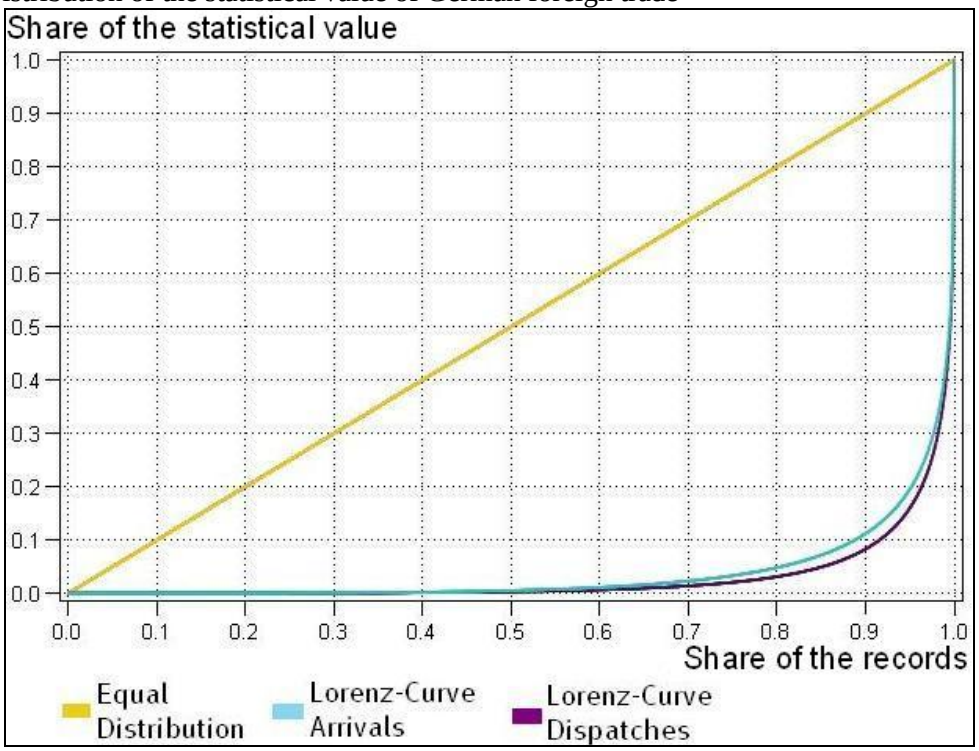
## I. INTRODUCTION

1. German foreign trade statistics has to disseminate first results in a relatively short time after the end of a month. First results for trade statistics have to be submitted to Eurostat at the latest 40 days after the end of the relevant month. Considering that up to 18 million records have to be processed per month, the data volume in the data editing process of foreign trade statistics can exceed the capacity limit. When, as in the case of the Federal Statistical Office, the staff numbers are falling, too, then it is a great challenge to keep up the quality level of statistical results. As a consequence, the data editing process at the Federal Statistical Office has been substantially reorganized by the new ASA (German abbreviation for automated processing in foreign trade statistics) system in order to increase the efficiency of data editing.

2. One key element to achieve a higher efficiency of the data processing is setting the right priorities for the data editing process by focussing the whole process on the results of foreign trade statistics. In the editing process, the records can be corrected either manually or by an automated procedure. Since it is hardly possible to correct the whole data manually it is necessary to select the records in important and less important for the results. The important records must be corrected completely manually because it is crucial for the quality of the results that all items of these records are correct. Approaches which correct certain items manually and others automatically cannot guarantee the quality of these records since it is often unclear which item is the error source of an erroneous record. In contrast, a procedure which informs the staff involved in the editing process about the importance of the records and thus enables them to determine which records must be corrected manually makes it possible to ascertain that the important records receive adequate attention and have a high quality level.

3. In the new ASA System individual threshold values for every commodity code are used to differentiate between important and less important records. Important records must be corrected manually to guarantee high quality results and rather less important records are corrected mainly by automated procedures. This processing system makes sense since the vast majority of records have almost no effect on the statistical results of foreign trade statistics. The distribution of the statistical value in German foreign trade statistics is shown in Figure 1:

Figure 1: Distribution of the statistical value of German foreign trade




4. As can be seen in Figure 1, about half of the records taken together account for about 1% of the total statistical value. In contrast, one tenth of the records with the highest values account for about 90% of the total statistical value. That may be due to the fact that the values of the goods traded are very heterogeneous (there are about 10000 commodity codes), but it shows that in most cases the statistical value is distributed very unevenly also within the commodity codes.

5. What relevance a record has for the results of foreign trade statistics is determined by numerical items. When these values are unevenly distributed, a promising approach would be to select the data into records which are either important or less important for the results in order to assure the quality of the important records. This contribution presents the new ASA data processing system in foreign trade statistics and shows how the records are prioritized by non-uniform threshold values.

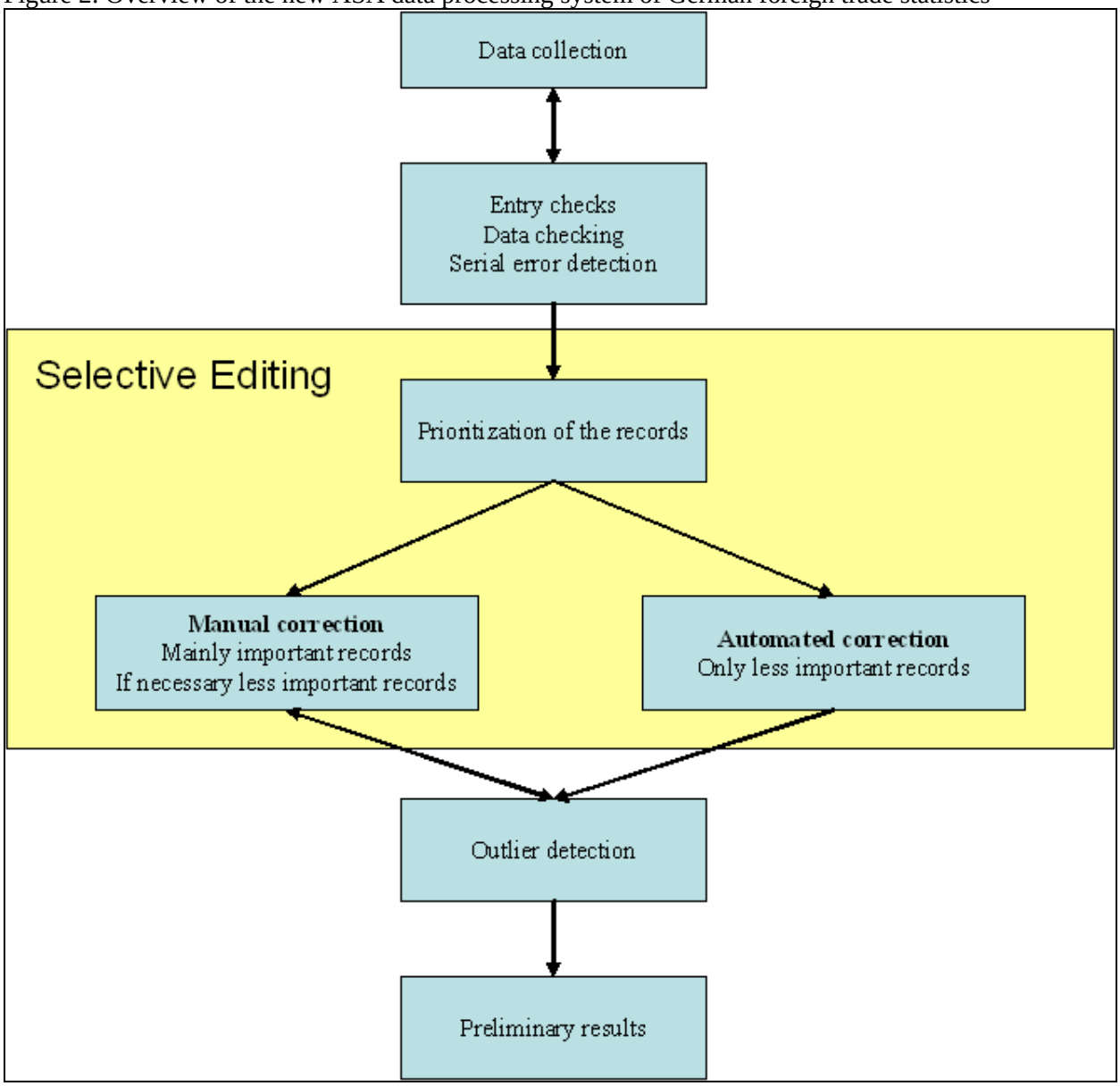
## II. SELECTIVE EDITING

6. In the new ASA data processing system of German foreign trade statistics, records are classified as being important or less important for the results. According to this distinction, the records are treated differently. Data processing and the classification of records are presented in this section.

### A. The new data processing system of German foreign trade statistics

7. The new data processing in the ASA-System is shown in Figure 2:

Figure 2: Overview of the new ASA data processing system of German foreign trade statistics




8. After having been submitted to the Federal Statistical Office, the records are checked for errors and plausibility. At the stage of where entry checks are made, serial errors are searched in the records delivered. Whenever a serial error is found, the delivery is rejected and the enterprise is contacted. If a delivery of records contains no serial errors, the records will be released for data editing.

9. During the data checking the records are classified in priority groups. The records are selected in 4 different priority groups. Priority group 1 consists of the records which are very important for the total and detailed results and the records in priority group 2 are important for the final results. In contrast, the records in priority group 3 are rather less important and the records in priority group 4 have almost no effect on the final results.

10. The allocation of records to priority groups makes it possible to control the different data editing methods. The main task of clerks dealing with manual data editing is to manually correct all the important records with errors in priority groups 1 and 2 since these records are not corrected by automated procedures. However, the clerks are also able to correct the rather less important records in priority groups 3 and 4. Towards the end of the editing period, the records in priority groups 3 and 4 are corrected by automated procedures. A nearest neighbour hot-deck imputation method and a deterministic

error correction approach are used for the automated error correction. If for some reason a record is still erroneous after automated error correction, it must be corrected manually.

11. After manual and automated error correction, the clerks check the results of data editing and search for outliers by comparing the results with those of previous months. If they find unusual results, they must check the edited data again and correct them manually if necessary.

## B. Prioritizing the records

12. The new ASA data processing system requires a classification of the records according to their relevance for the results. Hence, the records are selected to different priority groups. The records which are important for the results of foreign trade statistics are identified by threshold values. Due to the heterogeneity of the traded goods, different threshold values are calculated for every commodity code. The use of only one threshold value for all goods would lead to an inconsistent quality of the statistical results for the different commodity codes since goods with a high value would mainly be edited manually and low-value goods would almost completely be corrected by automated procedures. By contrast, the non-uniform threshold values assure that the results of every commodity code have the same quality because the records are prioritized within the commodity code. According to the distribution of the statistical value within the commodity codes in the past, the threshold values are determined by the following procedure:

(a) At first the records are sorted by the statistical value for every commodity code *j*:

$$x_{1j} \leq x_{2j} \leq \ldots \leq x_{nj}$$

(b) The second step is to calculate the records shares of the total statistical value for every commodity code *j*:

$$S(x_{ij}) = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}$$

(c) The third step is to cumulate the shares of the records:

$$F(x_{kj}) = \sum_{i=1}^{k} S(x_{ij})$$

(d) With the cumulate share function it is possible to classify the records into different groups. The records which fulfil the condition $F(x_{kj}) \geq 0,5$ are allocated to priority group 1.These are the records that have the greatest impact on the results with a share of the total of the statistical value of over 50%. If records are not in group 1 but fulfil the condition $F(x_{kj}) \geq 0,25$, they are allocated to priority group 2. Taken together, the records in group 1 and group 2 account for over 75% of the total statistical value. If records are not in group 1 or 2 but fulfil the condition $F(x_{kj}) \geq 0,1$ they are allocated to group 3. The remaining records are assigned to group 4. The records which are allocated to group 4 together have a share on the total of the statistical value which is below 10%.

(e) The thresholds for a priority group are determined by the mean of the lowest statistical value in a group and the highest statistical value of the group below. According to the procedure to calculate the threshold values for the priority groups they can be interpreted as follows: The records which were classified in priority group 1 because they have a higher value than the threshold value for priority group 1 should together account for a share of the total statistical value of about 50%. The records which have a lower value than the threshold value for priority group 1 but exceed the threshold value of priority group 2 should together have a share of the total statistical value of about 25%, etc.

13. We use the SAS software to analyse the distribution of the statistical value and to generate a reference file with the calculated threshold values. Of course it is possible to calculate threshold values for even more detailed groups with this procedure. Specific threshold values are generally calculated for every commodity code and flow. For the processing of Extra-EU trade statistics, specific threshold values

are also calculated for the different types of transaction. If a sufficient number of observations is not available for the calculation of the threshold values in the analysed time period, all threshold values are set to 1. In this case all data of the concerning group (such as a certain commodity code, flow, type of transaction) will be classified as very important.

14. For the prioritization of the records the collected data are compared with the relevant threshold values. For this procedure, the so called “fictional value” is calculated to avoid a misclassification. The fictional value is the maximum of the statistical value, the supplementary unit multiplied by the average value of the supplementary unit or the net mass multiplied by the average value of the net mass. The prioritization of the records by means of the fictional value has the advantage that it assures a correct classification even if the statistical value is incorrect, provided an average ratio of the correct statistical value to the supplementary unit or net mass. The fictional values of the records are compared with the threshold values in the reference data within the framework of the plausibility checks.

15. For the introduction of selective editing the data of the year 2008 were used to calculate the threshold values. This partially produced threshold values which did not work as expected. The main reason was that in the new processing system the data are much more aggregated than before. We tried to simulate the new aggregation with the data of the year 2008 but for some commodity codes this led to even more aggregated data than in the actual processing system. The consequence was a distorted distribution of the statistical value and thus partly too high threshold values especially for priority groups 1 and 2. Another factor that causes problems for the prioritization by threshold values was the sharp decline of German foreign trade especially in the first months of 2009. To solve these problems new threshold values were calculated on the basis of the first 3 months of 2009 for imports of non-EU countries. Table 1 shows the effect for the classification of the records for June when the new thresholds were used:

Table 1: Relevance of the priority groups for the statistical value and the data volume of the Extra-EU trade, Imports

| | Share in the statistical value | Share of the records | Share in the statistical value | Share of the records |
|---|---|---|---|---|
| | May 2009 | | June 2009 | |
| Priority group 1 (very important) | 46.0% | 2.8% | 66.1% | 5.4% |
| Priority group 2 (important) | 15.5% | 1.9% | 18.9% | 6.0% |
| Priority group 3 (rather less important) | 17.6% | 4.5% | 7.3% | 5.4% |
| Priority group 4 (less important) | 21.0% | 90.8% | 7.7% | 83.2% |

16. As can be seen in Table 1, the new threshold values calculated on the basis of the new processing system had a huge impact on the prioritization of the records. The records classified in priority group 4 accounted for only about 8% of the total statistical value in June compared with about 21% in May. By contrast the share of priority group 1 in the total statistical value rose from about 46% in May to about 66% in June. However, the prioritization still does not work as expected. Normally the share in the total statistical value should be defined by the threshold values and only the share in the total number of records should vary. According to that, the share of priority group 1 is much higher than the expected 50% and, by contrast, the shares of the other priority groups are much lower than expected. One possible reason could be the unstable situation of the foreign trade this year. An increase in the variance of the reported values results in a less precise prioritization with the threshold values. Now the threshold values are updated every month to solve the problems mentioned and we carefully monitor the actual developments.

17. Although the prioritisation of the records did not work exactly as expected, the main principle still works very well after the first adjustments. As can be seen in Table 1, only about 11.5% of the records account for about 85% of the statistical value in the different commodity codes (Extra-EU trade, imports). By contrast about 83% of the records together have an impact on the results of the different commodity codes which is below 8%. This clearly shows that even within the commodity codes the statistical value is distributed very unevenly and that it makes sense to focus the data editing process on the important records. Furthermore it shows that the prioritization of the records by non-uniform threshold values works in principle. However, we also experience the problems of this prioritization

system, for example in the case of a structural break (like the sharp decline of foreign trade at the end of last year and the beginning of this year) and therefore will also further improve the system.

## III. AUTOMATED ERROR CORRECTION

18. Near the end of the editing period when manual editing is mostly finished, the remaining implausible records in priority groups 3 and 4 are corrected by automated procedures. With the implementation of the new ASA system the automated error correction were also reorganized since the automated procedures become more important in the new processing system. The biggest change was the adaptation and implementation of a nearest neighbour hot-deck imputation method in the processing of foreign trade statistic. The hot-deck imputation complements the deterministic error correction and model based imputation which were the only methods used for automated correction until the ASA System was implemented. During the automated correction procedure the records are first corrected by the hot-deck imputation method. Records which are still erroneous after the hot-deck correction are corrected by deterministic and model based correction methods. If records cannot be corrected by automated procedures they must be corrected manually.

### A. Hot-Deck Imputation

19. The hot-deck imputation was adapted and implemented because the deterministic and model based error correction cannot correct all implausible records in an adequate manner. Especially the categorical items like country of origin/destination can hardly be corrected by pre-defined rules without the risk of distorting the results considering the great heterogeneity of the traded goods. For errors concerning the categorical items a nearest neighbour hot-deck imputation method is used in the ASA processing system. For this method a SAS program was developed which corrects the records outside of the main ASA system. The most similar donor record for a receiver record is determined by the following distance function:

$$D_{XY} = D(X,Y) = \sum_{k=1}^{r} w_k \left| x_k - y_k \right|$$

The distance of a donor record $X$ to a receiver record $Y$ is related to the sum of the differences of the items $k$ which are relevant for the similarity of the records. The distance value of categorical items will be set to 0 if they are equal ($x_k=y_k$) and to 1 if they are unequal ($x_k \neq y_k$). With the parameter $w_k$ it is possible to weight the influence of the items. This weighting parameter is used to avoid unnecessary changes of important items. Especially the item country of origin/destination has a high weight to avoid changes of valid countries. We try also to avoid unnecessary changes by identifying common errors which only necessitate the correction of a certain item and by generally correcting implausible records with a donor of the same commodity code. A basic overview of how hot-deck imputation is used in the processing of foreign trade statistics is provided by Table 2:

Table 2: Overview of the use hot-deck procedure for the processing of the June 2009

| | |
|---|---|
| Corrected records | 308604 |
| Corrected records with special error (1 item correction) | 237022 |
| Corrections of the item country of origin/destination | 47020 |
| Share of the corrected records in the total statistical value | 0.57% |
| Potential donors | 14211814 |

20. As it can be seen from table 2 the impact of the hot-deck procedure on the results of foreign trade statistics is very small since the records which are corrected with this procedure accounted for not even 1% of the total statistical value. However it is likely that the impact of the hot-deck imputation for certain commodity codes is much higher. The hot-deck procedure corrected 308604 erroneous records by over 14 million potential donors for the data processing of the June 2009. Therefore there are in general enough potential donors available for the hot-deck procedure. However for individual commodity codes the ratio of donor to receiver records can be problematic, since the erroneous records are corrected mainly by a donor of the same commodity code. After the hot-deck imputation method had been used for the first few times our experience and the feedback coming from the clerks were used to make some adjustments

to the new automated error correction. Certain errors were identified which can be corrected more easily (by a single item) than the normal errors and the determination of potential donors was also adjusted in special cases. As a consequence of these adjustments over three-fourths of the records corrected by the hot-deck procedure needed only a correction of a single item during the processing of the June 2009. The important item country of origin/destination was corrected by the hot-deck procedure at about 15% of the records. We currently analyse the consequences of hot-deck imputation for the results of foreign trade statistics. The first insight is that significant changes of the values of the item 'country of origin/destination' are extremely seldom even at commodity code level. To evaluate the changes made by the hot-deck method the results of manual data editing will be compared with the results of automated error correction by the hot-deck method. On the basis of this analysis the application of the hot-deck method will be further improved.

### B. Deterministic and model based error correction

21. Erroneous records that are not completely corrected by the hot-deck method are corrected by the deterministic or model based error correction. Possible reasons for records not being completely corrected by the hot deck procedure are that a donor may not be available or that this specific type of error is not corrected by the hot-deck method. For example, numerical items are not corrected by the hot-deck method. Instead, model based error corrections are used for numerical items. When the value of a categorical item is clearly determined by other items, then an error concerning this relation is not corrected by the hot-deck method, too. In these cases pre-defined rules are still used to correct the errors.

## IV. OUTLIER DETECTION

22. After the automated error correction procedures have been run, the results of the editing process are checked. To detect unusual developments, the clerks check the results for outliers. Therefore the so called "acceptance factor" is calculated:

$$AF_{idkj,t} = \frac{V_{idkj,t} - AV_{idkj,t-1}}{SD_{idkj,t-1}}$$

23. The acceptance factor $AF$ is the deviation of the value $V$ of a month from the average value $AV$ of the previous months divided by the standard deviation $SD$ of the previous months. The acceptance factor is calculated for every combination of commodity code $i$, flow $d$, type of transaction $k$ and country of origin/destination or federal state $j$. With this procedure it is possible to detect unusual developments even at the level of the different countries or states. The acceptance factor is also calculated without the country of origin/destination or federal state $j$ to verify if the total value of a commodity code itself is unusual. The outlier detection is not only confined to the statistical value since the acceptance factor is also calculated for the supplementary unit, the net mass and their ratio to each other and the ratio of the statistical value to the supplementary unit and net mass. The results with the highest acceptance factor are checked first. When the acceptance factor is greater than 2, which means that the current deviation from the average is more than twice as high as the standard deviation, then the value is classified as unusual. In this case the clerks have to verify respectively falsify the previous process steps. If necessary the records must be corrected again manually by the clerks.

## V. CONCLUSION

24. All in all, the first experiences with the new ASA data processing system are very positive, although we experienced some problems in the introduction phase. Meanwhile the prioritization of the records by non-uniform threshold values works in principle and therefore we are able to focus the manual data editing on the records which are most important for the results of foreign trade statistics. The rather less important records are corrected by a new automated correction system. With this new automated error correction, the automated data processing has also improved. The correction of the categorical items of the rather less important records by the new implemented nearest-neighbour hot-deck imputation method is normally convincing and often better than a deterministic error correction approach. However, in some cases the deterministic error correction is preferable to the hot-deck imputation and therefore is

still used. The new outlier detection enables us to find unusual developments even on a very low level. Currently we are analysing the results of selective editing and automated error correction. On the basis of the results of this analysis, the new data processing system will be further improved.

**References**

Blang D., “Neuausrichtung der Aufbereitung der Außenhandelsstatistik”, Wirtschaft und Statistik 12/2006

Blang D., Helmert T. “Verwendung von Hot-Deck-Verfahren in der Außenhandelsstatistik“, Wirtschaft und Statistik 11/2008

Chen J., Shao J.: “Nearest Neighbour Imputation for Survey Data” in Journal of Official Statistics, Vol. 16, No. 2, 2000

Giles P., Patrick C.: “Imputation Options in a Generalized Edit and Imputation System” in Survey Methodology, Vol. 12, No. 1, 1986

Kalton G., Kasprzyk D.: “The Treatment of Missing Survey Data” in Survey Methodology, Vol. 12, No. 1, 1986

Sande I.: “Hot-Deck Imputation Procedures” in Madow, W./Olkin, I. (Edt.): “Incomplete Data in Sample Surveys”, Vol. 3: Proceedings of the Symposium, New York 1983