STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working paper No. 24
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Work Session on Statistical Data Editing
(27 – 29 May 2002, Helsinki, Finland)

Topic (iii): Editing of administrative data

EDITING AND IMPUTATION OF TAX RETURN FILE – EVALUATION OF APPLIED METHODS

Contributed paper

Submitted by Statistics Finland[1]

I. INTRODUCTION

1. Tax Return File (TRF) is an important administrative data source used for construction of the Structural Business Statistics database at Statistics Finland. Generally the data for enterprises with 20 or more employees are collected with an inquiry while information for smaller enterprises is gained from TRF, which includes financial statements of every enterprise that is taxed due to business taxation act. Data is received annually to Statistics Finland from the National Board of Taxes.

2. Editing of TRF has been under development in recent years and three different methods have been applied to data. These methods in order of development are ratio imputation, scaling-method and mixed method. This paper describes briefly the methods used in editing and imputation of TRF, the creation of experimental data and the methods used for comparing three different editing and imputation methods. The evaluation is based on profit and loss account of the TRF of year 1998. Profit and loss account consists of 18 variables ($x_1,...,x_{18}$) and three subtotal variables (y_1, y_2, y_3).

II. EDITING

3. The main edit rule that is used to determine if a record is correct or incorrect is

$$\left|\sum_{i=1}^{18} x_i - y_3\right| < 1000FIM ,$$

where y_3 is the profit and loss for the year. If this inequality holds then the record is considered to be correct. Otherwise the record is edited to fulfil a condition.

4. Subtotal variables y_1, y_2, y_3 are used to locate the error in enterprises that do not follow the main edit rule. Subtotals should fulfil the following equations:

$$\sum_{i=1}^{7} x_i = y_1, \sum_{i=1}^{10} x_i = y_2, \sum_{i=1}^{18} x_i = y_3.$$

For example we can test the condition

$\left|y_1 + \sum_{i=8}^{18} x_i - y_3\right| < 1000FIM$ and if it is fulfilled we can conclude that the error(s) are located in variables $x_1,\ldots,x_7$. This basic rule is used to determine the localization of error for every incorrect enterprise. Group of erroneous variables is denoted *E*.

---

[1] Prepared by Janne Ikäheimonen (janne.ikaheimonen@stat.fi).

5. After the main edit rule and error localization some deterministic edits are applied for the incorrect enterprises. These edits are mostly sign checks of those variables that can be either positive or negative. If main edit rule is satisfied by changing the sign of the value of variable then the sign is changed and the record is considered to be correct. About 25 per cent of incorrect TRF records can be edited with these deterministic edits.

A. Ratio Imputation

6. Ratio imputation is a quite commonly used method in surveys (e.g., Shao, 2000). Method is based on ratios of sums of variables in correct records C_h, where *h* denotes imputation cell. Imputation cells are based on enterprise's industrial class and size class of turnover. Imputed value for variable x_j of incorrect record *i* that belongs to imputation cell *h* can be formularized

$$\hat{x}_{ij} = \left(\sum_{k \in C_h} x_{kj} \right) \Big/ \left(\sum_{k \in C_h} x_{k1} \right) x_{i1}^{*}, \; j = 2,...,18,$$

where x_{k1} is turnover data of TRF for correct records and x_{i1}^{*} is turnover data of business register for imputed record *i*.

7. Ratio imputation has been applied to TRF in such a way that it does not take advantage of localization of error so the method is applied to all *x*-variables except turnover (x_1), which is replaced with data of business register.

B. Scaling-method

8. The erroneous variables can be located to a certain part of account with the help of subtotal variables. The scaling-method is used to adjust these erroneous variables so that they sum up correctly.

9. The error between erroneous variables *E* and subtotal *y* is $\varepsilon = \sum_{i \in E} x_i - y_j$. The absolute sum of incorrect variables is calculated as $S = \sum_{i \in E} |x_i|$. Then the scaling factor is defined as $k = \varepsilon / S$. The scaling is performed as follows:
$\hat{x}_i = (1-k)x_i$, if $x_i \geq 0$ and $\hat{x}_i = (1+k)x_i$, if $x_i < 0$.
This rule works while $-1 \leq k \leq 1$. If $k > 1$ then *S* is calculated as a sum of negative values of x_i's ($i \in E$) only and these negative values are multiplied by (*1+k*). This applies inversely if $k < -1$.

C. Mixed Method

10. The mixed method is a mixture of outlier detection, nearest neighbour imputation and ratio imputation. The method takes advantage of localization of error the same way as the scaling-method. Firstly, the outlier-method is applied to all records. Then nearest neighbour imputation is applied to those records that could not be corrected with the outlier-method. In the last phase, ratio imputation is applied to remaining incorrect records.

11. The outlier method is used to correct erroneous records by changing only 1 variable in a record. Distributions of correct records are calculated for different industrial classes. The method is based on using 1. and 9. deciles of distribution as boundary values for outliers. If the value of a record compared to its own industrial class distribution is an outlier then that record is corrected by some rather simple rules.

12. Nearest neighbour imputation (e.g., Chen and Shao, 2000) is based on the distance function between incorrect record *i* and correct records *j* of the same industrial class. The method takes advantage of the hierarchical structure of industrial classification. This method is called hierarchical hot-deck imputation in Kalton and Kasprzyk (1986). Distance is calculated

$$D_{ij} = \sum_{k \in F} \left| \log(x_{ik}) - \log(x_{jk}) \right|,$$

where F is group of variables used in distance calculation. Values of correct record j with minimum distance $\min D_{ij}$ are imputed to incorrect record i. After imputation these values are scaled with scaling-method to sum up correctly.

III. EXPERIMENTAL DATA

13. Statistics Finland carries out an annual inquiry for largest enterprises. An experimental data can be constructed by comparing erroneous TRF enterprises with same enterprises of the inquiry that is considered to be correct. There are R=1093 erroneous enterprises in TRF-data for which the correct values can be obtained from inquiry. With this knowledge, experimental data can be constructed.

14. Experimental data is created by using error coefficients which are just proportions of values of the TRF and the inquiry. The error coefficient for variable x_i is

$$c_{x_i} = \frac{x_i}{x_i^*},$$

where x_i is erroneous in TRF-data and x_i^* is data from the inquiry.

15. The error vector is a 'vector' of error coefficients from comparison of one pair of TRF-record and inquiry record. So we have R=1093 error vectors that have value 1 in cells where there is no error and in erroneous cells the value is $\neq 1$.

16. There is an exceptional error type that can not be generated by the above described way. If TRF-data has a value different from zero and the correct value is zero then the error is introduced as a ratio $s_i = x_i / x_1^*$, where the TRF-data value is divided by the correct turnover of inquiry data.

17. The experimental data is based on restricted subset of 116146 correct enterprises in TRF-data. The proportion of erroneous enterprises in TRF-data is 14.9 per cent so nearly the same proportion is generated by introducing every error vector 16 times. That is M=16*1093=17488 (15.1 per cent) enterprises are generated erroneous.

18. First 1093 error vectors and 116146 correct enterprises are sorted randomly. Then the coefficients of first error vector are multiplied with corresponding values of first correct record to generate error(s) $x_i^{gen} = c_{x_i} x_i^{ok}$.
With the above mentioned exceptional error type the generation can be formularized $x_i^{gen} = s_i x_i^{ok}$.

19. In the next phase it is tested if the correct record has become erroneous. The first error vector is introduced to correct records in order until 16 records are made erroneous. Then the same is done for the second error vector continuing from the point next of last introduced error in correct data and so on until every error vector has been introduced 16 times into correct data. If the correct data runs out of records before every error has been introduced, generation is started from the beginning of correct data so that errors are not generated into same records more than once.

20. In generation procedure it is checked that every error in the error vector is introduced as it is meant to. For example, if the correct value is zero and error coefficient is also zero (which means that error is of type 'missing item') then the generated value is the same as the correct one which is not acceptable. These cases are rejected and the error is tried to introduce into next correct record as described above. This checking preserves the structure of errors in generated experimental data. This generation procedure is repeated 10 times to get 10 different experimental data sets. Every data set is edited and imputed using three different methods.

IV. EVALUATION

21. Let

$$\mathbf{x}_i^{ed} = \begin{pmatrix} x_{i1}^{ed} \\ x_{i2}^{ed} \\ \vdots \\ x_{i,18}^{ed} \end{pmatrix}, \quad i = 1,\ldots,M,$$

be the imputed data, where *ed*=ratio imputation, scaling or mixed method, and let

$$\mathbf{x}_i^{ok} = \begin{pmatrix} x_{i1}^{ok} \\ x_{i2}^{ok} \\ \vdots \\ x_{i,18}^{ok} \end{pmatrix}, \quad i = 1,\ldots,M,$$

be the correct values for enterprise *i*. The Mahalanobis distance (see Mardia et al. 1979)

$$D_i^2 = \left(\mathbf{x}_i^{ed} - \mathbf{x}_i^{ok}\right)' \Sigma_{ok}^{-1} \left(\mathbf{x}_i^{ed} - \mathbf{x}_i^{ok}\right), \quad i = 1,\ldots,M,$$

is used to compare every imputed record with its correct values. Σ_{ok} is the covariance matrix calculated from correct TRF-data enterprises. The final statistic for comparing methods is sum $D = \sum_{i=1}^{M} D_i^2$, which is calculated for every method in comparison.

22. The results of 10 simulations are presented in Table 1. Distances calculated from erroneous data are also presented for comparison. The smallest distances are bold. Table 2 shows the distances only for those enterprises which were edited by outlier-method.

23. Comparison is made also on aggregate level. Standard Industrial Classification (SIC) is used to divide data into *H* classes. Sums of variables are calculated for these classes. Because correct sum is known, edited and imputed data sets can be compared to it. The proportional error is calculated

$$e_{hj} = \left| \frac{\sum_{i \in h} x_{ij}^{ed} - \sum_{i \in h} x_{ij}^{ok}}{\sum_{i \in h} x_{ij}^{ok}} \right|$$

for class *h* and variable *j*. Notice that these proportions are calculated for erroneous part of data only so they do not present error proportions of the whole TRF-data. Results are presented in Tables 3 and 4. The numbers are averages of proportional errors for three different methods over 10 simulations. The bottom line of the table shows overall averages that are weighted by number of records in each SIC-class.

V. CONCLUSIONS

24. The results show that the scaling-method works most efficiently overall and ratio imputation is the least powerful method. The reason for the weakness of ratio imputation is obvious: because the method does not localize the error, it edits all the variables.

25. Although the scaling-method is best of these methods overall, it is not efficient for all variables. 'Direct taxes' is an example of a variable that can be either positive or negative. Because the scaling-method cannot change the sign of a value, it does not work well for that kind of variable. Another weakness of scaling is that it cannot correct missing values.

26. One big problem in editing of TRF is the amount of edited data. Because error can usually be localized only into a group of variables, it is obvious that also many correct values are edited. The outlier-method used in the mixed method is very efficient and about 20 per cent of real incorrect TRF-data can be

edited with it. However some quite big errors still remain in data edited with the outlier-method. Despite the limitations of use, the outlier-method it is a good way to decrease the amount of edited data.

27. Based on the results of this simulation, a new method has been developed. It combines the outlier-method, scaling and nearest neighbour imputation. Also a cold-deck imputation where the previous year's data are used, has been implemented in the current method.

Simul.	*Erroneous*	*Ratio*	*Scaling*	*Mixed*
1	44574764	857696,9	**317748,3**	815637,5
2	43206489	1336549	**458302,3**	619360,7
3	30291503	638394,5	**373586,9**	417596,7
4	1,01E+10	1370065	**358685,5**	494344,6
5	35954998	1145850	**221881,9**	373434
6	4,59E+08	535890,9	**222419,5**	492122,7
7	45929455	803510,8	**235561,4**	409376
8	75473679	746226	**456001,6**	544214,4
9	68124298	918637,3	**357682,7**	812359
10	1,62E+08	621577,7	**193952,1**	414941,8

Table 1. Mahalanobis distances. 10 simulations.

Simul.	*Erroneous*	*Ratio*	*Scaling*	*Outlier*
1	6826822	76079,4	73758,4	**10868,7**
2	13017611	35999,4	35230,4	**7023,2**
3	2408896	73733,5	147030,9	**21712,8**
4	5082577	85233,8	60061,6	**21500,2**
5	5290024	41323,8	47509,4	**16142**
6	9877289	84695,7	41871,3	**10775,7**
7	2090742	98586,7	41996,2	**30634,1**
8	35898796	67228,8	155028,7	**8193,5**
9	1275331	182925,9	**21265,7**	21878,4
10	1883007	68352,6	21206,7	**6443,2**

Table 2. Mahalanobis distances of records edited with outlier-method in mixed method. 10 simulations.

SIC	*Ratio*	*Scaling*	*Mixed*
1	**0,0916**	0,2336	0,1119
2	**0,0489**	0,2118	0,0947
3	0,0871	0,2243	**0,0577**
4	0,0698	0,219	**0,0307**
50	**0,0719**	0,2112	0,118
51	0,1229	0,1712	**0,1025**
52	0,0742	0,1858	**0,0607**
55	0,1932	0,2162	**0,1473**
6	**0,0804**	0,2423	0,141
7	**0,1692**	0,19	0,1815
Weighted aver.	**0,0997**	0,2066	0,1029

Table 3. Averages of proportional errors of variable 'direct taxes'.

SIC	*Ratio*	*Scaling*	*Mixed*
1	0,1905	0,0637	**0,0299**
2	0,067	**0,0136**	0,0371
3	0,1533	**0,0302**	0,0908
4	0,0674	**0,0101**	0,0587
50	**0,0596**	0,0598	0,0648
51	0,1107	**0,0363**	0,0806
52	0,0481	**0,0125**	0,0259
55	0,1738	**0,0305**	0,0952
6	0,1721	**0,0248**	0,0971
7	23,5593	**1,4724**	25,9007
Weighted aver.	4,1666	**0,275**	4,5408

Table 4. Averages of proportional errors of variable 'profit and loss for the year'.

References

Chen, J. and Shao, J. (2000). Nearest Neighbor Imputation for Survey Data. *Journal of Official Statistics* 16, 113-131.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology* 12, 1-16.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

Shao, J. (2000). Cold Deck and Ratio Imputation. *Survey Methodology* 26, 79-85.
