WP. 19
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ljubljana, Slovenia, 9-11 May 2011)

Topic (iii): Macro editing methods

Selektive Editing in the International Trade in Services

Invited paper

Prepared by Karin Lindgren, Statistics Sweden

I. Introduction

1. Statistics on international trade in services, wages and transfers are based on a quarterly sample survey (in the following called *International Trade in Services*) involving about 5 000 enterprises and organizations. *International Trade in Services* is an important part of the Balance of Payments statistics of Sweden and is therefore somewhat influenced of the regulations set up by IMF and EU regarding the international trade in services contents in the Balance of Payments. Statistics Sweden has been responsible for collection of necessary data since 2002, taking over responsibility from the Swedish Central Bank. The survey has for the last years had problems with excessive micro editing and has also been forced to do a lot of output editing under severe time constraints. It was obvious that something had to be done.

2. At Statistics Sweden a generic tool for selective or significance editing, SELEKT, has been developed, and is being implemented in different surveys including *International Trade in Services.* The purpose of this paper is to illustrate this implementation. The paper will begin with a section about the *International Trade in Services* and the properties of the survey, important in the implementation of selective editing. The second part deals with the features of SELEKT and what conditions which have to be met when implementing SELEKT. The final part describes the implementation of SELEKT in *International Trade in Services* and some of the problems and solutions in the implementation are reviewed.

II. About International Trade in Services

3. The data in *International Trade in Services* are collected on a quarterly basis from about 5000 sampled enterprises, central government authorities, municipalities, county councils and non-governmental organisations (NGOs) situated in Sweden. They report values concerning their income and expenditure as a subsequent of trade in services, compensation of employees and transfers, with foreign parties. The respondents are asked to provide data on the three following levels:

(a) Total expenditure and income originating from trade in services

(b) The expenditure and income specified by type of service or transfer

(c) The expenditure and income specified by country of the foreign trading partner for each type or service or transfer

4. The second level data are the primary input for the main statistical output and is decisive for the design of the selective editing. The information is collected for about 100 service or transfer categories (in the following called *Service Codes*) for income and expenditure respectively. The main output consists of thirteen service and transfer aggregates for each direction, i.e. 26 domains. Level one, total

value, is not used in any output but is merely collected as auxiliary information to be used in the editing. The estimated totals in the statistical output are based on aggregations of the values reported on the second level.

5. The sample is renewed once a year and at that point about a third of the sample is replaced. Important major enterprises and organizations in each stratum are chosen with the inclusion probability one which enables long time series for a large part of the sample. This is beneficial in the editing process. Only larger enterprises and organizations are asked to provide the data on the third level, i.e. country specific data, to keep the response burden down. The country specific data are used in an allocation model for the whole population but the results are not published as an important output table. In the future, the country specific data might be used in a more extensive way and the editing must be able to be adapted to such a situation.

6. The data are stored row-oriented with one observation of a service or transfer on each row. One observation contains the variables (columns) Corporate ID, Year, Quarter, Service Code, Direction, Country and Value. The submitted totals are stored with Service Code ‘090’ and the totals for each Service Code are stored with Country Code ‘A1’. This is exemplified in table 1 below.

Corporate ID	Year	Quarter	Service Code	Direction	Country	Value
165500000000	2009	4	090	1	A1	500 000
165500000000	2009	4	090	5	A1	30 000
165500000000	2009	4	410	1	A1	200 000
165500000000	2009	4	410	1	GB	100 000
165500000000	2009	4	410	1	PO	100 000
165500000000	2009	4	411	1	A1	300 000
165500000000	2009	4	411	1	NO	200 000
165500000000	2009	4	411	1	US	100 000
165500000000	2009	4	432	5	A1	30 000
165500000000	2009	4	432	5	BE	30 000
165500000000	2010	1	090	5	A1	40 000
165500000000	2010	1	432	5	A1	40 000
165500000000	2010	1	432	5	NL	40 000

Table 1.

7. The country specific values with the same Corporate ID, Year, Quarter, Service Code and Direction should of course sum up to the corresponding value with Country Code = ‘A1’. The same goes for the sum of values for specified Service Codes and the corresponding ‘090’-value.

8. Many of the previously used traditional edits are based on the comparison of different rows and the comparisons of aggregates of observations. Some examples are:

- Comparison of the sum of the values for the specified Service Codes and the values on the ‘090’-rows for each direction
- Comparison of the sum of the Country specific values for a Service Code and the corresponding totals for each direction
- Comparisons of new ‘090’-values and the corresponding ‘090’-values from previous quarters for the same enterprise/organization and direction

- Comparison of expenditure value and income value for the same enterprise/organization and Service Code

9. The previously used edits were not run in one context and each edit resulted in separate error lists. To keep the amount of editing down only the totals (‘090’-rows) were used in some edits, e.g., as in the third edit in the list above. Errors in values for specific Service Codes were then easily overlooked. Many of the current edits also had a restriction that the grossed up value must be more than 5 million SEK for the edit to flag the value. The unfortunate consequence was that erroneously low values not were flagged which could have a large impact on the statistical output.

III. About SELEKT

10. SELEKT is based on the idea that the potential impacts of suspected and fatal errors on the statistical output are estimated in order to determine which errors have the largest anticipated impact on the statistics. These errors are the most important to deal with in the manual follow-up including re-contacts with respondents, also considering the total response burden. This means that as much anticipated impact as possible shall be reduced by a minimum number of re-contacts.

11. The *suspicion* of an error can be calculated in different ways. SELEKT uses the difference between the observed value and an expected value for a so called test variable, relative to a measure of dispersion, to set the level of suspicion. In SELEKT the suspicion is thus computed as a continuous probability-like measure between 0 and 1. Another alternative is to set the suspicion beforehand for the current set of traditional edits. Different query edits can associate with different levels of suspicions. The choice of suspicion level can preferably be based on the computed hit rate of the edit. Fatal errors thus have suspicion = 1.

12. The expected values, and their dispersion measures, for the test variables can be calculated in two ways. Either using time series data for the enterprise/organization and Service Code or by using cross sectional data based on homogenous edit groups. In SELEKT time series data are used when available and cross sectional data otherwise.

13. To calculate the *potential impact* of a suspicious or fatal error the prioritized statistical output must be available. Since the final edited data are not available for the survey round that is to be edited, final edited data from one or several earlier survey rounds are used to estimate the sum for all domains. Based on these edited cold deck data, an expected value is also calculated for each combination of enterprise/organization and Service Code. The potential impact on the estimates is calculated as the weighted difference between the expected value and the observed value. The *anticipated impact* is calculated as the suspicion multiplied with the potential impact. Each observation gets a local score by relating this anticipated impact to the estimated sum or standard error of estimated sum, for a domain. The *local scores* are the anticipated impact multiplied with weights for different sets of domains and are used to rank the flagged values. The *global score* is a function of the local scores for a particular respondent and helps prioritizing the manual follow-up.

14. When domains are constructed in more than one way, for example by direction and service or by direction only, the anticipated impact must be computed for all such groupings. The relative importance of the classification, from the client’s point of view, should be set in SELEKT. The impact of erroneous data on the most important output tables are then prioritized in the calculations, while less consideration is made to tables of minor importance.

IV. SELEKT in International Trade in Services

15. *International Trade in Services* has, as stated, many traditional edits based on the comparison of different rows. In SELEKT all edits are made in one data step and therefore a lot of adaptations had to be made to the row-oriented data structure before the old edits could be incorporated in SELEKT. New variables were created and put on each row to enable the comparisons. One example is the sum of all specified Service Code values that were to be compared with the corresponding ‘090’-value. The new

variables *Sum_S* and *Value090* were created, as seen in table 2. The suspicion of each edit was decided based on the previous hit rates. The edits constructed to find fatal errors got a suspicion = 1. The thresholds for grossed up value are not included when the old edits are implemented in SELEKT since the use of potential impact replaces the need for thresholds in order to keep the amount of editing down. The traditional edits in SELEKT are mostly based on each Service Code and not only on the '090'-rows as they were before.

New variables. These are compared, discrepancies implies erroneous data

Corporate ID	Year	Quarter	Service Code	Direction	Country	Value	Sum_S	Value090
165500000000	2009	4	090	1	A1	500 000	500 000	500 000
165500000000	2009	4	410	1	A1	200 000	500 000	500 000
165500000000	2009	4	411	1	A1	300 000	500 000	500 000

Table 2.

16. Having three different levels of data in the same dataset enables all three levels to be edited at the same time. The "total levels" for each enterprise/organization are edited simultaneously with the single observations and all the re-contacts with the respondents can be made at the same time which is time-saving and reduces cost and response burden. The potential impact originating from errors on the '090'-rows and the country specific rows are not possible to calculate straightforwardly since they are not part of any statistical output produced today. To be able to calculate the anticipated impact of errors in the '090'-rows, a pseudo output table containing the estimated totals has been created. In the current implementation of SELEKT in *International Trade in Services* it was decided that the country specific rows should not be edited as individual observations. They are only edited through their aggregate for each Service Code and enterprise/organization. If more detailed output tables will be produced in the future with country specific data, the editing can be extended to involve the country specific values.

17. In the implementation of SELEKT both the current traditional edits and new SELEKT-type edits are incorporated. In the SELEKT-type edits the observed unedited values are compared to the expected values for each combination of Service Code, Direction and enterprise/organization and if they differ too much, SELEKT flags the observed value as a suspected error and computes a measure of suspicion. The expected values are computed as medians if there are at least three observations in cold deck data and last observation if there are less than three observations. When such "time series data" are not available cross sectional data are used. The cold deck data are ordered into a hierarchical structure of homogenous edit groups in order to calculate solid suspected values and dispersion measures. One major issue in the implementation is that the values are hard to predict based on the available background variables. Size of enterprise/organization has a clear impact on the values but otherwise neither line of business nor stratum are good predictable variables which makes it difficult to create efficient edit groups. The time series based expected values are preferred and are fortunately available for many observations. During the first two quarters of the year, however, all new enterprises and organizations in the sample have less than three observations in cold deck data. The homogenous edit groups were determined by the nine levels found in table 3.

Hg-level	Hg-variable	Number of digits	Description
1	Level	1	'1' = specified service, '2' = '090'-row
2	Direction	1	Income and Expenditure
3	Corp. Size	1	Six size groups for enterprises and organizations
4	Service group	2	The 100 service Codes are grouped into 50 groups, as homogenous as possible
5	Stratum	5	The strata are based on line of business and sector
6	NgS2	2	Line of business, two digit classification
7	NgS3	3	Line of business, three digit classification
8	Service Code	3	Service Code
9	Corp. ID	12	Enterprise or organization ID.

Table 3.

V. Possible Missing Data

18. One issue in the implementation of selective editing in *International trade in Services* was to solve the problem of missing values on certain Service Codes. If a respondent has reported a total income or expenditure in a '090'-row and no specified services, there are most likely missing values for one or several Service Codes. The problem is to decide which services that are missing. If a respondent had reported values for earlier quarters we might be able to assume which services that were missing. But what if the company never had reported any services before? It might be new in the sample or has recently started trading with services abroad. Another obvious error is when the value in the '090'-row and the sum of the Service Codes does not coincide. If the '090'-value are higher, then important information belonging to one or several services is probably missing.

19. There is no way of knowing what row or rows which are missing information, maybe a reported value is too small or maybe the row does not exist. The question is which observation to flag. Three alternatives are available. (1) The '090'-row can be flagged if the sum and the value do not sum up. (2) The Service Codes included in the sum can be flagged. (3) The third alternative would be to flag all rows involved in the edit. None of the alternatives is optimal. If a total consists of many Service Codes, each row would be suspected if the sum and the total differ. The error might not even be in any of those observations but originating from a missing observation. If all the observations were flagged then the global score will be highly "overestimated" for the enterprise/organization. The potential impact of the error will be calculated for the Service Codes that are submitted and the impact gets misleading. If only the '090'-rows were to be flagged, the flag indicates that there is an error on the '090'-value. The impact of an erroneous total is calculated which is misleading. Ideally the impact of the missing values should be calculated but that is not possible without further assumptions. The solution to the problem was to not calculate the impact of these types of errors. All '090'-rows flagged for this type of error go straight to the final error list without selection. A consequence is that even very small errors get edited but it might be good to deal with all these types of problems as soon as they appear, especially if the respondent is new in the sample. Rounding errors are of course allowed in these edits.

VI. Issues still to be resolved

20. Only values separated from zero are evaluated when deviating from the expected values. Submitted zeros are ignored. Ideally some zeros should be edited and some ignored. It is likely that some enterprises only trade with certain types of services during specific quarters, and not during all quarters of the year. Some services are only bought once a year or maybe even less frequent, i.e. some zeros are not to be suspected because they are supposed to appear once in a while. To this day we have yet not found a way to formulate an edit for this problem and incorporate it in the selective editing process. There is one edit that finds missing Service Codes when the enterprise/organization have

reported them for two straight years and another edit which controls if a service that always has been zero suddenly have a reported value is also incorporated, but no less blunt edits. A SELEKT-type edit which suspects zeros based on time series data would be preferred.

VII. References

Statistics Sweden, 2010. *A General Methodology for Selective Data Editing, version 1.0*.
Statistics Sweden, 2011. *User's Guide to SELEKT 1.1, A Generic Toolbox for Selective Data Editing,*