WP. 45
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

EVALUATION OF SELECTIVE EDITING IN SLP 2008

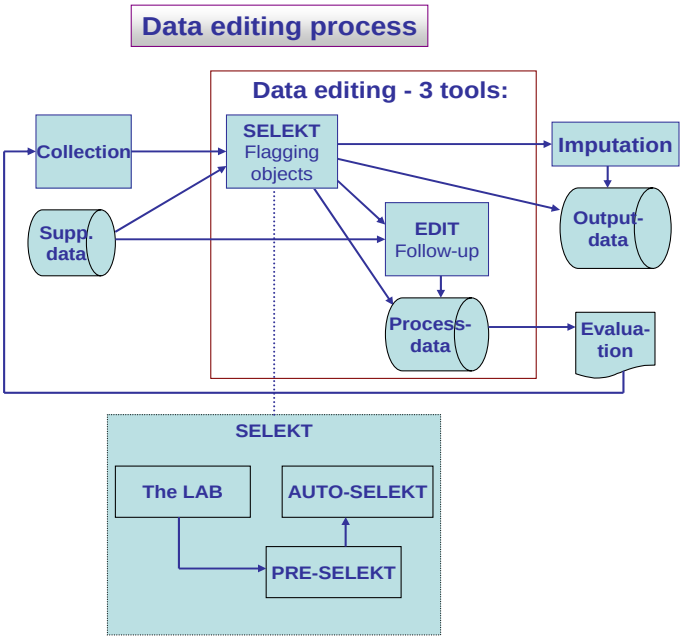
Supporting Paper

Prepared by Chandra Adolfsson and Peter Gidlund, Statistics Sweden

I. INTRODUCTION

1. The National Mediation office (MI) is the responsible authority for the official statistics of wages and salaries in Sweden. Statistics Sweden (SCB) is engaged by MI to produce the statistics. MI has recently demanded a reduction of the price for the services of Statistics Sweden. One possibility for SCB to reduce the price is to decrease the amount of production editing being performed and this may be possible to carry out by implementing selective editing with score functions. It was suitable to start with the relatively expensive and complex survey Wage and salary structures in the private sector (SLP).

2. At the present there is an extensive amount of development going on at Statistics Sweden to create and build general collection and editing tools, SELEKT[1] and EDIT. SELEKT is the environment where all the calculations of the scores take place while EDIT is the working space which will be used by the editing staff for manual review. The included components of the future tool box of Statistics Sweden regarding editing is shown below:



[1] A description of the method being used in SELEKT can be found in *A General Methodology for Selective Data Editing, preliminary version* (SCB 2009).

SELEKT and EDIT are not yet in place and because of this a decision was made to build a prototype of SELEKT and to implement it in the already existing SLP-production system. The prototype was implemented in SLP 2008.

3. The main purpose of this evaluation is to create basic data for decision-making regarding whether the prototype should be implemented in other already existing IT-production systems or not. The basic data for decision-making consists of an analysis of the implementation process from both a qualitative and a quantitative perspective. The experience of the implementation process of everyone involved is described in the report[2]. Furthermore, the results are presented in terms of changed amount of production editing and changed costs.

4. Now it is up to Statistics Sweden to decide if the prototype should be completed into a first real version of SELEKT and implemented in other similar business surveys' production systems or not.

II. EVALUATION OF SELECTIVE EDITING IN SLP 2008

5. The implementation process was characterized by time pressure because a lot of work had to be done in a short amount of time before the production of the SLP 2008 could begin. The work that was done consisted of adapting the already existing SLP-production system and to find appropriate parameter settings. The time pressure attributed to that several adjustments and improvements had to be made during the survey round, as well in the SLP-production system as in the SELEKT-prototype. One of the things that were done was an update of the threshold values when about half of all the individuals had been reviewed. Many of the problems and teething troubles that occurred during the survey round complicate the analysis. Besides this there were two survey variables excluded from the selective editing, Occupation and the Swedish Social Security Number (SSN). These were instead traditionally edited. The reason is that they are part of a register of Statistics Sweden and imputed values are not allowed in the register[3]. If Occupation and SSN had been included in the selective editing the reduction of the amount of production editing being performed would have been even greater.

A. Description of the SLP survey

6. The purpose of SLP is to provide information about the level of salaries, salary structure and changes over time for different categories of employees in the private sector.

7. SLP is a yearly survey[4] where the target population consists of all employees at the enterprises in the private sector in September each year. The main parameters are average monthly salary for non-manual workers and average hourly wage for manual workers. Within each category of employees the most demanded publication tables are created by cross sections of the variables Gender and Industry and Gender and Occupation.

8. Statistics Sweden only collects and reviews about 10-15 percent of the complete SLP-data material on the level of individuals while a number of Employers' Associations (AO) collect the rest. Nevertheless the data material that Statistics Sweden collects represents almost 35 percent of the weighted number of employees because of the fact that the enterprises from whom Statistics Sweden collects the data are as a rule small and often have high design weights. Selective editing has only been

[2] This paper is a shortened version of the more extended project report *Utvärdering av selektiv granskning i SLP 2008* (SCB 2009) which is only available in Swedish.
[3] In SLP the Swedish SSN is used for deriving the variable Gender which is given by the 11th digit of the SSN. If imputations were allowed it is the variable Gender that would be imputed, when necessary, not the whole SSN.
[4] The survey design is one-stage cluster survey sampling where enterprises are stratified by Industry and Number of employees in 324 different strata. Within each strata there is an SRS drawn. The enterprises are the primary objects while the employees are the secondary objects. The enterprises provide data of their employees regarding variables such as Occupation, Salary, Number of hours worked and Variable supplements.

applied on the 170 000 employees at the about 4 300 enterprises from whom SCB collects data. SCB collects the data via postal or web questionnaires and by file transfer. After the editing process has taken place the data that has been collected and edited by Statistics Sweden and the data delivered from the Employers' Associations are put together. Output editing on the complete data material is performed before the estimations are made.

9. Hopefully selective editing can be used on both the data material collected by Statistics Sweden and by the Employers' Associations in the next survey round (SLP 2009). This would enable Statistics Sweden to get an indication of how the data quality varies in different parts of the complete data material. Furthermore, Statistics Sweden would get support, even for the data material from the Employers' Associations, regarding identification of individuals with great impact on the estimates.

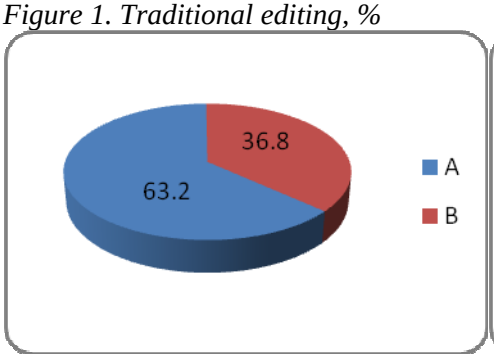
B. Reduced amount of editing

10. The results that would be reached in an upcoming survey round are best reflected by the results based on the updated threshold values. This is because of the fact that the updated threshold values will be used until a new analysis of the parameter setting is made in SLP.

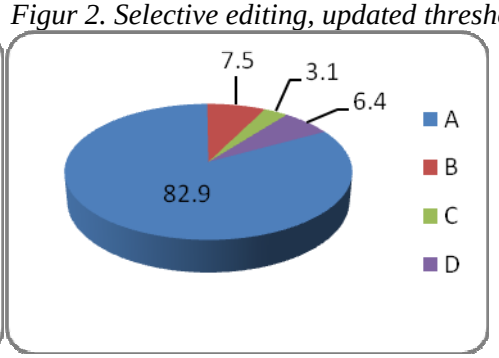
C. Individuals

11. With the updated threshold values the fraction of individuals that is manually reviewed is decreased from 36.8 percent with traditional editing to 17.0 percent with selective editing of all variables excluding Occupation and SSN. This corresponds to a decrease of 54 percent. If no variables had been excluded only 10.6 percent of the individuals would have been manually reviewed. The decrease compared to traditional editing then would have been 71 percent.

Figure 1. Traditional editing, %



Figur 2. Selective editing, updated threshold values, %



12. In *Figure 1* category *A* consists of individuals that are not error-flagged with traditional editing. Category *B* consists of individuals that are error-flagged by at least one edit check. These individuals would have been manually reviewed with traditional editing.

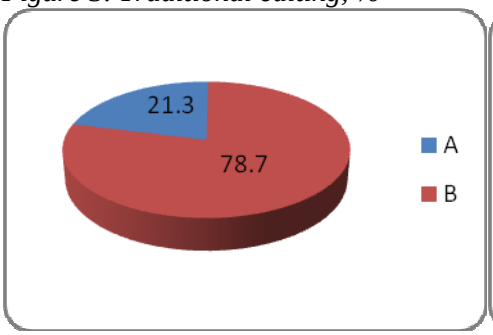
13. In *Figure 2* category *A* consists of individuals that are not selected by SELEKT. Category *B* consists of individuals that are selected by SELEKT and have valid values on Occupation and SSN. The individuals in category *C* and *D* have non-valid values on Occupation and SSN. The individuals in both category *C* and *D* were manually reviewed because of the exclusion of these variables from the selective editing. Among the individuals in category *C* and *D* it is only the ones in category *C* that would have been selected by SELEKT if imputation regarding the excluded variables was allowed.

14. It is thus the individuals in the categories *B*, *C* and *D* that were manually reviewed when the updated threshold values were used. Together these individuals correspond to 17.0 (= 7.5 + 3.1 + 6.4) percent of all collected individuals. If Occupation and SSN were included in the selective editing the fraction of individuals to review manually would have been reduced to 10.6 (= 17.0 – 6.4). The manual review of the individuals in category *D* would then have been replaced by imputation.

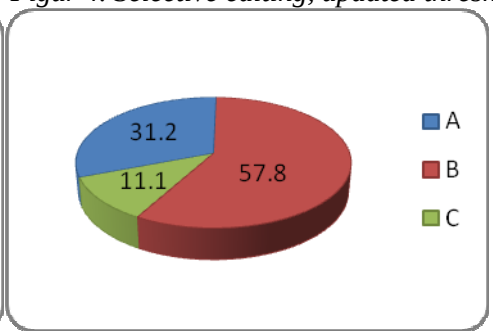
D. Enterprises

15. With the updated threshold values the fraction of enterprises that were manually reviewed were decreased from 78.7 percent with traditional editing to 68.9 percent with selective editing of all variables excluding Occupation and SSN. This corresponds to a decrease of 12.5 percent. If no variables had been excluded 57.8 percent of the enterprises would have been manually reviewed. The decrease compared to traditional editing then would have been 26.6 percent.

Figure 3. Traditional editing, %



Figur 4. Selective editing, updated threshold values, %



16. In *Figure 3* category *A* consists of enterprises without any traditionally error-flagged individuals. Category *B* consists of enterprises with at least one error-flagged individual regarding at least one traditional edit check.

17. In *Figure 4* category *A* consists of enterprises that were not selected by SELEKT while the enterprises in *B* were selected by SELEKT. Category *C* consists of enterprises that were not selected by SELEKT, but were still manually reviewed because of the traditional editing of Occupation and SSN.

18. The enterprises in the categories *B* and *C* were manually reviewed when the updated threshold values were used. Together these enterprises corresponds to 68.9 (= 57.8 + 11.1) percent of all collected enterprises. If Occupation and SSN had been included in the selective editing the fraction of enterprises to manually review would have been reduced to 57.8 (= 68.9 – 11.1) percent. The manual review of the enterprises in category *C* would then have been replaced by imputation.

19. It is likely that the reductions in the upcoming survey round (SLP 2009) will be greater. The reason for this is that the presented results are only based on the part of the collected data for which the updated threshold values were used. This part consisted of larger enterprises that were error-flagged to a higher extent than the first part of the data.

E. Quality aspects

20. One indication of quality is given by the size of the standard errors of the most important estimates. The number of main publication cells that were not published in SLP 2008 due to large sampling variance was, more or less, equal compared to earlier survey rounds. The number of imputed values, though, was substantially increased compared to earlier survey rounds. This is explained by the fact that imputation in some cases is a substitute for manual review when using selective editing.

F. Cost-savings

21. The total cost-saving in SLP 2008 was, expressed in hours, 1 100 which corresponds to one fourth of the total costs for collecting and production editing. About 1 000 of these hours were spent on adapting the existing SLP production system and developing the SELEKT- prototype. The latter cost will not occur in future, but during SLP 2008 demands on the existing production system have been raised. This type of cost is mainly expected to occur in the nearest survey rounds. A reasonable estimation of the future cost-savings in each survey round of SLP, compared to using traditional editing, is about 25 percent.

When the general production system EDIT is in place it is likely that the cost-savings can be further increased. The main reason for this is that EDIT is likely to be more stable and user-friendly than the existing SLP-production system. An estimation regarding how the costs are affected by access to EDIT has not yet been made.

G. Experiences

22. The editing staff thinks that their work feels more meaningful than before. Their experiences of working with selective editing is summarised by the following words from one member of the editing staff: "Now we know that the manual review actually is important and that our work really matters". The staff that was involved in the implementation thinks that the work has been burdensome at some times, but when summing up the experience of the implementation everyone thinks that is more positive than negative.