UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (ii): Managing and supporting changes related to editing and imputation

Implementation of selective editing methods at Statistics Finland using innovative and efficient team work methods

Prepared by Saara Oinonen, Statistics Finland, Finland

I. Introduction

1. On recent years Statistics Finland has invested a lot to improve and enhance editing practices of statistics. An editing model has been introduced (Ollila & Rouhuvirta, 2011), as well as a new SAS-based software tool with broad selection of editing methods called EG EDIT (Oinonen, 2014). A SAS-based program for selective editing called SELEKT, developed in Statistics Sweden (Nordberg et al. 2010), and compilation of several SAS-based editing modules from BANFF, developed in Statistics Canada (Statistics Canada, 2007), are included in EG EDIT, complemented with various macro programs developed at *Standards and Methods* department in Statistics Finland. From September 2014 a project started to implement selective editing methods to eight statistics. This implementation task requires effort from both statistical production staff and experts from methodology and data collection units with occasional support of the IT staff.

2. With efficient team work and agile project methods borrowed from *scrum* development methodology (Sims & Johnson, 2011), supported with wiki information and training events, promising results have been achieved: New editing methods have been very well received at participating statistics and an editing program EG EDIT including selective editing methods has been successfully implemented to be used in production in five chosen statistics so far. Main implementation tasks for each statistics were carried out in two-week sprint-sessions, which started on surveying the current editing practices of the statistics and led to testing and simulating the program EG EDIT and chosen parameters into production. The implementation work carries on to the end of the year 2015 and in 2016 all the eight chosen statistics will have EG EDIT as a part of their statistical production process.

II. Implementing selective editing methods

A. Background and planning

3. Statistics Finland has carried out several projects for developing statistical editing methods in past six years. Starting by surveying the editing practices of statistics, researching international editing methods and producing editing model on 2009-2011, to constructing editing program EG EDIT on 2012-2013 and leading to the implementation project launched on September 2014. The EG EDIT program is very versatile and flexible, so it is suitable for most statistics produced at Statistics Finland. However it was a request of the management that to increase the efficiency of statistical production the selective editing methods should be implemented first (Pyy-Martikainen, 2014). This is why the implementation started on statistics that are suitable for selective editing methods.

4. Eight statistics were chosen to be part of the implementation project. It was evaluated that with these eight statistics the implementation of selective editing methods would be the most beneficial in

terms of time and cost savings. It was also confirmed from statistical production staff that the timing of this implementation work was suitable and timetables were tailored to take account the production schedules. A project leader was chosen outside of the editing methodology expert team to carry out administrative tasks so that experts could concentrate only on implementation tasks. Comprehensive wiki site for EG EDIT as well as the concept library were produced well before implementation project was launched to support statistical production staff.

5. The eight chosen statistics are: *Innovation statistics*, *Environmental protection expenditure in industry*, *Research and development statistics*, *Waste statistics*, *Energy use in manufacturing*, *Local government sector wages and salaries*, *Private sector wages and salaries* and *Goods transport by road*. In May 2015 first five statistics have already had implementation phase and on three the production with new editing methods is on-going. Rest of the statistics will have implementation phase by the end of the year 2015 and year 2016 will be the first when all the eight statistics will have their production with new editing methods.

6. Before the intensive implementation phases an educative seminar was arranged for all the participating statistics. On this seminar experts of editing methodology explained the logic and terminology of selective editing and introduced the EG EDIT program for the future users. Also the nature of intensive implementation periods were explained. More education and support is needed when EG EDIT is first taken into production.

B. Implementation phase

7. Implementation of EG EDIT program is partially done by agile team work methods borrowed from *scrum* frame work. On this scrum-like approach experts of editing methodology, data collection and statistical content dedicate two weeks only for implementation work, so the work is done efficiently and without interruption. This intensive two week session is called *a sprint*. The team of editing methodology and data collection experts which carry out the implementation practises of the EG EDIT is called *development team*. Development team has a team leader called *scrum master* who takes care that development work is running properly and distributes tasks among developers.

8. In original scrum framework the scrum master does not participate the actual development work but has more facilitative role, but in our adaptation the scrum master is also one of the developers. It is the project leader's responsibilities to facilitate the implementation phase, e.g. taking care of data access permissions and recruiting the staff from statistics. On scrum methodology a person who represents the future users of the product and is responsible for aligning the development work is called *product owner*, but there has been difficulties to adapt this member on our project. Generally it is considered to be someone from participating statistics with adequate authority.

9. In the original idea of scrum the sprints would iteratively follow each other's, but in our development work it was not possible due to schedules of statistical production. Also it was essential to have time for other work, editing-related or not, between the sprints. On scrum methodology, it is also presented that after multiple rounds of sprints the development team would get more skilled, so that future sprints would be more efficient. This has proven to be correct also with the looser sprint schedule: Knowledge and skills of the development team have improved after every sprint. Every statistics have their own special features so the development team learn a lot during different implementation phases.

10. Implementation phase consists of eight sections (figure 1.):
    (a) Data preparations;
    (b) Analysing the current editing procedures of the statistics;
    (c) Defining the special features of statistics and statistics requirements for selective editing;
    (d) Defining the criteria and parameters of selective editing;
    (e) Testing the criteria and parameters of selective editing;
    (f) Choosing the final EG EDIT version to be taken to the production;
    (g) Launching EG EDIT in production of statistics;
    (h) Reporting.

Sections (a), (g) and (h) are not included in sprint session. Sprint session can be divided in three parts: *Review discussions* consists of phases (b) and (c), *Preparing the EG EDIT* consists of phases (d) and (e) and *Conclusions and follow-up* consists of sections (f) and (g), though (g) is done after the sprint. To addition of these one sprint session involves start-up meeting, half-way inspection and a demo, where the results are presented. These are adapted from scrum methods.

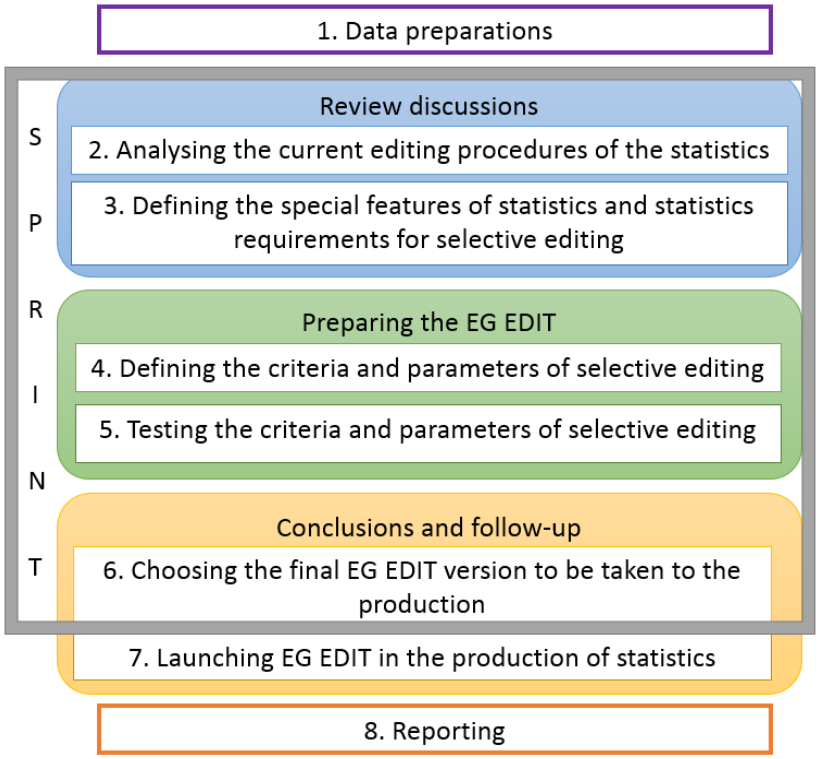

Figure 1. Implementation phase consisting of eight sections, translated and modified from *Pauli Ollila*s pp-slides in Finnish. Sections included in two week sprint session are outlined with grey box.

11. Before the sprint the statistics are asked to prepare data sets to be incorporated to EG EDIT with the help of the data collection expert. For some statistics data were easily constructed, but for some it required quite a lot of work. Also the technical framework where the statistical production takes place varies a lot and might create extra challenges for retrieving wanted data sets. Usually data needed to be converted to the form that EG EDIT is able to handle. For some cases the frame and sample information were difficult to reconstruct, as well as defining the unedited data. An expert from data collection unit was irreplaceable at this part of the implementation work.

12. The two week sprint starts by analysing the statistics current composition and editing practises thoroughly by review discussions. To make the start more rapid the statistics are given a form to fill in advance, which covers all the aspects of statistical production that would be associated with editing (e.g. units and variables, sampling, production schedules, quality requirements etc.). The form offers a good basis for dialogue between methodologists, data collection experts and statistical content experts. Open dialogue plays a key role in the beginning of the sprint since the methodology experts usually do not know all the aspects of the statistics. The form supplemented by methodology experts' notes also offers good basis for constructing new editing process for statistics. Usually statistics also offered extra documentation for experts to study. In the beginning there are usually long meetings for few days, so that all the aspects of the statistics are taken into account before planning.

13. After the basis of statistics production has been reviewed, the implementation team need to have consensus of the variables to be included in selective editing procedures. The main idea of selective editing is to focus on most important units and variables and explaining this to the statistical production staff is very important. Sometimes also the level and structure of units can cause a lot of speculation, especially if data has complicated structure or it contains a lot of important categorical variables.

14. For example, on *Waste statistics*, there are almost seven hundred different waste types that 2 300 different waste management units can report. Original data form has one waste type on each row, yearly data has about 40 000 rows. Thus the data is quite dispersed. Data is viewed and corrected by managements units. For EG EDIT, the waste types were aggregated to 51 + total waste classes and data were formed as clusters of waste classes on every waste management units. Since some of the most important variables were categorical, waste treatment for example, the amount of waste were allocated to each category, so that waste treatment variable with six categories became six continuous variables in the data (treatment 1, treatment 2 etc.) with reported amount of waste distributed on each treatment variable. Similar alterations were made to other categorical variables like management unit role, type of waste, origin of waste etc. Data alterations are illustrated on figure 2.

Original data form in statistics editing system

Role	WM Unit ID	Waste ID		Amount	Type of waste		Type of waste treatment	
Rooli	Kp Tunnus	Vahtitunnus		Määrä	Jätetyyppi		Käsittely	
1		03010500	Muut kuir	85 160,00	1	Tavanomainen jät	R035	Orgaanisen jätteen materi
1		03010500	Muut kuir	80520	1	Tavanomainen jät	R035	Orgaanisen jätteen materi
1		03010500	Muut kuir	40 200,00	1	Tavanomainen jät	R035	Orgaanisen jätteen materi
1		10010100	Pohjatuhl	3 585,40	1	Tavanomainen jät	R052	Epäorgaanisten jätteiden r
1		20014000	Metallit	192,39	1	Tavanomainen jät	R041	Metalli- ja metallipitoisten

Scores from EG EDIT to newly structured data

					TOTAL AMOUNT				TREATMENT3				TREATMENT4			
WMU	id	Flag	SCORE2	SCORE3	SCORE4	Susp	Value	Prev_Last	SCORE4	Susp	Value	Prev_Last	SCORE4	Susp	Value	Prev_Last
006	42	F	2 032 939	1 017 154,86	5 537,12	1,00	21 716 918	18 295 502	1 773,26	1,00	3 628 439	4 181 666	4 483,91	1,00	8 148 598	4 804 782
006	52	F	2 032 939	1 015 763,11	4 962,30	1,00	21 727 183	18 306 306	1 414,81	1,00	3 638 349	4 191 069	4 224,62	1,00	8 148 953	4 806 183
006	39	F	2 032 939	19,91	7,90	1,00	9 910	9 403	12,01	1,00	9 910	9 403	0,00	0,00		
006	07	F	2 032 939	0,60	0,19	0,37	309	269	0,00	0,00			0,25	0,37	309	269
006	35	F	2 032 939	0,31	0,07	0,12	29	141	0,00	0,00			0,18	0,12	29	141
006	43	F	2 032 939	0,06	0,02	1,00	5	2	0,00	0,00			0,03	1,00	5	2
006	03		2 032 939	0,00	0,00			977	0,00	0,00			0,00			977
006	02		2 032 939	0,00	0,00	0,00	12	12	0,00	0,00			0,00	0,00	12	12

Figure 2. Data alterations and error scores to *Waste statistics* data. Original data has one waste id with different attributes on each row and many categorical variables. Data in EG EDIT has 51+total waste aggregates (id) and waste amount has been distributed to categorical variables (type of waste treatment). Score 2 refers to waste management unit (WMU), score 3 to waste aggregate (id) and score 4 to variable. For every variable there is suspicion (Susp), new value (Value) and previous value from edited data (Prev_Last). This is one example of a table that is used as a basis for error list design.

15. Next the whole team focuses on defining parameters, edit rules and edit functions for EG EDIT. On some cases statistics have been able to provide existing editing rules in programming language or at least some documentation of them. These existing rules are very valuable for development team and quite often they can be transferred straight to EG EDIT. Data preparations are usually done prior to the sprint so it is possible to start testing the parameters and edit rules right away. EG EDIT also has many tools to help developers to find powerful edit rules. For example, for each edit rule the units that are caught by the rule are listed, both erroneous and non-erroneous, so it is easy to evaluate the efficiency of edit rules. Program also lists all those erroneous units that are not caught by any rule, so developer can look for those to find inspiration for new rules. After several sprints the development team has learned multiple methods to improve the efficiency of edit rules and selective editing, e.g. including generally well working edit rules in EG EDIT.

16. Possibility of automatic correction is discussed and some statistics have included some automatic correction functions to their editing process. For example respondents of *Innovation statistics* often reports innovation expenditures and other money related variables in euros instead of thousand euros asked, so a few rules were created to fix most obvious decimal errors before selective editing. The effect of automatic correction is illustrated on table 1. Previously all these corrections were made manually.

Sums of variables (€)	**Unedited data**	**Automatically corrected data**	**Final data**
Variable 1	54 572 002	4 891 682	3 712 132
Variable 2	6 395 936	565 971	467 270

Table 1. Power of automatic correction in Innovation statistics (decimal error)

17. Defining parameters, assessing existing editing rules and creating new rules is usually done separately by each developer on its own and the tasks are divided by the scrum master. Developer team is able to get help and comments from statistical content experts at any time. After suitable parameters and editing rules are decided the final EG EDIT version is constructed to be as efficient as possible. The chosen parameters and rules as well as the EG EDIT version are presented to the statistics and final changes can be made if they request.

18. The main product of EG EDIT is an error list. The error list is designed together with statistical production staff and edit rules are labeled with meaningful descriptions to be presented as a part of error information. The form of the error list is primarily subjected to the wishes of statistical production staff and development team offers a lot of options to be included in error list. One very simple basis for error list design is illustrated in figure 2 (latter table), and one error list used in production is presented in figure 4. Statistics can choose the information to be presented from wide selection of variables, values, figures, scores, flags, suspicion scores and other error information. By designing their future editing tool statistical production staff become very committed to new editing practices.

19. As the methodology team improve the EG EDIT versions, the data collection expert prepares data connections and linking programs between EG EDIT and statistics own IT environment. Very often the data that is run on statistical production must be modified to be able to run in EG EDIT. Also the possible frame and sample information must be retrieved if they are not included on production data. Sometimes implementation of EG EDIT requires modifying statistics own production system (usually an in-house program) and since this task requires assistance from IT unit, the modification work is done after the two week period. As the final EG EDIT version is completed, connections are tested and test-runs are done in production environment.

20. When the data connections and all linking programs are done the simulation study is operated. This tests the connections and linking programs in production-like situation and allows the data to move back and forth from statistics own system to EG EDIT. Again, this phase varies much from statistics to statistics. Generalisation of simulation process is illustrated on figure 3 with few optional functions. UDATA is new set of data arriving from respondents. It is transferred to SAS-environment and modified to be suitable for EG EDIT. Then process of error location and score calculation starts resulting error information and error list with prioritised units for manual editing. If the statistics is using a separate error list only, nothing is transferred back to statistics own system but manual editing can be done for UDATA with help of the list.

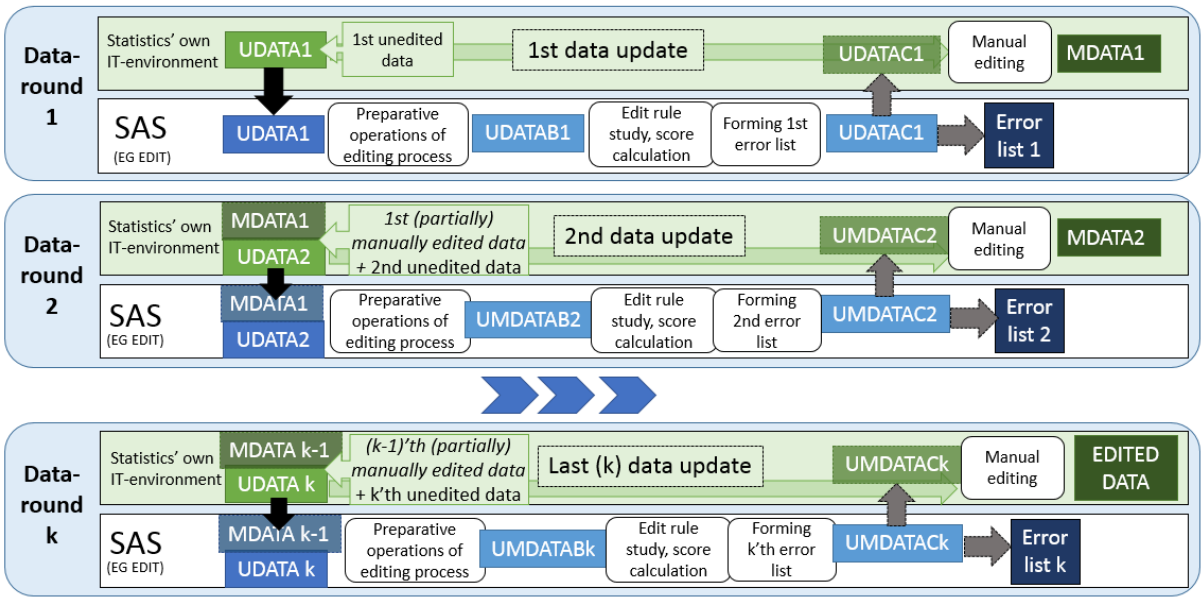

Figure 3. Simulation flowchart translated and modified from *Pauli Ollila*'s Finnish pp-slides.

21. If statistics want error information and scores to their own system, UMDATAC with scores and error information attached is transferred back to statistics own system. After manual editing there is first

(partially) manually edited data set MDATA (partially meaning that the new data set arrives before all the units subjected to manual editing are inspected and some are left unedited). For the next round, the new set of responses arrive. If error information and scores are transferred to statistics own system, previously edited data is not transferred again to EG EDIT. However, if the final result is a separate error list, the information of already edited units need to be transferred to EG EDIT so they will not come up to list again. Then also MDATA need to be transferred from statistics system to EG EDIT together with new UDATA. This goes on as many times as new responses arrive until the whole edited data is acquired.

22. For example, in *Research and development statistics* data arrives gradually as units respond and editing is started after first few responses are delivered. Since statistics production system is about to be renewed in near future, it made no sense to implement all features of EG EDIT to their production now. The statistics chose to use separate error list only for now, so arriving units are transferred to EG EDIT for scoring (UDATA in figure 3). However nothing is transferred back to statistics own system (UMDATAC leads only to error list and UDATA is manually edited). In case of *Waste statistics* the data arrives once as a whole, so there is no need of multiple rounds. Also the scores and error information is transferred back to the statistics own system (UMDATAC is manually edited).

C. Results and future plans

23. From eight implementation phases five has been completed so far. On three of the completed statistics the production phase has started and EG EDIT is in use. The main product of use is the list of prioritised units to be edited manually and information from edit rules and functions. Statistical production staff editing the data of *Research and development statistics* with new error lists thought that knowing the most influential units and focusing on them has been helpful and made the editing task clearer. Also a new person on editing staff mentioned that clear edit rule labels helped to get to know with the data and made it easier to audit the correct answers from respondent. Statistics Finland intends to gradually centralize manual editing practises on *Data collection* unit so similar error lists of different statistics may be helpful for staff who manually edit different statistics. Screen capture of a section of the error list of *Research and development statistics* is presented on figure 4.

			Previous edited HMENO var	Previous edited KMENO var	Previous edited MTKLKM var	
Scores	Error code	Error label				
pisteet	virhek	virheilm	Prel1_HMENO	Prel1_KMENO	Prel1_MTKLKM	Pr
1000553	000	SELEKT pisteet	143	5837	25	
1000553	003	0: Henkilökunnan lkm 351 vrt rekisterin sl	143	5837	25	
1000553	133	1.2: t&k-henk lkm 146 Muu t&k-henk. vain 121A. muu t&k-henk vähän	143	5837	25	
1000477	000	SELEKT pisteet	1122	7688	146	
1000477	003	0: Henkilökunnan lkm 1703 vrt rekisterin sl	1122	7688	146	
1000477	104	1A: T&k-hk yhteensä 286 naisia vain 14	1122	7688	146	
1000454	000	SELEKT pisteet	9,87654321	98,7654321	4	
1000454	133	1.2: t&k-henk lkm 20 Muu t&k-henk. vain 01A. muu t&k-henk vähän	9,87654321	98,7654321	4	
1000454	233	8: Ilmoitetun 2014 t&k:n ja tulevan vuoden arvion ero suuri	9,87654321	98,7654321	4	

Figure 4. Error list of *Research and development statistics* with labelled edit rule information and chosen variable values. Current value to be edited is visible on their own system.

24. Achieving the commitment of statistical production staff is essential for this kind of project to be successful. Explaining the idea of selective editing in common language and emphasizing the best bits of EG EDIT (e.g. prioritization, error information, error list etc.) has proven to be fruitful. Some participating statistics have been involved by their own request which is of course very good situation to start with. Also the decreasing resources and demands of higher efficiency forces the statistics to look up practices to help their work load. Even if some people might be suspicious of the new methods at first, as the implementation phase goes further the more enthusiastic they usually get, as they start to realize all the benefits of a systematic editing process. Statistical production staff have been impressed how much attention is paid to their statistics and their work. They have been generally very pleased having been in part of the project.

25. The project should reach its goal by the end of the year 2015 with implementation done in all eight statistics. Year 2016 will be the first one when all the statistics participating in the project are using EG EDIT and selective editing methods. Experts of the editing methodology must reserve some working

hours to help statistics on their first production round in case of any disturbance or uncertainties on the process. It is still unclear how the controlling of the selective editing methods should be done, but one idea is to survey some units from non-critical stream similarly that is done in UK (Lewis, 2014). With three statistics already on production these questions need to be solved quickly. Management of Statistics Finland is keen to continue the implementation work in future, but experts of editing methods would like to offer development work on editing for all statistics and improve imputation methods on EG EDIT.

26. One vision of future of EG EDIT is to make it even more generic and flexible and improve its user friendliness further. User interface should be built and all the code should be hidden. Also the implementation of EG EDIT could be done with less work if program is modified. This should be done with the help of IT department and SAS experts. One project running at Statistics Finland is defining common SAS architecture for statistics and that project has close connection to EG EDIT. The plan is to develop an add-in tool from EG EDIT to be included in SAS architecture but there is no timetable or resources yet to this kind of development work.

III. Conclusions

27. The EG EDIT program including selective editing methods of SELEKT and editing procedures from BANFF has been implemented so far in five statistics with three more to come by the end of this year. Implementation phase has been carried out with agile team work methods borrowed from *scrum* development methods. Focusing the implementation work on two week *sprint* period has proven to be successful: The work is done very efficiently, results can be taken to production immediately and the development team becomes more skilful as more sprints have passed. Participating staff from statistics have adapted to this new team work method very well and their response for renewing their editing practise have been mainly positive.

28. Participating statistics have been very different by their subject, data structure and IT environment. This has caused a challenge to maintain the two week timetable but working EG EDIT program with functional parameters and edit rules has been created to all statistics so far. Launching EG EDIT in production is not part of the sprint and so far three statistics have their production going on with new editing methods. New methods including error list with unit prioritisation has been very well received by editing staff. Possible savings in working hours or data quality improvements must be evaluated later as statistics get their figures published.

29. Future plans includes introducing EG EDIT to more statistics, either with or without selective editing methods. Improvement of imputation methods has been highly requested by many statistics so methodology experts would be keen to develop imputation methods on EG EDIT. The management however would like to see more improvements on efficiency on statistics so these hopes must be compromised with general vision. New SAS architecture is also planning to include EG EDIT as one of its tools. This requires improvements of usability of EG EDIT. Statistics Finland is revising its strategy soon and it will be interesting to see how EG EDIT and selective editing is places in new strategy.

References

Banff Support Team (2007): Functional Description of the Banff System for Edit and Imputation, Statistics Canada.

Lewis, Daniel (2014): Maintenance of Selective Editing in ONS Business Surveys, Work Session on Statistical Data Editing, Paris, France, April 2014.

Norberg A., Arvidson, G., Kraftling, A. & Nordberg, L. (2010): SELEKT - A Generic SAS™ System for Selective Data Editing, Statistics Sweden.

Oinonen, S. (2014): SAS Enterprise Guide project for editing and imputation, Work Session on Statistical Data Editing, Paris, France, April 2014.

Ollila, P. & Rouhuvirta, H. (2011): Process Model for Editing (draft), Internal methodology paper (in Finnish), Statistics Finland.

Ollila, P., Ahti-Miettinen, O. & Oinonen, S. (2012): Outlining a Process Model for Editing With Quality Indicators, Work Session on Statistical Data Editing, Oslo, Norway, September 2012.

Pyy-Martikainen, M. (2014): Renewal of Editing Practices at Statistics Finland, Work Session on Statistical Data Editing, Paris, France, April 2014.

Sims, C. & Johnson, H. L. (2011): The Elements of Scrum, Dymaxicon 2011.