9 November 2012
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Oslo, Norway, 24-26 September 2012)

Report of the September 2012 Work Session on Statistical Data Editing

Prepared by the UNECE secretariat

1. The Work Session on Statistical Data Editing was held in Oslo, Norway, from 24 to 26 September 2012 at the invitation of Statistics Norway. It was attended by participants from: Australia, Austria, Azerbaijan, Canada, Denmark, Estonia, Finland, France, Germany, Hungary, Iceland, Italy, Japan, Lithuania, Mexico, Netherlands, New Zealand, Norway, Republic of Korea, Russian Federation, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, and the United States of America. The European Commission was represented by Eurostat.

2. The agenda contained the following substantive topics:
(i) Selective and macro editing;
(ii) Global solutions to editing;
(iii) Editing and Imputation in the context of data integration from multiple sources and mixed modes;
(iii) Using metadata and paradata to analyse the efficiency of editing processes;
(iv) Software & tools for data editing and imputation;
(v) New and emerging methods;
(vi) Editing and imputation of census data.

3. Mr. Steven Vale (UNECE) opened the meeting, welcomed participants and thanked Statistics Norway for hosting the meeting. He also thanked the Organizing Committee members for their efforts in preparing the meeting. He outlined the role of the UNECE Statistical Information and Methodology Unit, and the UNECE in general. Statistical organizations are increasingly looking for ways to improve the statistical production process. Data editing is often the most expensive part of the process. It will, therefore, be a major benefit for all statistical organizations to reduce these costs. The High-Level Group for Strategic Developments in Business Architecture in Statistics (HLG-BAS) reports to the Conference of European Statisticians and is composed of 10 heads of National Statistical Offices concerned with the modernization of statistics. The topic of data editing is important to the HLG and it will continue to grow over the years.

4. Mr. Claude Poirier (Canada) was elected Chairman of the Work Session. He thanked Statistics Norway for their hospitality and creating excellent work conditions for the participants. The broad participation at the meeting is an indication of the importance that national statistical offices attach to the work in the field of statistical data editing. He also expressed his thanks to the members of the Organizing Committee, and stressed that the Work Session provides a forum for discussions on best practices and streamlining initiatives within statistical offices.

5. The following persons participated in the Organizing Committee and acted as Discussants/Session Organizers: Topic (i) – Messrs. Jeroen Pannekoek (Netherlands) and Pedro Revilla (Spain); Topic (ii) – Mr. Li-Chun Yhang (Norway); Topic (iii) – Ms. Orietta Luzi (Italy) and Ms. Felipa Zabala (New Zealand); Topic (iv) – Messrs. Rudi Seljak (Slovenia) and Philippe Brion (France); Topic (v) – Messrs. Claude Poirier (Canada) and Matthias Templ (Austria); Topic (vi) – Ms. Maria Garcia (United States of America); Topic (vii) – Mr. Daniel Kilchmann (Switzerland).

RECOMMENDATIONS FOR FUTURE WORK

6. Participants discussed the recommendations for future work on the basis of a proposal put forward by an ad hoc working group composed of Messrs. Philippe Brion (France), Gergely Horvath (Hungary), Elmar Wein (Germany) and Ms. Pilar Rey del Castillo (Eurostat). When preparing the proposal, the working group took into account suggestions made by other participants in side discussions during the meeting.

7. Participants considered that there are many issues that would deserve consideration at an international forum like the present Work Session. They recommended, therefore, that a future meeting on statistical data editing be convened in about 18 months' time, subject to the approval of the Conference of European Statisticians and its Bureau.

8. The future work group drew up a list of possible topics that was modified slightly during the discussion, resulting in the following proposal for the topics to be discussed at the next Work Session. The countries and organizations in brackets expressed a possible interest in contributing to these topics:

(i) Selective editing / macroediting (United Kingdom, Sweden, New Zealand, Slovenia, Netherlands, United States, Canada);

(ii) Other methods of editing – business registers infrastructures and in the context of multiple sources (Germany, Norway, Canada, France):

- Back to the question: what the concept of error consists of?
- Simple data editing
- Sub annual data editing – question of timeliness poses constraints
- Categorical variables

(iii) Getting the support of all people when implementing data editing (Finland, France, Norway, Canada):

- Quantitative assessment of the impact of data editing: including the question of costs and study of the total survey error?
- Use of paradata and metadata
- Getting the support of the editors staff
- Getting the support of top-level management
- A general question: industrialization within the global architecture of the information system of a statistical office.

(iv) New and emerging methods (France, Sweden, Netherlands, United Kingdom, United States, Mexico):

- Editing of data when using a direct access to company information (consumer price index, accounting variables)
- Electronic collections

(v) Editing of Census and social data (United Kingdom, Switzerland, New Zealand, France, Canada);

(vi) International collaboration and software and tools (Canada, Netherlands, UNECE, Slovenia, New Zealand, Eurostat):

- building tools
- defining common frameworks
- the industrialization of software

9. The representative of INSEE, France offered to host the next Work Session on Statistical Data Editing in Paris to be held in spring 2014.

FURTHER INFORMATION

10. The conclusions reached during the discussion of the substantive items of the agenda are contained in the Annex. All background documents and presentations for the meeting are available on the website of the UNECE Statistical Division (**http://www.unece.org/stats/documents/2012.09.sde.html**).

ADOPTION OF THE REPORT

11. The participants adopted the present report before the Work Session adjourned.

Annex

Summary of the Main Conclusions Reached at the Work Session

I. Selective and macro editing

Discussant: Jeroen Pannekoek (Netherlands)
Documentation: Papers by Spain, Italy, United States of America, Sweden (2) Germany and Italy

1. The objective of macro editing and selective editing is to limit time-consuming and costly manual editing (reviewing, treatment or re-contact) as much as possible without substantially decreasing output quality. Both macro editing and selective editing are selection methods, with the common purpose of selecting units with potentially influential errors for interactive (manual) review. Papers in this session covered applications and extensions of traditional selection methods and discussed issues in their implementation.

2. The presentation by Spain on optimization as a theoretical framework to selective editing, proposed to minimize the number of units to edit subject to a bound on mean squared measurement errors of non-edited units. They defined selective editing as an optimization problem where you have to choose the minimal subset of the records for editing such that the remaining measurement error is below a specified fraction. They are developing R packages and SAS macros to select and prioritize units according to this approach and adapting the approach to edit qualitative variables also.

3. Italy presented multivariate selective editing via mixture models and discussed the first applications of this approach to Italian structural business surveys. They showed the application of a scoring approach based on a mixture model to structural business statistics, based on the SeleMix software. The possible benefits were explained in terms of estimate accuracy and burden/costs reduction from the use as auxiliary information of external sources. Encouraging results have been obtained so far.

4. The United States Census Bureau presented an application of selective editing to foreign trade data. This includes a scoring method for selective editing of foreign trade data and an evaluation of pseudo bias. In traditional selective editing methods, data from the previous cycle are used to construct score functions; this is not possible for trade data due to large variations between periods. They adapted available score functions to using only current cycle data by computing an estimate of the anticipated value of the variables.

5. Statistics Sweden described a “tree analysis” approach – a method for constructing edit groups where soft edits are used to detect suspicious values. Large amounts of data are partitioned into relatively homogeneous groups in an iterative process, resulting in “branches” and, ultimately ending with “leaves”. The tree predicts a value for a new observation, based on the leaf that this observation would belong to if it were in the data set.

6. The Federal Statistical Office of Germany presented an automated comparison of actual statistics with plausible reference statistics (statistics of the previous period) for editing structural business survey data. There is an automatic checking of aggregates and flagging of suspicious records (drilling-down). This is based on principle components instead of original variables. A principal component analysis is used because it provides the opportunity to reduce the dimensionality problem as regards the detection of outliers.

7. A second presentation from Sweden reviewed implementation issues of selective editing, especially with regard to determining an acceptable level of selective micro editing. The aim of the presentation was to raise the question of what type of data material is suitable for evaluating how to set up the selective editing. The problems were based on implementations of the generic tool SELEKT but should be generally applicable in many types of selective editing.

8. Italy presented a second paper on selective editing as a part of the estimation procedure using the SeleMix approach to selection. Sampling unedited units, with probabilities proportional to scores, allows estimating bias, and correcting for it. They found that when data contain a few gross errors, selective editing

with SeleMix performs better than the two-step approach. In the presence of many small errors (inliers) bias correction can improve the estimate provided that a sufficient number of units are sampled.

9. Points raised in the general discussion of the model-based papers included:

- The importance of considering components of composite variables such as turnover. The application of selective editing methods to semi-continuous components is not straightforward, but a two-step approach could help.
- Prioritization by gradually decreasing the bounds to identify the highest priority units.
- Selective editing needs a good model. It doesn't work very well for unpredictable variables such as investments.
- Selective editing cannot be standardized. You have to adapt the method to the specific data sets and variables that you are working with.
- Selective editing models can be influenced by the errors in the data.
- It is useful to have a double data set, containing raw and edited data, to assist with developing an editing and imputation strategy and to fine-tune parameters.
- Whether the mean-squared error, or the number of records should be the main constraint for selective editing.

10. Following the presentations of innovative applications the discussion focused on the following issues:

- Evaluation data sets, which have been 100% edited, are very useful to determine the efficiency of different editing models, but can be very expensive. Can samples be an acceptable substitute?
- Learning about how respondents provide data can help to predict error patterns.
- In the tree analysis method, the errors seem to be easier to detect whereas selective editing looks at where the gains will be. How to reconcile these two aspects?
- The tree approach identifies groups of near neighbours to be used for predictions.
- All methods identify suspicious records, but how many of these records end up being adjusted? A 60% hit rate was reported by Sweden to be a target for edit rules.
- When suspicious records are found to be correct, a text comment on the reason can be very valuable for data users.
- There are many different approaches, but selective editing remains very relevant.

II. Global solutions to editing

Discussant: Li-Chun Zhang (Norway)
Documentation: Papers by New Zealand (2), Netherlands/Norway, Australia/UNECE, Sweden, Eurostat and Canada

11. This topic addressed the importance of developing stronger international collaboration to improve the efficiency of statistical production in view of the considerable resource constraints that national statistical offices all around the world face. A variety of directions can be followed. Concepts and methods are generally more portable between organizations than tools and systems. While the core of our work is about meeting short-term data requirements, there is a long-term obligation to store data in standard ways for future generations.

12. Statistics New Zealand presented work being done on the review of the UNECE *Glossary of Terms on Statistical Data Editing*. It is a product of collaborative work among participants of the Work Sessions on Data Editing over many years. It currently contains over 200 terms related to statistical data editing, including terms associated with collection, processing, and dissemination of data impacted on by editing. Participants were invited to give feedback on the proposed additions, changes and deletions of concepts.

13. Statistics Netherlands and Statistics Norway presented a view on the general flow of editing, taking into consideration some recent developments in the theory and practice of editing. The work flow consists of a breakdown of the overall editing process into sub-processes. Each sub-process involves up to 3 generic types of editing activities with its own purpose and ordering, and is composed of one or several statistical functions. The description of the general editing flow was illustrated and discussed using the editing practice of the Structural Business Statistics (SBS) in Netherlands and Norway.

14. The UNECE provided an update on the on-going development of the Generic Statistical Information Model (GSIM). GSIM is a cornerstone of the vision for the modernization of official statistics, endorsed by the Conference of European Statisticians in 2011. It is a model to describe information objects and flows within the statistical business process. Over 150 information objects have been identified so far. GSIM v0.8 has been released and will be followed by version 1.0 by the end of the year.

15. Sweden presented two paradigms for official statistics production which concern data and knowledge about the external world – not data and knowledge about producing statistics (but there could be consequences for the latter). This approach was inspired by the different discussions on on-going developments and initiatives within official statistics and may be relevant for editing. Paradigm 1 relates to a stovepipe approach, with editing for a specific purpose, whereas Paradigm II relates to an integrated system with automated editing for general use.

16. Eurostat presented a proposal of a revised approach for data validation within the European Statistical System. The issue of quality assurance is becoming increasingly important. Eurostat has undertaken a Vision Infrastructure Project on validation since the end of 2010. It will provide data validation solutions for use by both Member States and Eurostat.

17. The representative of Statistics Canada presented the development of a data editing and imputation tool set that identifies common requirements for both statistically collected data and administrative data. Dimensions of quality are defined in order to evaluate potential solutions. The desired features of a tool set are: functionality, quality criteria (relevance, accessibility, interpretability, coherence, accuracy and timeliness), and important software characteristics (adaptability, reliability, maintainability, and interoperability). The basic tool set proposed is composed of BANFF, CANCEIS and SELEKT. Other tools will be assessed to try to fill some gaps in functionality.

18. Statistics New Zealand presented their developments in statistical data editing. A new infrastructure was needed to produce statistics that are fit for purpose in a cost-effective way. Challenges encountered and lessons learned include: Balancing generic and specific needs; determining process elements suitable to be implemented as common services; moving from a process culture to a more constructive innovative culture; and the adoption of an “agile” project management approach for IT development projects. Future work includes a review of their generic business process model, implementation of a framework to measure and report on benefits from recent and on-going developments, and transformation of data collection processes.

19. The points raised in the discussion included:

- Are we getting anywhere *together*?
- What is the mandate / purpose of official statistics?
- Who are the users: external / internal?
- We have to keep end-users information about the methodology. The glossary is a useful tool for this, and will be developed according to country requirements.
- Is it important to harmonize terminology, or just the underlying concepts?
- GSBPM is intuitively understandable. GSIM is perhaps not so easy. We should work more on communication to make it clearer and show how it can lead to standardization and modernization. Task teams have been established to improve communication and develop a user guide.
- Both GSBPM and GSIM are necessary. We should shift towards components and granularity. It will make it easier to extend them to produce agency-specific models. This is the future of re-use and sharing.

- Do we need unique frameworks and models covering all kinds of statistics?
- Does "Grand Unification" provide a roadmap for standardization?
- ISO standards should be used to define quality and portability for software.
- Do we want to model everything or just concentrate on certain aspects? What is sufficient and what is necessary? The S is important in both GSIM and GSBPM.
- GSIM deliberately doesn't try to model quality frameworks, methodologies or reference metadata, as there are no global standards in these areas. Instead it aims to be easily extendible to cover specific methods and frameworks as necessary.

20. Participants agreed to provide feedback on the proposals to revise the Glossary to Felibel Zabala by the end of November.

III. Editing and Imputation in the context of data integration from multiple sources and mixed modes

Discussants: Orietta Luzi, Istat (Italy) and Felipa Zabala (New Zealand)
Documentation: Papers by Norway, New Zealand, United Kingdom (2), France, Hungary, United Arab Emirates, Netherlands and Italy

21. Integrating data from multiple sources (including administrative data) is becoming progressively more common in official statistics since it allows national statistical organizations to provide more information of good quality while reducing production costs and respondent burden. This can be considered as a "special case" of mixed-mode data collection where information on target variables is captured via different "tools" (e-questionnaires, paper questionnaires, direct access to companies systems, use of administrative data etc.). In these situations, the difference in quality of multiple sources/mixed modes data is a basic issue: the editing strategy needs to cover data from each source/mode, as well as to ensure the coherence of the integrated sources.

22. The issues covered in the presentations included:

- The quality of data from various sources and modes: how it compares across sources and over time.
- Issues emerging in the context of data integration from multiple sources and modes, to be aware of when developing editing strategies and systems.
- The impact on editing and imputation when integrating new types of data sources (including administrative data) into statistical production processes.
- Strategies for better alignment of statistical and other data sources, to improve the quality of output statistics.

23. Statistics Norway presented a paper on micro integration of register-based census data for dwellings and households. They proposed statistical methods and approaches to create a complete census data file of linked dwellings and households which meets the need for detailed tabulation and analyses required by the 2011 register-based population and housing census in Norway. The method of double nearest-neighbor imputation provides a general approach to the problem of micro-linkage between different types of units for which direct matching is impossible.

24. Statistics New Zealand presented the issues and solutions relating to the development of their Integrated Data Infrastructure (IDI) covering different statistical areas and allowing for complex multivariate analysis of social and economic phenomena. The editing strategy in the IDI aims to treat inconsistencies for the same unit from different sources, treat erroneous and missing variables in a record, and ensure consistency in variables across a record for a time period and over time. The planned next steps include extending the IDI and determining standard quality measures.

25. The presentation by the United Kingdom highlighted the editing challenges of using different data collection modes and focused on how to deal with edits in electronic questionnaires. It discussed the experimental design used to test edits in electronic questionnaires in terms of how to present edit checks to respondents, and the extent to which they should be used.

26. The presentation by INSEE France built on the presentations made at previous Work Sessions about Esane, a multi-source device for structural business statistics. It is based on the combined use of administrative and survey data. The presentation focused on the issues of field definition and the consequences of the use of multiple sources on estimation.

27. Hungary presented their experience of substituting a regular statistical survey with administrative data in the area of health statistics. An annual survey was compared with an administrative data flow. They found that despite reducing the administrative burden on a targeted group of data providers and obtaining overall improvements in data quality, the risk of introduction of biasing effects remains.

28. The representative of the Statistics Centre Abu-Dhabi (SCAD) described initiatives to develop mixed-mode data collections and automated error detection to enhance data quality. They found that automated data editing improved the quality and efficiency of surveys, and are investigating methods to handle missing and anomalous data in establishment surveys.

29. The United Kingdom presented methods for making greater use of administrative data in the production of business statistics. They described and evaluated the methods for detecting and imputing errors in VAT Turnover data to be used in the production of mixed-source short-term turnover estimates. The choice of methods to detect and correct errors in administrative data will ultimately depend on the data used to produce mixed-source statistics, especially in terms of coverage, timeliness and accuracy.

30. Statistics Netherlands presented joint work with several other European statistical organizations on imputing missing values when using administrative data for short-term enterprise statistics (STS). Analyses in Italy, Germany and Finland show that main revisions in STS-admin data estimates are caused by uncertainty of the units in the population. They recommended that research and development should focus on choosing the most adequate estimation method for the active population.

31. Italy presented improvements in the timeliness of the Italian Business Register through the imputation of missing information in administrative data. Provisional rather than final data can be used, if there is a sufficient mechanism to impute missing information. In order to validate the approach used, the two versions of the administrative data were compared and a good level of accuracy was obtained.

32. Points raised in the general discussion included:

- The importance of outlier treatment in administrative source files.
- The degree of follow-up on data quality issues with administrative source providers, and whether this can be at the level of aggregates or microdata.
- The impact of switching to administrative sources on different aspects of quality, particularly accuracy and timeliness. Timeliness often depends on the collection method of the administrative source.
- Contrary to expectations, it may be possible to influence the collection and content of administrative data.
- However, the extent to which statistical organizations determine the accuracy of administrative data remains an issue.
- When there are differences, survey data may not be any more accurate than administrative data.
- Why impute for missing values in administrative data sets, as they are often more complete than survey data? A more complete data set can increase flexibility and can be useful for determining the true population.
- Model-based approaches for the treatment of missing data can be an alternative to unit-level imputation.
- The quality of administrative data may be more dependent on issues of units than variables.
- Standards and guidelines on managing relationships with administrative source providers are being developed in some countries. It could be beneficial to share and “internationalize” these.

- There are emerging problems for editing and imputation linked to the additional types of errors arising from the use of integrated data sources.
- Longitudinal analysis of data poses additional problems in guaranteeing coherence.
- There is a need to share experiences on editing and imputation with multiple sources, to try to create a new framework and guidelines.
- Understanding the boundaries between the collection and editing processes becomes more important when multiple sources are used.

IV. Using metadata and paradata to analyse the efficiency of editing processes

Discussants: Rudi Seljak (Slovenia) and Philippe Brion (France)
Documentation: Papers by France, Finland, Canada and Sweden

33. According to the OECD glossary of statistical terms, metadata are data that define and describe other data. Reference metadata are metadata describing the contents and the quality of the statistical data. Paradata are automatically generated process data including audit trails and contact histories.

34. Metadata and paradata are becoming increasingly important in the statistical editing process with the advances in methods and automated applications. They are used especially as a source of information to analyse the adequacy and efficiency of processes, and to detect weak points and support continuous improvement.

35. France presented improvements to the data editing process based on metadata and paradata relating to the results of the selective editing process. This includes information on warning messages, the number of suspicious units and feedback from editing staff. Information coming from metadata and paradata could also be used for other methodological improvements, such as changes to questionnaire contents and structures. It can also be useful when explaining processes to users of the data, such as National Accounts staff.

36. The presentation by Finland described a process model for statistical data editing, together with ideas for indicators to help monitor the processes. Over 50 indicators are considered, but some are suitable only for specific types of statistics. It is, therefore, not possible to define a detailed list of indicators to be published with all statistics, but there are a few standard indicators for editing processes that should always be computed, for example concerning coverage and edit rates.

37. Canada presented the results of an experiment to test different approaches to improve the efficiency of the non- response follow-up for electronic questionnaires. From a collection perspective, the design of the experiment showed that significant response can be obtained without making phone calls for the first few months of collection for annual surveys. In addition, more frequent e-mail reminders are beneficial for obtaining responses. The strategy consisting of e-mail reminders every two weeks at the beginning of collection succeeded in obtaining the first 45% of response at lower cost. On the other hand, there is a point in time when it becomes necessary to start telephone follow-up, but starting this process later does not appear to impact negatively the final survey response rates.

38. Sweden presented an approach to gather qualitative information about editing through de-briefing sessions for editing staff. The approach uses interviews with editing staff, similar to focus groups, to provide qualitative paradata. These debriefings can identify various problems with the measurement instrument and the editing process, and can serve as a basis for further improvements. The editing staff often has useful ideas about what causes problems or errors. Considerable information is obtained at a modest cost, complementing information obtained by other means.

39. Points raised in the general discussion included:

- The possibility and desirability of developing a common framework of indicators about the data editing process. GSIM and GSBPM should be taken into account if such a framework is developed.

- Are editing staff debriefings used several times for the same survey? Some may have been repeated for one specific survey. The aim of de-briefings of editing staff is to obtain information to be used for expert review or cognitive test to make improvements to the process. They could be repeated at 5 year intervals.
- An important purpose of editing is not just to correct the data but to learn from the errors.
- In some countries, editing staff are reluctant to adopt selective editing. Training and support are essential. Focussing on the most important errors using paradata and metadata can make their work more meaningful.
- Metadata and paradata can be collected at the micro and macro levels, and for elements of the statistical infrastructure, such as registers, but how to decide what to collect and what to keep?
- The human component is an important aspect – how people respond to surveys. Statistical organizations should adapt to the needs of respondents by talking directly to them. This will improve the quality of data at the source.
- Different types of users want different metadata and paradata. For internal users, the aim is to improve the efficiency of the production process; for external users, the aim is to provide information about data quality
- A database of process metadata / paradata could be useful to help analyse production processes and answer questions about output quality.
- Some paradata are published via quality reports, others could be added.
- Process owners should be clear about the purposes for which metadata and paradata will be used.
- Studying industrial quality control methods and models could be useful.

V. Software & tools for data editing and imputation

Discussants: Claude Poirier (Canada) and Matthias Templ (Austria)
Documentation: Papers by the United States of America, Netherlands (2), Austria, UNIDO, Lithuania and Japan

40. Most of the presentations in this topic were based on tools developed on the R platform. R can also be a mediator that allows communicating with other platforms and databases.

41. The representative from the U.S. Census Bureau presented TEA – a generalized system to unify and streamline demographic survey processing from raw data to editing to dissemination of output. It is available as an unofficial R package, and has been used by the U.S. Census Bureau for processing several surveys. TEA provides a single platform for editing, inference control, and imputation, but keeps these procedures separate. The modeling framework offers flexibility, and makes imputation easy, while the R platform facilitates a range of visualization techniques.

42. The Netherlands presented two innovative visual tools for data editing – `treemap` and `tableplot`. `treemap` is a visualization method for hierarchical data, for example business data at different levels of an economic activity classification. `tableplot` can be used to detect outliers and unusual patterns in large data sets, and to monitor data quality during data editing and imputation. Both tools are implemented in R and are freely available on CRAN (the Comprehensive R Archive Network).

43. The presentation by Austria discussed the interactive adjustment and outlier detection in time dependent data in R. The X12-ARIMA seasonal adjustment software of the U.S. Census Bureau is wrapped in a graphical user interface (GUI) developed in R. Using the `x12` R package offers the possibility to employ R for pre- processing time series data, for managing `x12-arima` parameters and output and for presenting diagnostics in a more approachable and accessible manner. Very little prior knowledge of R or the respective packages is required for using the GUI.

44. Austria also presented joint work with United Nations Industrial Development Organization (UNIDO) on current and potential screening methods in the UNIDO database. The methods discussed varied from logical relationships in the data to highly sophisticated outlier detection procedures. While the methods are discussed in detail in the paper, the presentation focused on visualizations. One conclusion was

that with a data set of mixed quality, sometimes very simple screening methods can be more effective than very sophisticated ones.

45. A second paper by the Netherlands presented automatic data editing with open source tools. Several capabilities of R extension packages `editrules` and `deducorrect` were shown. The `editrules` package has been set up as a toolbox for edit rule management and manipulation. Future work will focus on the application of those packages in actual statistical production processes.

46. To improve and standardize the work of statisticians at Statistics Lithuania, a SAS Macro program was developed for editing and imputation. It consists of five parts: detection of errors, detection of outliers, imputation using the nearest neighbor method, imputation using models, and imputation using distributions. It will be extended to offer additional functionality.

47. Japan presented work on the multiple imputation of turnover to improve data quality in the economic census. Standard single imputation techniques have various limitations, which can be overcome using multiple imputation. The R-based package `ameliaII` (a general-purpose multiple imputation tool) has proved successful for implementing multiple imputation in large data sets. The impact of outliers on imputation will be considered in future research.

48. The following points were raised during the discussion:

- At the prototype stage it is important to have realistic objectives and a well-defined scope.
- At the production stage, user confidence will be greater if there are visible maintenance and related research activities.
- R or SAS – why use one and not the other? Other alternatives include general purpose languages like Java, C, C++ and Python.
- The human factor is important, many statisticians are familiar with SAS, but increasing numbers of students are learning to use R. A pragmatic solution may be to use both in parallel, at least in the short-term.
- R is not a package but an environment. It can be used with packages developed in other languages.
- One strength of R is the sheer number of statistical packages now available.
- New environments such as Julia offer greater speed than R, but are unfamiliar to most statisticians, and it takes a lot of effort to change from one environment to another.
- The official community can leave the choice of platforms to evolve gradually, unless instructed differently. This would require a commitment from statistical organizations, but would result in increased portability of solutions.

VI. New and emerging methods

Discussant: Maria Garcia (United States of America)
Documentation: Papers by Sweden/Estonia, Eurostat, Slovenia and Netherlands

49. Papers under this topic presented new ideas and advances in the development of methods and techniques for solving, improving, and optimizing the editing and imputation of data. Contributions covered probability editing, machine learning methods, model-based imputation methods, and automatic editing of numerical data.

50. The presentation by Sweden/Estonia proposed selecting units for editing using a probability sampling framework. Their method applies to all types of data and using this approach makes it possible to address the statistical properties of the resulting estimators which is not possible with selective editing alone. In addition, bias due to measurement errors is eliminated when using probability editing.

51. Eurostat described the use of machine learning methods to impute categorical data using either a neural networks classifier or a Bayesian networks classifier. Comparison studies of machine learning methods results with results obtained from logistic regression and multiple imputation showed that machine

learning methods are easier to automate and can offer substantial improvements over the traditional method. The machine learning approach also seems to be scalable to very large data sets.

52. Slovenia described their implementation of a Bayesian approach to imputation. The Bayesian model and linear regression are combined into a method for imputing annual gross income in a household survey. They also replicate their method multiple times to properly account for the variance due to imputation. However, the method gives weaker results when the data are less successfully described by the data partitioning within the selected model.

53. Statistics Netherlands has recently developed a new automatic editing method that can take soft edits into account when finding solutions to the error localization problem. A prototype application has been developed in R that uses an existing R package for error localization. They described some of the first empirical results obtained with this new method. Using soft edits improves error localization, but more work is needed to further investigate the possible added complexity and the effect on the quality of the solutions.

54. The following points were raised during the discussion:

- The approach described by Eurostat used several different platforms.
- How to evaluate the performance of multiple imputation.
- Sound statistical methods should include both point estimates and measures of uncertainty.
- We are more familiar with traditional statistical models, but nothing prevents us from considering different approaches such as machine learning methods.
- The extent to which probability editing can be seen as similar to two-phase sampling methods.
- The approach in the Swedish paper can be applied to all types of data in theory, but there may be issues of determining scores for categorical data.
- Whether macro editing is still necessary in relation to the approach described by Sweden.
- Whether categorical variables are a real problem, as they are often combined with numerical variables in practice.
- Whether selective editing is appropriate for categorical data.
- Issues relating to the combination of soft and hard edits. The approach described by the Netherlands allows identification of the variables to change to satisfy the edits.
- Coverage rates for multiple imputation, and the possibility of a model to predict missing data.
- For some complex imputation problems (i.e. large number of variables, different types of variables, large data files) the default settings within commercial software procedures may be inappropriate. What type of diagnostics, graphical or analytical, should be examined to ensure the procedure is working properly?
- Adding soft edits increases computational complexity, but this has not yet resulted in problems in the Dutch work.

VII. Editing and imputation of census data

Discussant: Daniel Kilchmann (Switzerland)
Documentation: Papers by Slovenia, Austria, United Kingdom (2), United Arab Emirates and Mexico

55. This topic focused on the methodological advances of editing and imputation techniques applied to census data. A number of statistical organizations are replacing “classical” census data collection, fully or partially with the use of administrative or register data, with significant effects on their editing and imputation strategies.

56. The presentation by Slovenia concerned editing of multiple source data in the 2010 Slovenian agricultural census. This census was based on a mixture of survey and administrative data, managed using a metadata-driven process, and a database approach to integrate the different sources. Although the increased amount of editing required meant that overall costs were fairly similar to previous censuses, this approach led to clear improvements in quality, as well as reduced response burden and information technology costs.

57. Austria presented the data imputation process of their register-based population and housing census. The 2011 census was the first based entirely on registers, and delivered major cost savings compared to previous censuses. A major challenge was developing prioritization rules for reconciling variables that are available from multiple sources. Methods used for editing and imputation included deterministic and distributional imputation (kind of hot-deck imputation) underpinned with correlation, clustering, and logical regression techniques. Quality indicators were derived for each attribute in each register.

58. Two presentations from the United Kingdom outlined the practical implementation of census imputation methodology and item imputation of census data in an automated production environment. The editing and imputation package CANCEIS was used for the first time in the 2011 census. A JAVA "wrapper" was used to move data between CANCEIS modules. This approach was more flexible and much faster than previous tools, reducing manual editing and imputation. Additional features in the most recent versions of CANCEIS should help to further improve imputation in future censuses.

59. The Statistics Centre of Abu Dhabi presented their experiences of editing and imputation in the 2011 Census, which was the first conducted by that agency. CANCEIS was successfully implemented, but societal differences in household compositions (large households, multiple wives etc.) meant that several common edit rules had to be relaxed or re-defined. Overall the use of donor and deterministic imputation kept the need for manual intervention to a minimum.

60. Mexico described their experience in editing census data with the assistance of geographic information systems. The latest census used a traditional data collection approach with six kinds of questionnaires. One of these covered urban environment issues, and was particularly well suited to the use of mapping techniques to identify potential errors and inconsistencies in a spatial context.

61. The following points were raised during the discussion:

- Census data have many uses. The application of selective editing techniques may not be optimal for all of these, for example use as a sampling frame, or re-use of microdata for research purposes. Automatic editing could be an alternative.
- A fully automated data editing system may not be feasible for censuses as there are always special sub-populations that need individual treatment.
- Fully automatic editing and imputation may reproduce already existing noise in datasets. However, score functions and macro-editing can indicate strange things that automatic editing has done. The treatment of the noise could therefore be performed after the fully automatic editing and imputation.
- The amount of editing and imputation needed when working with registers should not be underestimated.
- SAS can be more appropriate than CANCEIS for ad hoc deterministic editing, as it is quicker to develop the necessary routines.
- Under and over coverage for specific geographical areas can be a problem with a political dimension. The coverage of registers and other sources should be checked.
- The stability of registers over time could affect their suitability as a source for census data. Close cooperation with register authorities can reduce the risk.
- Lack of consistency between sources remains an important challenge in multi-source data collections and leads to the need of non-trivial prioritizing of the sources in order to retrieve one value per variable under the condition of maximum consistency between the variables.

* * * * *