

## Topic ii – Implementing Editing strategies and links to other parts of processing

**Discussants:** Orietta Luzi (Italy) and Carsten Kuchler (Germany)<sup>1</sup>

### Introduction of the topic

Under this topic papers focusing on the integration of single editing techniques with respect to comprehensive data editing strategies, like particularly reflected by the (re-)design of editing processes, were mainly solicited. Contributions considering the impact on data quality, on the organization of work and application flows or management issues were expected. In order to share experiences, reports of ongoing or already finished implementation projects were especially welcomed. The working session especially encouraged contributions focusing on issues concerning the impact of editing strategies on related stages of the survey process, such as:

- survey design (adapting sample design, improving questionnaire design, harmonizing and optimising editing at the data capturing stage, efficient use of auxiliary information and external data sources in data editing processes, ...)
- advanced estimation methods (modelling background knowledge from the editing process, assessing the impact of editing on the statistical properties of estimators, ...)
- release of micro data (supporting disclosure elimination with information from data editing, providing background knowledge to users, multiple imputation, ...)

The 11 invited and supporting papers received under the topic discuss the above-mentioned issues from theoretical, methodological, technical and operational points of view. Given the heterogeneity of the subjects involved in the topic, for a more structured discussion papers have been grouped depending on their focus on some sub-areas that can be identified in the large research and application field covered by the topic. In particular, the following four sub-topics have been identified:

1. *Sub-topic a – Links to other parts of processing: editing and variance estimation*
2. *Sub-topic b – Links to other parts of processing: editing and data dissemination*
3. *Sub-topic c – Implementing editing strategies: (re-)designing editing processes*
4. *Sub-topic d – Implementing editing strategies: using external data sources*

For an overall scheme of sub-topics and the corresponding papers, see annex 1 at the end of this document.

### Sub-topic a – Links to other parts of processing: editing and variance estimation

#### *Overall Summary*

Under this sub-topic two contributions are grouped:

- WP 10 “*Assessing and dealing with the impact of imputation through variance estimation*” by Eric Rancourt (Canada);
- WP 19 “*Performance of re-sampling variance estimation techniques with imputed survey data*” by Dolores Lorca and Félix Aparicio-Pérez (Spain).

Both papers deal with the problem of variance estimation in presence of non-response and imputation. The advantages of using imputation to compensate for (partial) non response are well known, but if imputed data are used as if they were actual responses, the estimate’s variance obtained using standard formulae will result underestimated. Therefore, for many imputation methods the variance of estimates will be inaccurate and the confidence intervals invalid. Measuring the non-sampling components of variance due to non-response and imputation is becoming a mandatory requirement in NSI’s in order to allow final users for correct analyses and inferences. A number of methods and approaches have been proposed in literature to deal with this problem (among others, re-sampling techniques like jack-knife and bootstrap, two-phase sampling, multiple imputation and so on). Recently, resources were concentrated on the development of standard frameworks

---

<sup>1</sup> Prepared by Orietta Luzi (luzi@istat.it)

and tools for allowing statisticians and subject matter experts for assessing the estimates variance in a standardised way under different scenarios, hence for assessing the quality of their own imputation strategies.

Under this issue, the Canadian contribution illustrates the framework developed at Statistics Canada to allow statisticians to measure the quality of imputation methods and assessing the accuracy of estimates computed on them. Generalized tools supporting researchers and statisticians in assessing the impact of imputation under different theoretical and practical scenarios are illustrated.

The Spanish paper illustrates the results and the outcomes on the application of some re-sampling techniques for assessing the impact of imputation on survey estimates and illustrates the main results obtained.

### ***Main Issues***

- Estimating variance under imputation allows for gathering information on data and process reliability, thus allowing for several specific analyses (Rancourt, 2005): more precise estimation of total variance, the possibility to make correct inferences, better quality reporting to users, evaluation and choice of the better imputation models, production of separate estimates of sampling and imputation variances to plan/adjust available resources between the sample size and the imputation/follow-up effort.
- Concerning imputation, its impact may be either direct (on the process, on estimates) or indirect (on the next designs, on secondary analyses). Developing standard strategies and quality measures for assessing imputation impact is essential regardless of the adopted model/strategy. If there are competing models method for treating non-response, estimating the non-response variance can be used as a criterion to make a decision on the “best” non-response treatment to be adopted.
- In order to statistically assess the quality of imputation all the stochastic mechanisms affecting the survey results (sampling, non-response, editing, imputation) are to be evaluated in order to correctly measure the resulting quality of estimates and make correct inferences.
- In general, the availability of an evaluation framework allows developing better imputation strategies. Statisticians can verify the reliability and the usefulness of the adopted methodologies and assumptions, in order to plan revised or new approaches for improving the performance of imputation.
- When estimating variance in presence of imputation a method has to be chosen among the available approaches. Some basic elements to be taken into account are:
  - the complexity of the design and estimator;
  - the complexity of the imputation strategy;
  - the approach implemented in production under full response;
  - the need for separate components for sampling and imputation variance;
  - the users characteristics (group producing imputation,, external users);
  - the simplicity
- An important aspect in this area relates to the methodological and operational effort needed at Statistical Agencies for designing and implementing tools allowing statisticians/data producers for easily evaluations in a standardized way. This effort implies a re-organization of both the data production and the data dissemination processes to integrate production/analysis/release of the additional information in the current processes and data dissemination plans.
- Estimating the variance in presence of imputation is only one dimension of measuring the overall quality of survey estimates. Other aspects need to be investigated, due to the other data processing activities generally performed at the editing and imputation stage (e.g. manual imputation, editing).

### ***Points for Discussion***

- How to choose among different imputation strategies and variance estimation approaches when dealing with the problem of minimizing the effects due to non-response, non-sampling errors and data processing on survey results.
- How to evaluate the impact of imputation on multivariate data and relationships.
- How to evaluate the impact of imputation on hierarchical data.

- How to evaluate the impact of imputation on matched or linked data.
- How to evaluate the impact of imputation in case of complex sampling designs.
- How to evaluate the impact of imputation in case of combined use of different imputation techniques.
- How to estimate variance in presence of both editing (manual and/or automatic) and imputation.
- In many Statistical Agencies currently adopted tools for variance estimation does not take into account non sampling components like editing and imputation. Managers and subject matter experts are to be convinced about the need of changing the usual way of estimating variance, methods and tools are to be provided to subject matter experts, and they are to be trained on them.

## *Summary of Papers*

### WP.10 Assessing and dealing with the impact of imputation through variance estimation (Canada) (Invited paper)

#### *1. Summary*

The paper contains an overview of methods and tools developed at Statistics Canada to provide statisticians with a standard framework to cope with the problem of assessing the estimates variance under imputation. In particular, the structure and functionalities of the software GENESEES (Generalized Simulation System) and SEVANI (System for Estimation of the Variance due to Non-response and Imputation) are illustrated. GENESEES allows statisticians for performing simulation studies in the presence of imputation. It contains a different modules for: sampling design (involving the most common sampling designs and, for several designs, the Horvitz-Thompson, ratio and regression estimators); simulation of missing data under different mechanisms (MAR, MCAR, NMAR); imputation under different models (Previous value, mean, ratio, regression, random hot-deck, nearest-neighbour); variance estimation under different approaches (depending on the imputation model and the non-response mechanism, the two-phase approach and the reverse approach; Monte Carlo evaluation measures are also available); imputation/re-weighting class identification (cross-classification, score method, together with Monte Carlo evaluation measures).

The SEVANI software allows for estimating the non-response and imputation variance portions in a survey context when a domain total or mean is estimated under some specific imputation models. Variance estimation is based on the quasi-multi-phase framework. Non-response variance associated to different errors models can be estimated. Variance estimation under different treatments (weightings or different imputation models, like deterministic linear regression, random linear regression, auxiliary value and nearest neighbour) can be obtained. The Author also proposes possible extensions of the evaluation approaches described in the paper to other fields: evaluating the quality of manual imputation, evaluating the quality of editing.

Through the description of methods and tools the Author discusses some relevant aspects about the following aspects:

- which method(s)/tool(s) should be used to perform variance estimation under imputation;
- which quality measures should be developed based on these;
- how to interpret results coming out of them.

#### *2. Main issues*

- Some basic elements to be taken into account when identifying the “best” approach to deal with the problem of variance estimation under imputation are:
  - the approach implemented in production under full response;
  - the need for separate components for sampling and imputation variance;
  - whether users are internal or external to the agency or group producing imputation;
  - the simplicity;
  - the complexity of the sampling design and estimator.
- The availability of a method for assessing the impact of imputation on variance makes it possible not only a more accurate estimation of total variance, and the possibility to make correct inferences, but also to obtain other “products” like better quality reporting to users, useful information to plan/adjust available resources between the sample size and the imputation/follow-up effort. An important aspect relates to the relevance of variance estimation under imputation for evaluating and improving the

effectiveness of imputation and/or for choosing the best imputation model for a given statistical context: if there are alternative models, estimating the non-response variance can be used as a criterion to make a decision on the best treatment method to be adopted.

- The framework developed at Statistics Canada represents a relevant example of how to develop a standardized environment for assessing the impact of non-response and imputation on variance in a unified theoretical, methodological and operational environment. The implementation of tools involving different solutions in different application scenarios requires a consistent effort for the definition of the elements to be taken into account (survey designs, estimates and estimators, variance estimation approaches, non-response mechanisms, imputation models, quality measures). The effort is focused on modelling the random mechanisms (sampling, non-response, editing, imputation, and so on) to be taken under control in the variance estimation process.
- The proposed framework and approaches can be extended in order to assess variance taking into account additional sources of variability such as manual imputation and editing.

### *3. Points of discussion*

- How to develop a comprehensive framework for evaluating non-response, editing and imputation effects on survey estimates, taking into account the following aspects:
  - the sampling contexts (design, estimates and estimators, available auxiliary information,...);
  - the data model;
  - the survey objectives;
  - the non-response mechanisms and incidence.
- In the context of variance estimation under imputation the author indicates the following areas of research:
  - combining survey and administrative data through direct replacement of survey data whether they are adjusted or not;
  - rolling surveys and censuses, where data are missing by design for some areas or groups at any given point in time;
  - multivariate aspects of imputation and relationships between variables;
  - multi-level aspects of imputation when imputation classes are sequentially used in a hierarchical setting.

## WP19 - Performance of re-sampling variance estimation techniques with imputed survey data (Spain) (Supporting paper)

### *1. Summary*

In this paper a Monte Carlo simulation study for evaluating the performance of jack-knife and bootstrap for estimating variance and confidence intervals coverage under imputation is illustrated. The starting data are from one structural business survey (the Structural Industrial Business Survey) and one short-term business survey (the Retail Trade Index Survey). Either random or stratified samples are considered. As for response mechanism, it is assumed either uniform or similar to that observed in the data. Imputation methods used are ratio and mean imputation. Evaluation indicators are (relative) bias and (relative) mean square error of the variance estimator, and the coverage rate for the (95 percent) confidence intervals.

### *2. Main issues*

The paper contains an example of how to carry out a simulation study for evaluating the joint effects of non-response and imputation on estimates precision and statistical inferences. The adopted approach consists in iteratively simulating different random mechanisms influencing results (sampling, non-response, imputation) in order to assess their combined impact on estimates. Evaluation is carried out taking into account the different scenarios in terms of non-response, data and imputation models.

### *3. Points of discussion*

- how to choose the most appropriate variance estimation approach under imputation taking into account the survey design;

- how to choose the most appropriate variance estimation approach under imputation taking into account the survey objectives and the data characteristics (e.g the sample size, the data model);
- how to choose the most appropriate variance estimation approach under imputation depending on the adopted imputation model;
- how to choose the most appropriate variance estimation approach under imputation depending on the non-response characteristics (non-response mechanisms, amount of missing data).

## **Sub-topic b – Links to other parts of processing: editing and data dissemination**

### ***Overall Summary***

Under this topic two papers are grouped:

- WP11 “*Preserving edits when perturbing microdata for statistical disclosure control*” by De Waal and Shlomo (Netherlands and United Kingdom);
- WP13 “*Release of macro survey data. How to do this ideally?*” by Laaksonen (Finland).

The two papers deal with different substantial issues in the area of disseminating statistical micro data. While in the past mainly macro data (e.g. tabular data) were provided to users, the request for micro data has progressively increased in all Countries. In order to meet the external demand, Statistical Agencies had to face a number of specific problems, in terms of legal requirements (e.g. confidentiality), statistical requirements (data accuracy/quality, data utility, data accessibility), costs (e.g. methodological development, implementation of infrastructures and of supporting tools for final users). In particular, Authors of invited and supporting contributions deal with two substantial issues in the area of links between editing and data dissemination: how to protect micro data against the risk of re-identification while preserving not only data utility, but also data consistency in terms of coherence with respect to predefined constraints (edits); how to implement suitable infrastructures and tools for providing (different types of) final users with micro and meta information maximizing the amount of accessible data and the data utility while preserving data coherence and security. These issues are discussed from a theoretical, methodological and operational point of view.

The Netherlands/United Kingdom paper faces the very challenging problem of preserving data confidentiality while maintaining data suitability and reliability with respect to micro and macro edits. The links between editing and data disclosure are deeply discussed, and a novel strategy to cope with the problem is illustrated. The efficacy of that strategy is proved through an experimental application on real survey data.

The paper from Finland deals with the problem of the accessibility and relevance of statistical (micro and meta) data provided by Statistical Agencies, including Eurostat. Available micro and meta data are characterized by different levels of accessibility, due not only to confidentiality, but also to organizational and economic reasons. This fact determines low levels of data utility for the different types of potential users.

### ***Main Issues***

- When providing users with micro data that are secure (in terms of risk of re-identification) and reliable (in terms of loss of information), the additional problem of preserving the data consistency with respect to (micro and macro) edits constraints has to be taken into account. Providing data with records that fail edits damages the utility of the data and increases the risk of re-identification. Furthermore, the perturbation of variable values has an impact on joint distributions, associations among variables, ability to make statistical inferences based on the protected microdata. A trade-off has to be found in order to meet all the different users and producers requirements (data accessibility, confidentiality, utility, reliability).
- The requirement of data consistency while preserving data confidentiality poses new and challenging methodological and technical problems for the development/adaptation of the statistical disclosure control methods/tools integrated with data editing and imputation methods. To this aim, new methods are to be found or already existing approaches are to be revised and properly modified.
- Editing at the statistical disclosure control phase is different from classical editing in terms of both the type of edits to be used and the editing and imputation objectives. Concerning edits, specific types of

edits are generally subject to failure as a consequence of data protection; furthermore, additional edits are used to check the logical consistency for all derived variables or a decision can be made to automatically recalculate all derived variables after the perturbation process. The edit and imputation for perturbed microdata is more complex than edit and imputation carried out at the data processing stage since variables that are perturbed have to remain fixed in addition to the control variables, while other variables need to be changed in order to obtain a logical and consistent record. For this reason, the selection of methods is not necessarily guided by “classical” criteria such as the minimum change requirement: alternative criteria might be preserving as much as possible “macro” properties of the data, such as marginal and joint distributions, data associations, data variability and so on.

- Evaluation strategies are to be designed and performance measures are to be defined in order to be able to evaluate and choose among different statistical disclosure strategies that take into account editing constraints.
- When disseminating public use files, additional (meta) information has to be provided in order to increase data utility. Additional information includes descriptions of target populations, frame population information on respondents and (at least implicitly) on non-respondents, auxiliary information (used in the sampling design, estimation and/or data cleaning), survey weights, documentation about the methodologies used in the different data processing stages (including editing and imputation), meta data on the survey variables of highest importance.
- Public use files are to be disseminated taking into account the different types of potential users in terms of their statistical ability, the objectives of their analyses on data and the available technologies. Practical examples/data analyses might be useful, data available in different formats can be exploited by a larger number of users, well designed data infrastructures for data dissemination can improve data accessibility.

### ***Points for Discussion***

- How to identify thresholds in order to find the optimal balance between acceptable levels of disclosure risk, data utility and data reliability.
- How to combine different methods of disclosure control in order to lower the disclosure risk while preserving the integrity of the data.
- How to identify the optimal techniques for statistical data editing and imputation in order to improve the overall quality and utility of the perturbed microdata by changing fewer variables to correct failed micro edit constraints.
- Which is the minimal set of information NSI have to provide when disseminating public use files
- How to construct public use files that are secure, accessible and useful for different types of users, in terms of contents, and infrastructures.
- How to construct public use files that can be linked/matched cross-sectionally, hierarchically, longitudinally.

### ***Summary of Papers***

#### WP11 - Preserving edits when perturbing microdata for statistical disclosure control (Netherlands/United Kingdom)

*(Invited paper)*

##### *1. Summary*

In the paper a method for protecting categorical micro data from the risk of re-identification while preserving data coherence with respect to micro and macro constraints is illustrated.

The proposed algorithm is based on the general perturbative method called PRAM (*Post-Randomization Method*) This method adds “noise” to categorical variables by changing values of categories for a small number of records according to a prescribed probability matrix and a stochastic process based on the outcome of a random multinomial draw. The proposed method allows for masking records using appropriate controls in the perturbation process in order to minimize the amount of edit failures in the protected file of

microdata. The goal is to assess and minimize disclosure risk while maintaining high utility data. The optimum trade-off is determined through the use of quantitative measures of disclosure risk and data utility. Given the same levels of disclosure risk, when perturbing microdata with PRAM, higher utility data can be obtained by putting more controls into the perturbation process. This causes less micro edit failures and therefore less imputation is needed to correct inconsistencies in the data. In addition, macro editing constraints which alert the data protector to loss of information beyond acceptable thresholds need to be taken into account. Micro edits involves both original edits (i.e. edits used at the data editing and imputation stage), and additional edits required given the controls used in the perturbation process. Macro constraints take the form of information loss measures aiming at controlling the effects of perturbation on aggregates (in particular, joint distributions, associations between variables, statistics used for statistical inferences). The results of an experimental application on statistical survey data drawn from the 1995 Israel Census sample data are presented. In this application, different statistical disclosure methods developed for implementing PRAM and using in different ways control variables in the perturbation process are described. Records that after data perturbation fail edit constraints need to undergo imputation procedures to correct the inconsistencies. The editing and imputation strategy and techniques used in the perturbation process are illustrated. A comparative evaluation of the different strategies proposed in the paper for data protection preserving editing constraints is performed. Evaluation measures are both amount of failure rates and values of macro edit constraints, and typical disclosure risk measures. The latter measures are needed in order to assess the disclosure risk of the perturbed microdata in order to ensure that the risk of re-identification is not greatly increased as a result of placing controls in the perturbation process.

## 2. Main Issues

- The statistical data disclosure process for protecting clean and fully edited microdata generally result in records failing edit constraints (inconsistencies will reoccur between the perturbed and original variables): for Statistical Agencies, releasing microdata containing records that fail edits it is not suitable for different reasons:
  - this damages the utility of the data;
  - an inconsistent and illogical record will immediately pinpoint to potential attackers that the particular record was perturbed and attempts can be made to unmask the record.
- An additional theoretical, methodological and operational effort has to be done for developing strategies allowing for gathering different requirements: minimizing the risk of data re-identification, maximising data utility and data reliability. A trade-off has to be found among these aspects in order to meet all the different users and producers requirements.
- When implementing statistical disclosure methods using controls to minimize the risk of data failures, critical elements are:
  - identifying the appropriate set of micro and macro edits;
  - identifying the appropriate control variables;
  - implementing controls in the statistical disclosure control method and find the “optimal” balance between the different quality requirements.
- Once data protection/perturbation has been carried out, editing and imputation techniques are used for treating unacceptable data (i.e. for identifying not acceptable data relations and restoring data consistency at the data perturbation stage).
  - editing and imputation for perturbed microdata is more complex than edit and imputation carried out at the data processing stage since variables that are perturbed have to remain fixed (in addition to the variables used as control in the perturbation algorithm), while other variables need to be changed in order to obtain consistent records: an algorithm is proposed to meet these requirements;
  - the optimal editing approach and the most appropriate imputation technique are to be identified taking into account the characteristics of statistical data to be processed. This can be done by comparing results obtained by adopting different strategies.
- Determining the optimal trade off between risk of re-identification and data utility on the protected and consistent micro data requires the definition of an overall evaluation framework involving different types of measures. Edits failures and macro indices are to be added to classical disclosure risk measures.

- The evaluation framework allows for comparative evaluations of different statistical disclosure approaches in order to select the one that guarantees the better balance between the above mentioned requirements.

### 3. *Points of discussion*

- How to identify thresholds in order to find the optimal balance between acceptable levels of disclosure risk, data utility and data reliability: elements to be taken into account in this phase are the specific practices, policies and protocols governing the release of microdata.
- How to combine different methods of disclosure control in order to lower the disclosure risk while preserving the integrity of the data.
- How to identify the optimal techniques for statistical data editing and imputation in order to improve the overall quality and utility of the perturbed microdata by changing fewer variables to correct failed micro edit constraints.
- Concerning statistical data disclosure, the main problems that remain to be solved are how to assess disclosure risk based on realistic scenarios, how to measure the quality and utility of microdata that has undergone masking techniques, and what is the optimum balance between minimizing the disclosure risk and maximizing the utility of the microdata.

## WP13 - Release of micro survey data. How to do this ideally? (Finland)

*(Supporting paper)*

### 1. *Summary*

In the paper the Author discusses the problem of how to develop frameworks for releasing public use files, what types of infrastructures and supporting tools are to be provided to users in order to allow them to exploit statistical information as much as possible. The basic requirements Statistical Agencies have to satisfy when releasing micro data are described: they include both statistical and operational aspects, i.e. data confidentiality and data consistency from one hand, data utility and accessibility from the other hand. As an example of construction of public use files and related infrastructures, in the paper the experience of the European Social Survey is described. This survey has been realized in 22 European countries and lot of information is available on the website in the form of micro data and meta information. A large amount of resources effort has been spent in building a framework for facilitating the use of data: this framework involves legal, statistical and technical elements. Secure and coherent micro data can be selected by final users together with all needed meta information and documentation about definitions, classifications, coding, and so on. Data can be downloaded in different formats and are free. The data file includes additional information for users interested in specific analyses such as assessing the interviewer effects and analysing non-response. The Author encourages the discussion on the problem of releasing micro data for public use, and suggest some starting elements to be considered.

### 2. *Main Issues*

- The basic requirements to be fulfilled when releasing micro data are described:
  - micro data have to be consistent and complete, meta data documentation on population, sampling and sampling units, respondents, variables definitions and coding, weights, data processing, data formats are to be provided to final users in order to increase data utility;
  - data are to be secure, and information about loss of data utility due to confidentiality are to be provided to users;
  - NSI's have to develop infrastructures and supporting tools that facilitate users in data access and use, avoiding mistakes in handling data and allowing efficient data use also to not expert operators and analysts. The SSE experience represents a good example of how to organize public use files from both a methodological and technical point of view.
- Users might be willing to get different sets of data/variables, and the data are to be organised in such a way that this requirement be fulfilled as much as possible, taking into account the main potential users and uses of the data.

- Variables definitions, categorization and coding are to be clearly documented in order to avoid errors in analyses.
- Fieldwork information, e.g. relating to respondents/non respondents, weights, interviewers, and so on can be useful for expert users that might be interested in more sophisticated analyses.
- Users are to be informed about how to use data, also through examples and documentation, as well as about data formats.

### 3. *Points of discussion*

- Which environment and tools are to be developed to construct secure and accessible public use files easily accessible and exploitable.
- How to construct public use files that fulfil requirements of different types of users, i.e. that allow for different types of analyses by users having different statistical/methodological background.
- Which information has to be put in a public use data file.
- How to design and implement a database for public use in order to allow users to get different portions of information he is interested in.
- Why NSI's do not proceed more quickly in organising themselves in order to meet the increasing request of statistical micro information: the main reasons can be either legal (confidentiality), methodological, technical, economical (required resources and time).

## **Sub-topic c – Implementing editing strategies: (re-)designing editing processes**

### ***Overall Summary***

Under this sub-topic 5 contributions are grouped:

- WP 12 “*A selective editing method considering both suspicion and potential impact, developed and applied to the Swedish foreign trade statistics*” by Anders Jader and Anders Norberg (Sweden);
- WP 20 “*An editing procedure for the low pay domain in the annual survey of hours and earnings*” by Salah Merad and Heather Wagstaff (United Kingdom);
- WP 14 “*Introducing and implementing a new data editing strategy*” by Elmar Wein (Germany);
- WP 18 “*The embedding of a uniform statistical process*” by Jeroen van Velzen (Netherlands);
- WP “*Linking Data Editing Processes by IT-Tools*” by Carsten Kuchler and Corina Teichmann (Germany).

The paper CRP.2 “*Selective editing using plausibility indicators and Slice*” by Jeffrey Hoogland (Netherlands) belongs to this group. It is an update of the paper presented at the 2003 Work Session on Statistical Data Editing (Madrid).

Papers under this sub-topic deal with the problem of (re-)designing, rationalizing and optimising editing performed at the post-data collection stage. In large-scale surveys conducted by Statistical Agencies the reduction of editing costs, in particular those due to manual/interactive editing, is a well-known problem. In short term business surveys the problem is even more serious, due to the additional requirement of timeliness. In general, resources and time represent fixed constraints, and an effort has to be done to optimise between them and the survey quality requirements.

The design of efficient editing processes balancing between accuracy, time and costs implies the re-design of the overall planning of the data editing. To this aim, specific projects are carried on at Statistical Agencies aiming at implementing “optimal” standardized editing processes for specific statistical areas/homogeneous classes of survey processes. Standardization involves both planning editing and developing/introducing (new) editing methods to be used for the data handling. Relevant changes in the survey planning and organization can be determined, due to the need of incorporating and harmonizing the new methods/data processing flows, depending on the available resources and organizational constraints.

Concerning methods and approaches for improving the efficiency of editing at the post-data collection phase, all papers propose the use of selective editing methods for balancing between manual and automatic data treatment. The general idea is to identify units with the highest impact on target estimates and limit to them

manual editing activities. The identification of “optimal” trade-offs in terms of data accuracy, timeliness and resources spent for manual editing (costs) is a critical point. In effect, implementing a selective editing strategy poses a number of key problems: among others, identifying the hierarchy among variables (target/non target), determining the hierarchy among units and errors (most/less influential), defining appropriate (score) functions and the corresponding thresholds taking into account the multivariate nature of statistical data and the specific survey quality requirements.

In the Swedish paper an experience in developing and implementing an overall selective editing strategy for short-term statistics is described. In this strategy, problems relate mainly to the identification of appropriate strategy to balance between suspicion and potential impact criteria taking into account the multi-purpose nature of the survey, and the complex quality requirements of the different surveyed phenomena.

The paper from UK deals with the problem of how to balance between costs and quality focusing on the optimal allocation of the available resources for manual editing under time and quality constraints. The proposed strategy aims at providing survey managers with a standard framework allowing them for periodic verifications of whether available resources are sufficient to cope with current and expected future workload. Furthermore, a method for identifying optimal thresholds and determining their periodic adjustment in order to satisfy editing capacity constraints whilst maintaining data quality is proposed.

In the German paper (WP14) the re-design of the editing process at the Federal Statistical Office is illustrated, and the main features of a new editing strategy implemented in the area of business surveys is described. The re-design gave rise to a new organization of the editing activities, supported by IT tools and carried on through the strict cooperation among appropriate specialised personnel. In addition, new selective and macro editing methods have been integrated in the data treatment in order to increase the efficiency of the editing process under quality, time and costs restrictions.

The paper from Netherlands illustrates a survey re-organization processes aiming at re-designing and standardizing the questionnaire, the logistic and the editing processes in the area of structural and short-term business statistics. Parallel to the redesign of the process the organisation of the Business Statistics Division changed, since the implementation of a general uniform process caused new internal dependencies. In the paper the new survey organization and its new interdependencies, well as the new editing strategy adopted are described.

### ***Main Issues***

- Implementing new data editing strategies/methods in survey processes has an impact on the overall organisation of statistical production processes. Organisational changes are determined by available know how, qualified personnel, financial means, existing strengths and weaknesses, and equipment.
- The improvement of data processing efficiency is not only responsibility of the data editing methodologists but requires a broad approach, involving many organizational and methodological aspects (e.g. re-design of (paper and/or electronic) questionnaires, exploit as much as possible already existing information, anticipate editing at the data collection phase, and so on).
- Since new methodologies are generally involved in new standardized processes, there is a need for convincing statisticians about the possible advantages deriving from their use, and as well as for training them on the use of the new methodologies.
- The design of strategies for optimal resources allocation under time and quality constraints is survey-dependent: the definition of the elements involved in selective editing approaches depend on elements like the (multiple) objective of the survey, its periodicity, available auxiliary information (e.g. historical data) and its reliability, the level of data publication. Therefore, generalized formulas cannot be defined: only general principles can be provided to methodologists.
- When designing selective editing strategies the multivariate nature of the data has to be taken into account. One way to cope with the problem is reducing the complexity of the error localization problem by identifying statistically homogeneous phenomena. Another possibility is defining hierarchies among variables.
- When designing selective editing strategies the multivariate nature of errors has to be taken into account: (relevant) errors affecting variables are not independent; different types of errors may jointly affect data.

- Determining the best selective editing strategy in terms of balance between costs and quality requirements implies the definition of an evaluation framework for assessing the performance of the different possible methodological, technical and operational alternatives. A wide range of measures of impact has been proposed in literature. The choice of the indicators depends on the specific survey quality requirements and constraints.

### *Points for discussion*

- How to preliminarily estimate costs for re-design editing strategies (for planning, developing, implementing training on the new editing concept/methods) and for extending/adopting the strategy to other scenarios.
- Integration of outputs cause new and more intensive interaction between units that publish different but related statistics: how these interactions can be efficiently managed?
- How to efficiently manage the possible updates of the standardized editing process due to changes in surveys contents, survey constraints and/or survey quality requirements?
- How to evaluate the usefulness and the reliability of historical information used in selective editing.
- How to put under control multivariate relations among variables when implementing selective editing.
- How to reduce the complexity of multivariate selective editing taking into account the quality requirements for the different target phenomena. In particular, how to implement selective editing when results are published at different level of detail/aggregation.
- Which are the key elements to be taken into account when adjusting a selective editing strategy for possible re-design or changes in survey organization/objectives.
- Which are the key elements to be taken into account to prioritise the quality dimensions (e.g. bias, variability, marginal and joint distributions, and so on) for evaluating the performance of a selective editing strategy.

### *Summary of papers*

#### WP12 - A selective editing method considering both suspicion and potential impact, developed and applied to the Swedish foreign trade statistics (Sweden)

*(Invited paper)*

##### *1. Summary*

In the paper a score function computed as a weighted geometric mean of measures of suspicion and potential impact successfully implemented in the editing process of the Swedish Foreign Trade Statistics is described. The survey, created for member states in the European Union (EU), covers trade of goods among states within the EU. The selective editing strategy designed and implemented for this multi-purpose survey is rather complex due to the high number of factors/parameters influencing the final results that are to be put under control. The proposed score function consists of three main elements: a measure of the “overall” suspicion, measured using the Hidiroglou and Berthelot approach properly adjusted to meet the specific data characteristics; a measure of the error suspicion, aiming at prioritising potential errors on target variables; a measure of the potential impact at domain level based on the use of historical information. The definition of each element poses theoretical and practical problems, due to the potential effects on the performance of the overall strategy. Methodological and application problems encountered in implementing the overall selective editing strategy and for the definition of its basic elements (parameters) are discussed: since a high number of possible combinations of parameters values are possible, experiments and applications have been carried on at Statistics Sweden for identifying the best solution. The good results obtained in terms of both costs reduction (number of units identified for manual editing) and expected quality (hit rates, amount of identified error) encouraged the implementation of the method in production (January 2004).

##### *2. Main issues*

- Concerning the definition of parameters required for estimating suspicion and potential impact:
  - Should historical data or current data be used for computing the quartiles, including the median, that are used in the measures of suspicion and potential impact.

- On which domains do quartiles are to be computed in order to get both relevant and accurate measures of mean and dispersion for each observation? Relevance depends on the homogeneity of the group, whereas accuracy depends both on homogeneity and number of observations.
  - What is the minimum number of observations needed in each domain, which variables are to be used to obtain homogeneous aggregations;
  - Should weighted or unweighted quartiles be calculated.
- Several indicators are proposed for evaluating the performance of the editing method for different parameter values. The best choice of the parameter values depends not only on the target variable, but also on the specific indicator, as different combinations give rise to different conclusions depending on the analysed indicator.
  - Once the overall strategy has been set up, it has to be continuously monitored and parameters possibly adjusted to put under control possible modifications in the survey organization and/or data behavior.

### 3. Points for discussion

- A question left open for discussion is whether we will be misled when we try to construct a new editing method by analysing data that has been flagged by an old method. Are conclusions concerning the optimum parameter values dependent on the method that has been used to flag the observations?
- How to evaluate the reliability of available information used in the selective editing approach taking into account the multivariate quality requirements of the survey
- How to identify the best selective editing strategy taking into account the different quality requirements of the investigated phenomena, particularly in terms of aggregations/level of detail of figures to be published.
- To what extent the proposed approaches can be generalized to other scenarios.

## WP20 - An editing procedure for the low pay domain in the annual survey of hours and earnings (United Kingdom)

*(Supporting paper)*

### 1. Summary

In the paper the current activities at ONS for developing a methodological and operational framework for supporting the management of more efficient editing processes is illustrated. In particular, ONS undertook the development and the application of new methodologies to survey production processes in order to better allocate resources along the editing phase, and in particular for rationalizing manual editing taking into account available resources, time and quality constraints. A new selective editing approach to balance between editing capacity constraints and data quality under time restriction is illustrated. In this approach the identification of potential errors having a substantial impact on target figures is split into two phases: set up a decision rule informing data analysts at regular time intervals whether available resources are sufficient to cope with current and expected future workload; determining optimal thresholds and the amount by which they can be raised in order to satisfy editing capacity constraints, whilst maintaining data quality. In the first case, an approach that takes into account some parameters representing the competing constraints (resources, time, expected quality) is proposed. The threshold's determination and adjustment, that is based on the use of past pre-edited and edited data, takes into account the expected quality of estimates (in terms of editing bias ratio), and the resources constraints. In particular, the threshold values are set to ensure that the overall editing bias, which results from not editing all records potentially in error, of the estimator of a population parameter of interest is small in relation to the standard error of the estimate. The efficacy of the overall strategy and its main problematic aspects are discussed by analyzing the results of an experimental application on the British New Earning Survey data. The expected benefits in terms of number of records needing manual editing and editing bias are analysed.

### 2. Main issues

- In order to monitoring the performance of the editing process in terms of quality of data and editing capacity, the balance between available resources for manual editing, time and quality constraints has to be periodically verified and the available resources re-allocated, e.g. by raising selective editing

thresholds. Appropriate criteria are to be defined in order to properly determining the timing of the change to the thresholds and the amount by which they are to be raised.

- The determination of thresholds based on past data requires additional assumptions on respondents behaviour (response rates and their trend), on errors (types of errors, their incidence and trend), and on the behaviour of editors.
- When implementing selective editing the multivariate nature of the data has to be taken into account. Reducing the complexity of the influential error localization problem poses additional problems, such as the identification of restricted sets of relevant phenomena, the definition of groups of related variables that are homogeneous in terms of relevance, the identification of multivariate criteria to put under control the multiple quality requirements.
- In the proposed strategy, time is not only a constraint, but also a “covariate” that directly influences the determination of the optimal threshold, as the amount of available information needed for estimating the required parameters (amount of registered questionnaires) varies along the interval of allowed time for the data entry and editing phases.
- Selective editing strategies need to be constantly updated in order to take into account the possible changes in survey organization, such as the questionnaire contents and structure, the survey design, the available resources, and so on.

### 3. *Points for discussion*

- Integrating in the production process the periodical verification of the elements conditioning the thresholds raising.
- Integrating in the production process the periodical assessment of the assumptions (such as those relating to respondents, errors and editors) underlying the thresholds raising strategy.
- How to reduce the complexity of selective editing due to the multipurpose nature of surveys while avoiding biasing effects at variable or domain levels.
- To what extent the process can be standardized to face uncertainty in future survey cycles, especially because of the (possible) redesign of the survey.

## WP14 - Introducing and implementing a new data editing strategy (Germany)

*(Supporting paper)*

### 1. *Summary*

In the paper the process of re-designing and implementing a new editing strategy at the Federal Statistical Office (FSO) is illustrated. This strategy is based on a new data editing concept that:

- induces the introduction of new IT tools for the planning and management of data editing,
- implements new data editing methods based on practical test of a combined selective and macro-editing method in combination with an automatic error determination and correction,
- introduces a new training concept for data editing.

In the paper the changes induced by the new data editing concept on the planning of data editing, on the data processing and editing methodologies are illustrated. The data editing concept was implemented by introducing IT tools which support relevant activities as regards data editing, that means subject matter statisticians apply automatically the data editing concept by the use of IT tools. Several specialized figures interacts in the new organization (the IT production manager, the PL-Editor, the subject matter statisticians) with different responsibilities: subject matter statisticians with a specific degree of training specify the checks and use selective and macro editing methods; the IT production manager consults subject matter statisticians as regards the data processing of a survey; in addition to this he uses the specifications of the PL-Editor.

A very important part of the implementation strategy consisted in the introduction of selective and macro editing methods to increase the data processing efficiency in terms of balance between accuracy, timeliness, available resources and respondents burden. Combining selective editing, macro editing and automatic editing is recognised as the best strategy.

Not negligible aspects of the project were the need of training users on IT tools and convince managers and subject matter statisticians so that they accept modern data editing methods. In the paper actions performed to face these problems are illustrated. Among others, meetings illustrating the project aims and the possible advantages of adopting the new data editing concept, seminars and publications illustrating the new editing methods and approaches. Selective, macro, and automatic editing methods were presented and explained at a very simple level so that the participants could comprehend the statements and possible improvements deriving from its adoptions are showed by simple examples.

In the paper the characteristics and the results of the application of the new concept and new methods to the Annual Survey on Costs of the Producing Industry to test its performance are illustrated. In particular, a selective editing method combined with macro editing has been developed and implemented in order to cope with the specific survey requirements. Possible improvements of the editing planning and editing methods are identified and discussed through the analysis of the application results.

## 2. *Main issues*

- Implementing new data editing strategies/methods in survey processes has an impact on the overall organisation of statistical production processes. Organisational changes are determined by available know how, qualified personnel, financial means, existing strengths and weaknesses, and equipment.
- The re-design of the data editing planning requires the preliminarily identification of target needs and focal interests at the Statistical Agency. Activities which permit to satisfy best the needs of subject matter units and implement them in helpful IT tools are to be chosen. The addressees who are responsible for the tasks are to be identified and convinced of the benefits provided by the new IT tools and methods and the IT production managers have to integrate these methods in their work flows, pre-information on their needs are to be obtained and parts of the objects/IT tools to be implemented are to be tested.
- Changes in the planning of data editing processes required a change of the management culture, i.e. the replacement of individual best performance by a systematic standardized processing. Manager of the upper hierarchy should be informed and convinced first. Their support is an important precondition of a successful implementation but it facilitates only the start of the implementation. It has to be confirmed by a positive feedback from the specialists in the middle term to ensure that the new methods will be a part of the standard processes.
- Advantageous preconditions facilitate changes: for example, reductions of staff create an incentive for the introduction of modern data editing methods. Another positive precondition can be created if target group experts obtain information on the expected results (e.g. by the simulations of new data editing methods).
- The improvement of data processing efficiency is not only responsibility of the data editing methodologists but requires a broad approach, involving many organizational and methodological aspects: re-design of (paper and/or electronic) questionnaires, reducing respondent burden exploiting the already existing information, anticipating editing by integrating checks at the data collection phase, and so on. Data editing methodologists have to collaborate and to be able to communicate with IT managers, questionnaire methodologists, subject matter statisticians, organisers and lecturers.
- Concerning selective editing, an optimal trade off has to be identified among the components of the proposed score function in order to optimize the error detection and the selective editing method priority setting. Experiments and analyses are to be performed taking into account the multivariate nature of data, the survey characteristics and quality requirements.
- Concerning macro editing, important factors that have been considered for the priority setting among data domains (strata) were:
  - the need of producing projected preliminary results
  - the need to rapidly improve the data plausibility
  - the structure of the costs (heavily determined by the size of the units - enterprises),
  - the need to maintain the continuity of the results in terms of comparability with previous year's data.

- Selective/macro editing approaches needs to be carefully tested in order to identify possible inefficiencies and make improvements, and continuously monitored for updating/adapting them to changes in the survey organization/contents.

### 3. Points of discussion

- How to preliminarily estimate costs (for planning, developing, implementing training on the new editing concept/methods) for extending/adopting the strategy to other scenarios.
- Is the selective editing strategy robust against strong changes in the survey organization and/or modifications of the survey contents?
- How did missing (i.e. latecomer) records affect the behaviour and power of the selective editing method? The influence of important missing erroneous records on preliminary results has to be checked.
- Did the selective method work properly under completely changing conditions (new sample)?
- Identifying approaches for efficient test of the selective/macro editing strategy to set parameters so that the best trade off between quality, cost and time requirements is obtained for each specific application.

## WP18 - The embedding of a uniform statistical process (Netherlands) (Supporting paper)

### 1. Summary

In the paper an example of how the efficiency of survey processes (including editing) can be improved by re-designing and re-organizing the overall statistical production processes is presented. The Author presents the re-organization at Statistics Netherlands of survey processes of the division of Business Statistics at Statistics Netherlands. The new organization has been carried on during the project IMPECT (*IM*plementation of *E*conomic *T*ransformation *p*rocess). The project aimed at re-designing and standardizing the questionnaire, the logistic and the editing processes in the area of structural and short-term business statistics. A new standard process defined according to the structure of processes characterized by a homogeneous structure, not dependent on the specific statistics (branch) has been developed.

In the new standard process all the survey phases are revised and improved from an operational and methodological point of view: the sample is optimized using a Neymann allocation; the layout and structure of all questionnaires are harmonized; a central logistic system for data collection is created, consisting of a central database in which all contacts with the respondent are registered and the preferred method of data collection is recorded (e.g., paper questionnaire, electronic questionnaire; a new editing strategy aiming at improving the efficiency of data processing activities, and at exploiting as much as possible the available tools and the most recent methodological advancements is developed, consisting of the combined use of selective editing, (based on the use of a global plausibility indicator), interactive editing (based on the use of the software Blaise), automatic editing (based on the use of the software SLICE implementing based on the Fellegi-Holt paradigm and regression based imputation), robust outlier detection; the weighting strategy is standardized (the software BASCULA is used); the publication system is standardized by creating a central database filled with all records on both annual structural business statistics and short term statistics (this database is used to generate tables for Statline, the output database of Statistic Netherlands).

Parallel to the change of processes, the organisational structure also changed: the Author describes the new structure of the Statistics Netherlands division of Business Statistics aiming at supporting the new processes, and illustrate the corresponding responsibilities and interactions. In particular, the *Business surveys departments* responsible for the data collection and editing of the survey data has been created. The main critical aspects in the re-design of processes and organization are highlighted, and open problems are pointed out. Further generalization of questionnaires and logistic processes is supported by a new project called Prodonna.

### 2. Main issues

- Processes are too complex to be overseen by one person or even a small group of statisticians. Especially processes in which subject-matter knowledge is essential, require an organizational structure that combines overview and coordination with commitment and clear responsibilities for the statisticians.

Especially processes, in which subject-matter knowledge is essential, require an organisational structure that combines overview and coordination with commitment and clear responsibilities for the statisticians.

- The implementation of a general standard process caused new internal dependencies: managing interactions of the different actors in the new organization (the departments) not only in terms of output and resources, but also in terms of coordinating changes in requirements and maintenance of joined applications has to be organized.
- A major redesign of processes needs a strong, central control and coordination during implementation. Subsequently control and coordination have to be delegated. Requests of modifications in the standard process in specific applications are to be properly managed.
- Since new methodologies are involved in standard processes, there is a need for training statisticians in charge of different responsibilities on these methods: not all statisticians are willing and able to become sufficiently familiar with the new techniques. This caused both disproportional workload and risks of continuity.
- Centralized coordination of methods, questionnaires and output may cause the statisticians to lose the overall view over all relevant processes that lead to a publication. This may cause the incentive for maximal contribution to diminish. Tasks are to be assigned in order to make workload more equally divided among units.

### 3. *Points of discussion*

- Integration of outputs will cause new and more intensive interaction between units that publish different but related statistics: how these interactions can be efficiently managed?
- In the coming years statistical processes at Statistical Agencies will be characterized by increasing use of complex methods in large integrated applications. The developments in ICT enable to further intensify re-use of data which forces us to focus more and more on integration. An critical organizational issue is how to efficiently manage integration of outputs and intensified use of register data.
- How and to what extent standard processes and organizations can be extended to other statistics: in general the data collection processes are more easily expanded to other statistics, the processes of analyses show more statistic specific elements.

## **Sub-topic d – Implementing editing strategies: using external data sources**

### *Overall Summary*

Under this sub-topic two contributions are grouped:

- WP 16 *“Planning editing and imputation as an integral part of multi-source data collection”* by Olivia Blum (Israel);
- WP 17 *“Modelling the construction of a social accounting matrix in the context of statistical matching”* by Marcello D’Orazio, Marco Di Zio and Mauro Scanu (Italy).

Papers under this sub-topic deal with the general issue of designing and implementing editing and/or imputation strategies when data come from several sources of data. Due to the large amount of available (statistical and administrative) information and to the need of producing timely information at lower costs, the interest of Statistical Agencies in exploiting as much as possible already existing sources of data has significantly increased. This tendency determined the need of identifying theoretical, methodological and operational solutions to be integrated in an overall, more complex editing strategy. In this strategy, appropriate editing methods dealing with information collected for different purposes through different (statistical or administrative) production processes and characterized by different levels of quality are to be efficiently integrated. From a methodological and theoretical point of view, harmonizing the target populations, the variable definitions and classifications are the classical problems to be tackled. In addition, solutions to complex organizational and operational problems are to be identified.

The Israeli paper is an example of designing editing strategies when integrating both register and field-survey data for statistical purposes. In particular, the Author illustrates the editing requirements and the suggested strategy for implementing the integrating Israeli census of population using several (administrative and field-survey) sources of information. Editing is spread over all the survey process, from the early stage of data-source selection, through data-collection and integration, up to the formation of the final census files. Some relevant issues and critical points are highlighted by the Author in this area.

The Italian contributions deals with the specific problem of combining data from different sources of information that do not observe the same set of units, so that neither merging nor record linkage techniques can be applied. In this case, Statistical Matching is proposed as a peculiar form of data integration.

### ***Main Issues***

- The use of multiple sources of data implies the re-design of the overall survey and editing processes. Appropriate methodologies for measuring the accuracy and the reliability of the different sources of data, for the editing and imputation of the different combined information/populations, for statistical modeling, for record linkage, for statistical matching, are to be harmonized in the overall integrated strategy. If data are obtained through different modes of data collection (e.g. computed aided, traditional paper questionnaires and so on), the editing strategy results even more complex: different sub-processes are to be designed to face the different levels of quality of the different sources of data, and then integrated in a overall strategy.
- When combining different sources of data, since they are collected and processed for different purposes, statistical classifications and definitions are usually inconsistent each other. Therefore, a theoretical effort is needed in order to harmonize classifications and definitions of units and variables in order to produce the required statistical information.
- Planning editing and imputation is challenging when data come from different sources. Each source is affected by specific errors, the simple integration of data is a source of errors in itself. Errors in one phase of the production process influence the scope and content of editing needed in the following processes. Errors have to be treated at data collection process, at the data integration phase and after each data manipulation stage.
- Under resources constraints, the overall editing process has to be optimized at both local and global levels. When combining multiple sources of data costs relate to several problems: among others, identifying the most appropriate editing strategy for each single source/product (in terms of resources needed and expected quality), linking information, harmonize data in terms of definitions and classifications. The amount and type of editing that balance between the level of accuracy of the data, and the editing time and costs should be determined.

### ***Points for discussion***

- When combining different sources of data, harmonization of target populations, variables classifications and definitions has to be performed with great caution: in effect, an optimal procedure does not exist, and in general this operation produces changes in the original variables meaning, changes in the definition of the population target, and as an overall result changes in the initial informative power of the samples.
- The use/integration of more sources of information implies a highest complexity of the survey process/editing process. Determining the trade off between the reduction of costs due to using already existing information and the additional costs due to the increased complexity of the processes and the need for developing new strategies/methods to validate and harmonize information is a central problem. In case of using administrative data, the additional costs for obtaining data and making them usable for the specific statistical purposes are to be considered.
- When combining different sources of information, assumptions on data relationships are often unavoidable. How to formulate and test these assumptions and how to evaluate their effects on final results is a crucial problem to be discussed.

## *Summary of papers*

### WP16 – Planning editing and imputation as an integral part of multi-source data collection (Israel) (*Supporting paper*)

#### *1. Summary*

In the paper the product-oriented strategy adopted at the Israeli Statistical Agency for carrying on the Israeli Census of population. The main methodological and implementation problems faced for planning and implementing the Census are illustrated. In particular, the Author discusses problems relating to planning a global editing strategy in which multiple statistical and administrative sources of data are combined together for obtaining the final statistical product(s).

The final census products are the geo-demographic file and the socio-economic file. The first is a weighted integrated administrative file, calculated by the surveys estimates. This file is mainly used to supply the demographic margins of the census. The second file is the one that the vast majority of the users use. It has all census information of about 20% of the population. It is the source of socio-economic and household data, and its demographic information, although not exclusive, is required in conjunction with the other attributes. This file is adjusted to the demographic benchmarks, provided by the geo-demographic file, by weights and imputations, and has to meet the quality requirements of the census.

The final census products are obtained in a “product-oriented” approach, by integrating several intermediate products. Also the intermediate products are obtained by using and/or integrating different of administrative and/or statistical sources of data. In this process, data and intermediate products can be used for different purposes. The Author describes the overall structure of the processes and products underlying the Israeli Census and the links among them.

In the Israeli Census scenario editing is spread over all the statistical production process, from the early stage of data-source selection, through data-collection and integration, up to the formation of the final census files. The “local” editing processes are linked throughout the statistical production process in a “global” strategy, that has to be locally and globally optimized, since the statistical quality of intermediate products and the accuracy of the different sources of information have a critical impact on the quality of the final product, optimization of editing should be local as well as global. In the paper the local and global editing strategies adopted at the Israeli Census are illustrated.

#### *2. Main issues*

- The statistical production becomes more and more complex since it uses many sources of data and it leads to many end products. Consequently, editing has to widen its scope to include these ends and to encompass data manipulations that have not been part of editing in the past. Furthermore, a product-oriented editing-strategy is called for, because of the quantity and the diversified nature of the processes involved.
- Planning efficient editing in a product-oriented approach includes the definition of all products, within the main project and those related to it; their required quality; and the derived questions of implementation. A mixed editing strategy has to be designed, in which specific editing processes, dealing each with the quality requirements of a specific intermediate product, are integrated in order to meet the quality requirements of the final-product.
- Designing editing and imputation strategies is complex and challenging when data from different sources are used in an inter-dependent process. Each source carries its own unique errors, the mere integration of data is a source of errors in itself, and when the integrated file is an intermediate product, more errors are accumulated before the final file is obtained. Errors in one phase of the production process stipulate the scope and content of editing needed in the following processes. Errors have to be treated during the data collection process, during the integration of files and after each data manipulation.
- In the proposed widespread strategy, the active ensuring of the appropriateness of the data is done simultaneously with the data-source selection, i.e. editing and imputation serve not only as a validation and correction phase, but also as an integral part of the data collection process itself. This holds in particular for administrative files: since their accessibility is limited and usually involves costs, a source-

selection process is needed and has to be based on a careful evaluation of the statistical quality of the contents and the physical usability.

- The editing plan for obtaining the final/intermediate products starting from different sources can be very different. In effect, it depends on the source's characteristics and on the specific statistical purpose to be gathered by using the source itself. Since resources are limited, the overall editing plan has to be optimised. In a product-oriented project, where the statistical quality of the intermediate products has a critical impact on the quality of the final product, optimisation of editing cannot be done on a global level and on a local level.
- The different sources/products can be used for different purposes/products in the overall statistical project. Additional uses imply additional evaluations and hence, a broader editing plan that optimises additional quality dimensions and increasing costs, then they are to be carefully evaluated. The boundaries of the global editing plan of a specific project have to be well defined beforehand, and that it may include processes not directly related to the project at hand.
- Extended scope and diversity of editing processes may eventually lead to a complicated or even inapplicable implementation, under given resources. There is a need to identify the threshold beyond which excessive editing, carried out under simple universal rules, becomes less damaging than the selective, product-oriented, specific editing. This threshold should be defined, *inter alia*, in terms of the number of data files, variables and meaningful products involved, the marginal costs added in the transition to undistinguishing editing process, and the applicability of the plan. Beyond the threshold, a pure product-oriented strategy will not prevail and a different approach is expected.

### 3. Points for discussion

- Since the use of each additional source implies a highest complexity of the editing process (the inclusion of either a new source or intermediate product implies new uses and therefore a need to plan editing accordingly) which are the key elements to be taken into account for the preliminary assessment of the advantages/drawbacks due to using an additional source?
- How to identify the threshold beyond which excessive editing, carried out under simple universal rules, becomes less damaging than the product-oriented, source-specific editing?

## WP17 – Modelling the construction of a social accounting matrix in the context of statistical matching (Italy) (Supporting paper)

### 1. Summary

In the paper the Statistical Matching approach used for the construction of the Italian Social Accounting Matrix (SAM) is described. The SAM is a system of statistical information containing economic and social variables in a matrix formatted data framework. In Italy such archive is not directly provided by a specific source, but has to be built by means of the *integration* or *fusion* of two different surveys: the Household Balance Survey conducted by the Bank of Italy and the Household Expenditure Survey conducted by Istat. This integration process, that allows to put together information on household outlays and information on household entries independently observed, is carried on by means of information on the socio-economic characteristics observed on both the samples. This *statistical matching* process consists of three steps:

- i) the consistency of the two surveys is checked and, if necessary, the two surveys are harmonized;
- ii) the statistical framework where the sample surveys live has to be defined; and
- iii) according to the two previous steps, an appropriate statistical matching method has to be applied.

The final result is an integrated data set where all the variables needed for estimating the target distribution(s) are jointly present. This data set is obtained through imputation techniques following one of the two hypotheses so far discussed. The data set is constructed by imputing the missing variables (completely non-observed) in one file with values taken from the other data set (donor file). The imputation is done through the distance hot-deck stratified.

In the paper the methods adopted in the different steps of the construction process (including imputation) and the underlying assumptions are described. Authors underline some of the most critical theoretical and methodological aspects underlying the construction of sets of micro data starting from independent data

sources characterized by relevant differences, such as those considered for the Italian SAM. In the paper the different consequences, in terms of expected data and information quality, of the possible different assumptions made when performing such an operation are investigated through experimental applications.

## 2. *Main issues*

- Since a general optimal procedure for the construction of sets of micro data starting from independent data sources does not exist, the fusion process has to be performed by taking into account the specific characteristics of the sources and the data to be combined. In particular, harmonization of target populations and variable definitions (including definitions of variables categories) are basic problems.
- Harmonization must be performed with great caution, noting that does not exist an “optimal” procedure. In fact, the harmonization phase consists of a kind of “simplification” of some key characteristics of the different surveys. This operation produces changes in the original variables meaning, changes in the definition of the population target, and as an overall result changes in the initial informative power of the samples. Statistical matching output is greatly affected by these operations. A rule of thumb can be the following: change as less as you can during the harmonization step.
- When performing statistical matching operations using independent sources of information, assumptions like the Conditional Independence Assumption (CIA) and assumptions on data relationships are unavoidable. These assumptions have relevant consequences in terms of expected data and information quality. In particular, the CIA has to be formulated in order to reducing the risk of biasing effects on the estimated data relationships due to departures from the CIA, and exploiting as much as possible all the available information coming from the matched sources.
- One way of reducing the risks due to departures from the CIA is through the use of auxiliary information. Auxiliary information can be micro (additional set of micro data) and/or macro (e.g. contingency tables). The type and the use of auxiliary information have an impact on the quality of final results.
- In the proposed strategy, imputation is done through the hot-deck stratified, which means that the missing values have been imputed with the conditional mean (conditional to the parameters used in the computation of the distance). When using the hot-deck method, different assumptions about CIA implies a different choice of the strata and distance variables for selecting the nearest neighbor. Furthermore, when using hot-deck models, determining which file acts as recipient and which one as donor is not negligible: since the good behavior of non parametric estimates are essentially expressed in terms of asymptotic properties, generally the file containing more observations is selected as donor file.
- Since the CIA is not testable, how to evaluate the performance of different statistical matching strategies, i.e. which aspects are to be evaluated and how: authors suggest looking at the target joint distributions by using a simulation approach.

## 3. *Points for discussion*

- Since the Conditional Independence hypothesis highly influences the final results of statistical matching, general criteria are to be found in order to reduce the impact of CIA in terms of risk of biasing effects on the estimated data relationships.
- Concerning the use of auxiliary information, its type (micro/macro) and use (imputation models, other techniques) affect final results. Therefore, given a specific statistical matching problem, the following aspects need to be further studied:
  - how to select the most appropriate auxiliary information,
  - how to assess the reliability of auxiliary information;
  - given the (micro and/or macro) auxiliary information, which is the most appropriate way (model/method) of using it.
- When performing statistical matching, the adopted approach depends on the specific objectives of the analysis: e.g. imputation seems less appropriate than other approaches when the aim is obtaining preliminary estimates. There exist general criteria to be followed when choosing among approaches statistical matching?

- Concerning the use of imputation in statistical matching, how to select the best imputation model, given the specific assumptions and hypotheses made and the available auxiliary information:
  - which classification/matching variables to use as predictors (covariates);
  - which method/model to use in the imputation phase to better exploit the predicting power of covariates (hot-deck, regression models,...);
  - which constraints can be possibly imposed at micro/macro level and how to use constraints in the models.

## **Annex 1 – Sub-topics and corresponding papers for topic (ii): Implementing editing strategies and links to other parts of processing**

### ***Sub-topic a – Links to other parts of processing: editing and estimation***

#### ***Invited papers:***

Canada – Assessing and dealing with the impact of imputation through variance estimation WP.10

#### ***Supporting papers:***

Spain – Performance of re-sampling variance estimation techniques with imputed survey data WP.19

### ***Sub-topic b – Links to other parts of processing: editing and data dissemination***

#### ***Invited papers:***

Netherlands/United Kingdom – Preserving edits when perturbing microdata for statistical disclosure control WP.11

#### ***Supporting papers:***

Finland – Release of micro survey data. How to do this ideally? WP.13

### ***Sub-topic c – Implementing editing strategies: (re-) designing editing processes***

#### ***Invited papers:***

Sweden – A selective editing method considering both suspicion and potential impact, developed and applied to the Swedish foreign trade statistics WP.12

#### ***Supporting papers:***

United Kingdom – An editing procedure for the low pay domain in the annual survey of hours and earnings WP.20

Germany – Introducing and implementing a new data editing strategy WP.14

Netherlands – The embedding of a uniform statistical process WP.18

Germany - Linking Data Editing Processes by IT-Tools

#### ***Tabled papers:***

Netherlands – Selective editing using plausibility indicators and Slice CRP.2

### ***Sub-topic d – Implementing editing strategies: using external data sources***

Israel – Planning editing and imputation as an integral part of multi-source data collection WP.16

Italy – Modelling the construction of a social accounting matrix in the context of statistical matching WP.17