EDITING STRATEGIES WHEN ADMINISTRATIVE AND EXTERNAL DATA ARE USED IN THE STATISTICAL SURVEY PROCESS

Draft – September 22,2003

INTRODUCTION

For the topic of editing strategies when administrative or external data is used in the survey process, there are four papers: Paper No. 10 - *The junction between external data and statistics data. Is it possible to optimize?* by Seppo Laaksonen from Statistics Finland; Paper No. 12 - *Treatment and editing of tax data for Swedish Structural Business Statistics* by Johan Erikson from Statistics Sweden; Paper No. 34 - *Non-response recovery by imputation using temporal extrapolation of (administrative) Profit & Loss account data in the Structural Business Survey* by Guy Vekeman of Belgium, and Paper No. 38 - *Source point data editing in health surveys* by Kenneth Harris of the National Center for Health Statistics (USA).

All of the papers deal with the editing and preparation of administrative and external data in order to incorporate them into the statistical processes at the Statistics Agency. The use of accurate and coherent administrative data will increase the quality and the timeliness of the survey data and reduce response burden and costs which would have resulted from a survey investigation for producing the target indicators. No administrative data is automatically ready for use since they are collected and processed for different purposes and data suppliers are not necessarily responsible for the quality specifications that meet the requirements of the Statistics Agency. In addition, statistical classifications and definitions are usually inconsistent with the needs of the particular survey.

The goal is to fulfill as much of the statistical requirements as possible at the point where the data is collected by providing incentives and improving the interaction and cooperation with the data suppliers. Improved methodology for measuring the accuracy and possible biases in the data, statistical modeling, editing and imputation on multiple sources of administrative data will increase the quality of the data and can compensate for missing indicators or for target sub-populations in the survey process without having to carry out ad-hoc surveys. In addition, administrative and external data can also be used to identify coverage problems with existing frames and registers.

MAIN ISSUES

- Administrative data needs to be adapted to the survey process by fulfilling all of the statistical requirements with regard to classifications, definitions, harmonizing and standardizing the content of the variables, assessing the level of quality of the data, editing and imputing missing variables or target sub-populations. The goal is to try and fulfill as much of the the statistical requirements as possible at the junction where the data is obtained from the data supplier. Data suppliers usually have different conceptions of their data with respect to accuracy and content, and have different editing practices. There is a need for interaction and cooperation between the data suppliers and the Statistical Agency to get the data suppliers to perform edit and quality checks and conform to the requirements of the Statistical Agency.
- Multiple sources of administrative data are exploited and quality guarantees need to be taken under consideration. In particular, new methodologies need to be developed for modelling, for editing and imputing of administrative data, for record linkage and estimating error rates, for assessing new types of biases that may occur specifically when incorporating administrative data into the survey process and when compensating for missing values on

particular indicators or sub-populations. Methods of imputation, ie ratio or hot-deck on previous data of the non-responding unit or through responding units within homogeneous imputation classes, need to be analyzed.

- Costs and timeliness must be taken into consideration. Incorporating multiple sources of administrative data can be costly in terms of the resources needed for extensive editing, quality checks and adapting the data to the survey. Usually this involves performing manual checks on those units that are the most influential. These costs are considerably lower than carrying out surveys for the same information, and survey data may be of lesser quality because of increasing rates of non-response. However, the cost effectiveness should be weighed against the quality of the administrative data and the possiblities of introducing new biases when the data is insufficient. Sometimes, there is a need for ad-hoc surveys to gather information about phenomenon that are not available on administrative sources. In general, response burden is decreased on the units and a higher level of data accuracy is achieved with administrative data. The amount and type of editing should be determined that balance between the level of accuracy of the data, and the time and costs due to editing.
- Administrative data provides a valuable source of information for checking and quantifying coverage problems in existing frames.
- Mixed modes of data collection complicates the overall editing strategy of the survey. Editing practices for both paper data collection and electronic data collection, where sophisticated edit checks can be incorporated directly into the data capturing stage of the survey, need to be harmonized to take into account the different levels of quality and accuracy in the data after the collection stage. The need to define the edits for electronic data capture must take into consideration the trade off between data accuracy and respondent burden.

POINTS FOR DISCUSSION

- How to determine the balance between the maintenance of a complex survey processing system consisting of multiple sources of administrative and survey data, and reducing the timeliness and costs to a survey. In particular, the trade off between a high level of data accuracy that is needed and the costs incurred for editing and checking the validity of the data.
- How can statistical agencies increase cooperation and what incentives should be given to data suppliers in order to get them to improve their editing practices and conform to statistical standards, in particular with respect to the EU context where comparable data and uniform definitions are required .
- How can administrative and external data be better utilized for quantifying frame errors and for modeling, editing and imputing in the survey process. With a wealth of administrative data that can be used as covariates, improved statistical techniques can be developed for imputing missing and erroneous data. More methodology is needed to understand and exploit administrative data and to measure possible new biases that are introduced into the survey from the administrative sources.
- How can data with low quality be identified and can it be utilized in an optimal way. What is the amount and types of edits that are appropriate for balancing between the accuracy of the data and the timeliness of the survey and the high costs due to editing.
- More and more mixed modes of data collection are carried out. How can data with different levels of quality be integrated and how can an optimal mixed editing strategy be designed to take into account the different editing activities that were performed at the various stages of the survey process.

SUMMARY OF PAPERS

WP 10

The junction between external data and statistics data. Is it possible to optimize?

Seppo Laaksonen - Finland

1. Summary

The paper reports current experiences at Statistics Finland in using external (not statistical) sources of information for statistical purposes either separately or in combination with survey data. Examples of multi-level and longitudinal registers using external sources of information from a central population register, businesses, taxation authority, employer's and pensioner's organization are reported. One particular complex statistical register described in the paper, which is developed from multiple sources of external data, is the Linked Longitudinal Employment Data (LLED). In addition, particular attention is given to the use of administrative data in the business area for efficiently developing and updating business registers. In this area, the integrated use of external data and *profiling* is discussed: profiling aims at identifying enterprises, its legal and operating structure, and production units for subsets of (large and complex) units for which administrative sources do not give all the information needed to define and maintain the statistical units. Businesses that are profiled depend on their size, the propensity to change and complexity.

The focus of the paper is centered on the problem of finding a trade off between the gain in productivity derived from the multi-exploitation of external sources of information, and the quality guarantees on external data. How to find this balance depends on the type of data suppliers, the characteristics of information they collect for their own purposes, the type of quality controls they perform on their data, the adopted definitions and classifications, the coherence and quality criteria, and the quality of linkage.

In general, no external data can be considered reasonable enough for statistical purposes without further editing since they are collected and processed for different purposes, and data suppliers are not necessarily responsible for quality from its statistical point of view. The amount and type of editing activities must take into account both the Statistical Agency and the data supplier's quality requirements.

Statistical Agencies try to exploit as much as possible information coming from external sources and often more variables are required than are minimally needed by the data supplier. Some external data can only be partially used and need to be integrated with information coming from either other external sources or ad hoc statistical surveys to fill in the information gaps. Imputation is carried out when possible. Also, some external data cannot be considered efficient at all.

In the paper, the topics that were covered included:

- Completeness and usefulness of information provided for statistical purposes, especially with regard to coverage issues, and harmonization and standardization of statistical classifications.
- Type and amount of data checks performed by data suppliers for their own purposes
- Level of data suppliers availability and cooperation in changing/updating the type and amount of data checks performed on their own data when provided to Stat Finland

In the paper, some examples of editing practises adopted by data suppliers on their own data are described. The quality requirements can be very different in terms of target variables, type of edits

applied, and the detail of the editing. Statistical Agencies (e.g. due to current EU regulations) generally have more demanding targets for the data quality than those of the data suppliers, and this fact evidently influences the type and amount of editing activities that is performed on external data.

The complexity of the editing activity increases as the differences among purposes, definitions of units and variables, concepts, and classifications increase. In most extreme cases, statistical concepts are to be "created" starting from the definitions, concepts and classifications adopted by the specific data supplier. A further complication might derive from the fact that some data suppliers collect data from other registers and/or other administrative sources using sub-contractors.

With regard to the cooperation of the data suppliers, and the possibility of making them change their own editing procedures, the data suppliers are not necessarily responsible for data quality from its statistical point of view, so they are not obliged to modify their data processing procedures. On the basis of this premise, typical scenarios identified in the Finnish experience:

- Some external data suppliers are available in reviewing their own data if the Statistical Agency is not satisfied with the quality of some subsets of data ;
- Generally, data suppliers provide the Statistical Agency with all information needed for editing and analysis activities;
- External data suppliers are often asked to provide more variables than minimally needed for its purposes to allow the Statistical Agency to exploit as much as possible the provided information.

2. Main issues

- No administrative data are automatically ready for use in the statistical process and the goal is to fulfill as much of the statistical requirements as possible at the junction between the external supplier and the Statistical Agency, and in particular to obtain complete, harmonized and standardized administrative data.
- The trade off between the advantages of exploiting multiple external sources of information, and the quality guarantees on external data.
- Different data suppliers have different editing practices and there is a need for interaction and cooperation with the data suppliers to conform to statistical classifications and improve editing practices.
- Profiling of businesses.

3. Problems

- The need to determine the balance between the costs for maintaining a complex production system (consisting of a combination of external sources and statistical surveys) and the reduction of cost due to collecting information from suppliers other than respondents.
- A general problem is the discrepancies in classifications and definitions used by data suppliers (particularly in the business area, for example, occupation and wages definitions) and the balance must be found between the gain in productivity and the effort needed by the Statistical Agency to continuously adapt their practises to external data changes.

- The needs of the external data supplier with respect to the statistical quality of their data should be identified, and incentives should be offered for editing their data better, for conforming to statistical classifications and for increasing cooperation.
- The interaction between different data suppliers and the Statistical Agency should be improved, in particular for setting up better editing practices as well as identifying and correcting erroneous data.

4. Related Aspects

- One important aspect touched in the paper relates to the case in which many external sources and statistical surveys are to be integrated to obtain reasonable and reliable information on all aspects related to a particular investigated phenomenon. Once the available external information for a given statistical investigation has been identified, the following main problems arise:
  1. Assessing the quality of external information and identifying the appropriate editing approach to make it usable for statistical purposes;
  2. Identifying missing information on specific phenomena or for specific sub-populations in the target population. Define alternative solutions for collecting the missing information directly from statistical units through ad hoc surveys;
  3. Linking all information coming from all the different sources.

WP 12
Treatment and editing of tax data for Swedish Structural Business Statistics
Johan Erikson - Statistics Sweden

1. Summary

In the paper, the Author describes the editing and imputation methods used at Statistics Sweden to deal with the use of administrative tax data to estimate economic indicators produced by the Swedish Structural Business Statistics (SBS). The main characteristics of the data processing activities needed to make the external data appropriate for statistical use are described. The adopted solutions for compensating for gaps on particular indicators or on target population subsets are illustrated.

One interesting point discussed by the Author relates to cases in which given information is available on more than one external data source. In this case, it is obvious that the most appropriate source will be selected (in terms of quality and level of detail of available information, similarity between definitions and concepts, level of coverage of the statistical needs). The possibility of using the other sources of information in the editing process should in any case be evaluated

A wide discussion is devoted to the modelling effort made in order to integrate all the available information required to derive some target indicators. The need for combining information coming from different data sources and the modelling activities imply initial high survey costs, balanced by the increase of data quality (estimated values are often preferred to low quality information) and the decrease of respondent burden and other survey costs.

The editing and imputation procedures have similar characteristics and pose similar problems as those developed for traditional statistical surveys. Internal inconsistencies and extreme values potentially influent on target indicators are looked for among data, manual review is performed on critical units while for the smaller units they are either turned into non-respondents or are imputed

automatically. Imputation is based on the average values of the variables in homogeneous groups (activity code and size group). A validation of target indicators after this editing and imputation process is then performed. In this strategy, crucial points are represented by:

- Defining the amplitude of acceptance bound for both ratio and query edits (balance between how much editing and required level of data accuracy);
- Balancing between the amount of manual editing and automatic imputation (rationalisation of resources, reduction of time, costs and respondent burden);
- Identifying the appropriate type and amount of edits to be used taking into account the statistical quality requirements and the available resources: an efficient editing system should guarantee higher data accuracy and lower costs;
- Balancing between desired level of accuracy of final data and accuracy of manual data review;
- Identifying the most appropriate imputation model for the different missing patterns, exploiting as much as possible all the available information.

A wide discussion is devoted to the problem of quantifying frame errors, and specifically for checking the under and over coverage problems particularly crucial in the business area. The need for further research in this area, as well as on data modelling, editing and imputation for administrative data are underlined by the Author.

2**. Main Issues**

- Data processing of external data and making it appropriate for statistical use, including modeling, editing, imputing and combining information from other data sources to compensate for missing values on particular indicators or for target sub-populations.
- By incorporating more administrative data on the smaller units, more resources are available for enumerating larger units that have more impact on the target variables.
- Edit rules on administrative data check for extreme values, incoherence and large errors, and also validate calculations carried out by the Statistical Agency. Large units are checked manually and the smaller ones corrected automatically or turned into non-respondents.
- Imputation based on average values within strata. All data imputed even when some variables pass edit checks.
- Use of external data to check existing frames for coverage problems.

3. **Problems**

- How external data can be better utilized for quantifying coverage problems in frames, and for modelling, editing and imputing in the survey process. More methodology is needed to understand and exploit as much as possible administrative data and in particular for measuring possible biases that are specific to the use of administrative data.

4. Related Aspects

- The Author underlines some relevant aspects relating to its particular experience:

1) The high gain in survey data quality and timeliness due to the availability of accurate external data at the required level of detail for most of the economic indicators previously produced through a statistical investigation;
2) The low amount of work to be done to adapt tax data concepts and definitions to the statistical requirements;

3) The decrease of burden on enterprises while increasing the data accuracy.

- Critical aspects to be faced are represented by:

1) The need to integrate information not available on external data or external data with low quality. This can be achieved in different ways:
   - By using (when available) other sources of information. In the paper, other external sources are to used for computing some of the target indicators that are not available in the main administrative data source. This problem is common to most surveys making use of administrative data, where generally more than one source are to be integrated in order to cover all the required statistical data;
   - By estimating data through data modelling;
   - By imputation.

   Understanding the reasons and the origin of missing data allows the selection of appropriate compensating methods (ie., other external sources, modelling, imputation). This is recognised as an area where a more intensive exploitation of external data can provide good results.

2) The need for performing editing activities to check for administrative data coherence taking into account both possible low administrative data quality, and the specific statistical purposes;
   - The amount and type of editing appropriate for balancing between the level of data accuracy, and the time and costs due to editing (generally performed through manual review for critical units as large businesses or businesses with potential high impact on target indicators);
   - The need to better exploit as much as possible external sources of information for producing both the specific statistical estimates at the desired detail, and for assessing and improving the quality of frames (like Business Registers). In the paper, an interesting analysis of how under and over coverage problems can be identified through the use of tax data taking into account the Swedish economical situation is presented in the paper.

WP 34
Non-response recovery by imputation using temporal extrapolation of (administrative) Profit & Loss account data in the Structural Business Survey
Guy Vekeman – Belgium

1. **Summary**

In the paper an experience on using administrative data to compensate for non-response in Belgian Structural Business Survey is described.

Information from Profit&Loss accounts is used to recover missing information on accounting totals, then a breakdown of these totals are imputed using methods that exploit as much as possible the observed correlations between accounting totals and the missing total details. The proposed different imputation techniques are *temporal extrapolation* (ratio estimation on data of that same responding unit from the previous survey), and *additional appropriate collective data* (ratio estimation on grossed up data of respondent companies similar with respect to known auxiliary information).

The results of a comparative evaluation study of the two imputation approaches, as well as a good analysis of the data, are described. The target variables in responding units are imputed using both methods, and then true and imputed values are compared in order to identify the better technique. As expected, if previous data are available, the *temporal extrapolation* gives better results because it exploits as much as possible temporal correlations in the breakdown of accounting totals for each given company. The second technique imputes the mean ratio within each stratum, consequently its performance is dependent on the amount of available auxiliary information that can be used to identify homogeneous strata.

The imputation strategy adopted at the Structural Business Survey to deal with each particular type of unit (large, small, typical or anomalous), type of information available (previous information, level of correlation) is described. Attention is paid to correlation levels between target variables and variables used in the imputation strategy. While the larger units use temporal imputation, the smaller units and those units without previous data, use ratio imputation.

2. **Main Issues**

- The focus is on imputation techniques and comparing methods. A thorough analysis of the data and the methods are necessary before carrying out imputation.

3. **Problems**

- A crucial point in the imputation process is represented by the need for standardizing definitions and concepts not only between administrative and statistical sources of information, but also with respect to the EU context, where comparable data and uniform definitions are required at international level.

4. **Related Aspects**

- With increasing non-response levels for the survey, it is necessary to measure and control for the risk of higher biasing effects due to the increasing use of imputation.
- Another point of discussion relates to the appropriateness of hot deck techniques in the business statistics area. If the MAR assumption is valid (i.e. if strata are built efficiently), good imputations (in terms of preservation of correlations and variability) should be obtained with both ratio and donor (either random or nearest-neighbour) methods. The point stressed was that for quantitative variables, hot deck imputation may not be appropriate, especially when inequalities need to be maintained, but for some variables (number of work hours, number of employees, etc.) its use is recommended.

WP 38
Source point data editing in health surveys
Kenneth Harris – National Center for Health Statistics (USA)

1. Summary

In the paper, the experience in moving editing as close as possible to data providers at the U.S. National Center for Health Statistics (NCHS) is described. NCHS conducts a wide range of surveys and administers the national vital statistics registration systems to collect and disseminate vital and

health statistics. Surveys are grouped into four main families, depending on the nature of the investigated phenomena.

The NCHS's ultimate goal is to adopt in each survey the so called *source point data editing* approach, consisting in performing editing as close as possible to data providers, during the data capturing stage or very near to that stage.

At present, surveys in each family adopt different approaches to collect and control data quality, in some cases the *source point data editing* approach is already used, often combined and harmonized with other data capturing modes and different data editing strategies. The data quality control strategies in these surveys are described in the paper. These strategies include the use of edit alerts for erroneous and inconsistent data in the CAPI and CATI modes of data collection in surveys, electronic data capturing of administrative information with rigorous checks, and more in general re-contacting respondents, interviewers or administrative records within a limited time following the original interview or data collection. Edits include checking for inconsistencies, missing values, out of range or invalid values, the ability to amend rosters and reaccess and change data, automated skip patterns, and allowing explanatory comments.

In the paper, the Author provides a description on the possibility and the extent of performing editing activities during or near the data capturing stage in each data collection mode, which is strictly dependent on the following aspects:

- the type of information asked: sensitive (like attitudes, illness, social behaviours) or demographical (birth, deaths, and so on)
- the type of respondent (individuals or other Agencies)
- in case of Agencies, their technological capabilities and organization

The Author provides interesting examples of how to combine different modes of data collection, and then how to design overall editing strategies by harmonizing the traditional post-data capturing editing activities with editing performed either at or very "near" the data capturing stage.

2. **Main Issues**

- Summary of source point data editing practices carried out in the surveys and statistical registers.
- Good cooperation with suppliers with rigorous checking of the data prior to its receival at the NCHS.
- Mixed modes of data collection where some systems are more sophisticated than others. A final editing process after data collection to remove remaining inconsistencies.

3. **Problems**

- Often mixed modes of data collection (with and without edits) are to be adopted because of the complexity of investigated phenomena. This imposes two kinds of problems:
  1) Integrating data of different quality levels
  2) Designing mixed editing strategies, in which the editing activities performed at the different survey processing stages are properly combined.

4. **Related Aspects**

Important aspects pointed out in the paper are:

- When using information coming from administrative sources, the quality of provided information (and hence the editing strategy to be adopted on provided data), is highly dependent on:

1) The technological capabilities and on the organization of the provider (e.g., presence of electronic archives, hence automation at the data source: the possibility of receiving data electronically increases timeliness, improves quality, reduces the risk of errors during data conversion from paper to electronic files);
2) The similarity between the administrative and statistical specifications and definitions: if there is a "political" agreement between the Statistical Agency and the data provider about how data are to be filed, this requirement can be met easily.

- The need to identify the appropriate set of edits to be anticipated at the data collection stage in order to obtain an acceptable trade off between respondent burden, data accuracy and editing complexity. One suggestion is to concentrate attention at the data capturing stage on the crucial items to the study. In some cases, it could be appropriate to repeat some checks at both stages (data capturing and post processing) to ensure that they operated properly.
- There exists some data collection constraints due to specific providers characteristics or type of information required that do not allow adopting source point editing either efficiently or at all.